Microbial population genomics and ecology
Edward F DeLong

The origins of biological complexity in microbial ecosystems are encoded within the collective genomes of the community. Cultivation-independent genomic studies provide direct access to the genomes of naturally occurring microbes, cultivated or not. Genome-enabled approaches are now significantly advancing current knowledge of genome content, diversity, population biology and evolution in natural microbial populations.

Introduction
Most definitive microbiological studies have been conducted in laboratories using pure cultures. Such studies have been critical to the development of the microbiological science, and provide the basis for our understanding of the microbial world. However, the microbial species and interactions that really count in Nature do not occur in pure culture. In fact, most naturally occurring microbes exist in complex communities, and have never before been cultivated or characterized in the laboratory. Given the above, how do we gain a more holistic picture of the players and processes that define structure and function within natural microbial ecosystems?

Norman Pace and colleagues [1] first proposed cultivation-independent approaches to study natural microbial populations. Their strategy involved the use of nucleic acid-based cloning and sequencing technologies, to access naturally occurring microbial diversity. Those seminal ideas led to the now-common practice of cloning ribosomal RNA (rRNA) genes from mixed microbial assemblages, to determine the phylogenetic identity of population constituents. Over the past decade, such cultivation-independent molecular phylogenetic surveys have revealed an astounding number of novel phylogenetic lineages, as well as ecologically dominant types of Bacteria and Archaea [2] that were totally unanticipated from culture-based studies. Our current understanding of the extent and nature of natural microbial diversity has improved dramatically, thanks to the application of cultivation-independent rRNA surveys.

The past decade has also seen an amazing acceleration in genome sequencing capabilities. Just a few years after the first report of a complete microbial genome sequence [3], hundreds of full microbial genome sequences have become available, or are currently being determined. A large percentage of the genomes sequenced (about 75%) originate from clinically important microbes. Of the total prokaryotic sequences, about 90% are from the domain Bacteria [4]. To date, most full microbial genome sequences originate from pure cultures, although a few are derived from obligately symbiotic [5] or parasitic microbes [6].

Genomics in natural microbial assemblages
A logical extension of the tremendous genome sequencing capacity available today aims to genomically better describe the natural microbial world. This really is just an extrapolation of Pace’s approach [1], but with a twist. Instead of cataloguing just rRNAs (or other single genetic loci) recovered from environmental samples, it is now feasible to determine large portions of the genomic content found within naturally occurring microbial communities (Figure 1). There are many different avenues of pursuit for this new approach: bioprospecting (for example, the ‘mining’ of new genes, enzymes or natural products from environmental samples), characterization of uncultivated microbes, and microbial population genomics may all be advanced by its application. Recent signs of the vitality of this new field of microbial population genomics are evidenced in part by the references discussed in this review.

The general scheme of cloning of large DNA fragments from natural populations was originally proposed by Pace and colleagues [1,7]. Initial schemes used phage lambda as the vector to archive natural population DNA [1,8]. Early applications surveying the phylogenetic diversity in lambda ‘shotgun’ libraries from marine bacterioplankton demonstrated the efficacy of this approach [8]. In the early 1990s, two technical developments had a major impact on the field. Because it is so technically straightforward, use of thermostable DNA polymerases and the polymerase chain reaction (PCR) [9] became a major tool for phylogenetic diversity surveys of single genetic loci, especially rRNA genes [10]. At about the same time, bacterial artificial chromosome (BAC) vectors, whose replication is controlled by the F1 origin of replication [11,12], were being developed. By virtue of their low copy number, BACs can be used to stably propagate very large DNA fragments that are otherwise unstable [12]. One version of the BAC vector (originally termed ‘fosmid’, for F1-origin-based cosmid-sized vector [11]) is introduced into Escherichia coli via phage particles, after using bacteriophage lambda packaging extracts to package the clones. Transduction of the lambda particles is extremely efficient, facilitating construction of BAC libraries with uniform insert sizes closely bracketing 40 kbp [11]. Another approach to
construct BAC libraries uses electroporation to introduce BAC clones into *E. coli*, and can produce clones exceeding 300 kbp in size, depending on the source DNA [12]. Shortly after the invention of these BAC vectors, the development of high-throughput ‘shotgun’ sequencing strategies [3] set the stage fully for genomic studies of naturally occurring microbial genomes. (Note: in the past, there has been some erroneous indication that BAC and fosmid clones are somehow inherently different. In fact, BAC and fosmid clones are identical in both vector sequence and copy control, once propagated in *E. coli*. With some ‘environmental DNA’ samples, more total genomic DNA and approximately the same average insert sizes can often be archived using the fosmid cloning strategy to generate BAC clones).

**Beyond rRNA genes – large genome fragment libraries from natural populations**

One approach for cloning large genome fragments from environmentally derived DNA is to use high fidelity, low-copy-number BACs described above [11,12]. The first application of BAC vectors to recover DNA from a mixed microbial population was reported in 1996 [13]. In this study, a BAC library was constructed from microbes collected at a depth of 200m in the Pacific Ocean to characterize uncultivated marine archaia. A 40 kbp DNA fragment from a planktonic archaean was identified on a BAC and partially sequenced, providing the first glimpse into the gene arrangement, content and identity of marine archaia [13]. Large genome fragments derived from bacterioplankton of the order *Planctomycetales* were also later identified in the same library [14]. The same approach was later used to clone genome fragments from the uncultivated symbiotic archaean *Cenarchaeum symbiosum* [15]. The utility of expressing and characterizing specific gene products from uncultivated microbes was demonstrated [16]. Additionally, the nature and extent of population genome heterogeneity could be explored in some detail using *C. symbiosum* BAC libraries [17].

Recent advances include the recovery of greater overall amounts of DNA in environmental DNA libraries, and larger DNA. Rapid advances in high-throughput screening, sequencing and robotics have also greatly facilitated more thorough analysis of the recovered clones. These technological advances are vastly improving the economic and technical feasibility of cloning, screening and sequencing large numbers of clones derived from natural environments.

There has been a good deal of interest in recovering microbial DNA from soil, with most studies focusing on bioprospecting for drugs, enzymes and other natural products. This type of approach has been in use now for nearly a decade in the biotechnology industry [18]. Directed bioprospecting in small (5–8 kbp) insert environmental DNA libraries has been fruitful on several fronts, for example, for isolating novel lipolytic, 4-hydroxy butyrate dehydrogenase or chitinase genes [19–22]. More recently, a two-pronged approach of combining environmental discovery with laboratory evolution for identification and optimization of industrially important biocatalysts, in this case α-amylase, was reported [23].

In one recent report [24**], microbial DNA was extracted directly from soil and cloned into BAC libraries. Two different so-called ‘metagenome’ libraries were recovered and analyzed, one with an average DNA insert size of 27 kb, and the other averaging 44.5 kb. The phylogenetic composition of ribosomal RNA genes in the library was consistent with other phylogenetic surveys in soil. Although the average insert size of these libraries does not much exceed that of conventional lambda or cosmid library approaches, this report was one of the first to retrieve
large DNA fragments from soil microbial populations [24••]. The authors also showed that a small fraction of the BAC clones expressed identifiable phenotypes, including DNAse, lipase and amylase activities [24••]. Another recent study focused on a BAC library, recovered from soil microbes, that had an average insert size of 37 kbp and a range of 5–120 kbp [25]. Similarly, Brady et al. reported the recovery of the biosynthetic gene cluster for the broad-spectrum antibiotic violacein from a soil DNA cosmid library [26].

To assess the nature and function of naturally occurring planktonic microbes, many of which had resisted cultivation, Béjà et al. [27••] constructed a BAC library from surface-water microbes of Monterey Bay. These BAC libraries had an average insert size of about 80 kbp, with some clones exceeding 150 kbp in size. These results represent some of the most encouraging so far, and demonstrate the feasibility of generating very large insert BAC libraries from mixed microbial population DNA. Analysis of this surface seawater BAC library showed that the phylogenetic representation mirrored that of PCR-amplified rRNA clone libraries using PCR [27••]. These results suggested that the library contained significant genomic information that could provide insight into the biological properties of naturally occurring planktonic marine bacteria and Archaea, cultivated or not.

In a follow-up study on the Monterey Bay surface water BAC library, the sequence of a 130 kbp BAC clone derived from a SAR86 type bacterioplankton revealed the unanticipated presence of a rhodopsin-type protein gene (dubbed proteorhodopsin), a class of enzyme never before observed in the domain Bacteria. When expressed in E. coli, the novel proteorhodopsin proved to be a light-driven proton pump [28••], similar to bacteriorhodopsins found in extremely halophilic archaea. These data suggested that the SAR86 bacterioplankton represented a new, abundant type of phototroph distributed in oceanic surface waters worldwide. The presence of native proteorhodopsin in ocean waters, and the discovery of its spectral ‘tuning’ as a function of depth, was subsequently demonstrated in a verification study in the field [29]. In aggregate, this work shows that genomics in microbial populations is more than just an exercise in bioinformatics. Hypotheses and reagents are generated as part of the process, which can lead to the identification and verification of novel properties and processes of natural ecosystems.

The same surface-water Monterey Bay BAC library was screened for evidence of bacteriochlorophyll-containing aerobic anoxygenic phototrophic (AAP) bacteria [30] that had recently been reported to be abundant in seawater [31,32]. A number of BAC clones were shown to contain ~40 kbp ‘superoperons’ that encoded the photosynthetic reaction center, carotenoid and bacteriochlorophyll biosynthetic AAP bacteria genes. Analyses of the genomic structure of these AAP bacteria photosynthetic ‘superoperons’ showed that some of the most prevalent planktonic phototrophs were not the types predicted from knowledge of characterized and cultivated strains. Genome-based discoveries are bound to be plentiful, in the context of naturally occurring microbial populations.

**Variety is the spice of life – microbial population genomics**

Environmentally oriented microbial genomics is also catalyzing in-depth studies of microbial population genetics and placing them in an ecological context. These studies employ both cultivated isolates with habitat-specific phenotypes (for example, ‘ecotypes’ [33]), as well as large DNA fragments recovered directly from the environment [17,34•]. A timely example of blending comparative genomics with microbial population biology comes from recent comparisons of two closely related Prochlorococcus species [35••]. Prochlorococcus is a chlorophyll-b-containing marine cyanobacterial group that accounts for as much as 50% of the photosynthetic biomass in the open ocean. High- and low-light Prochlorococcus types differ in their chlorophyll a/b ratios, in their irradiance optima for photosynthesis, and in their relative distributions in the water column. Yet their small subunit rRNA sequences differ by only 3%. Recently, the complete genome sequence of two strains, a high-light-adapted strain (MED4) and a low-light-adapted strain (MIT9313), were determined [35••]. Comparative analyses revealed the genomic origins of some of the physiological and ecological variability in the closely related Prochlorococcus ‘ecotypes’. The low-light-adapted type has a significantly larger genome (2.4 Mbp) than the high-light-adapted type (1.7 Mbp), and has more genes associated with the photosynthetic apparatus, including phycoerythrin biosynthetic genes. The highlight-adapted type, in contrast, has more genes in the ‘high-light-inducible protein’ category, as well as more genes for UV-damage repair [35••]. Differences in the complement of nitrogen assimilation genes between these closely related Prochlorococcus strains may also influence their distribution in the water column [36].

In the past, most information on microbial population genetics has been inferred from multilocus enzyme electrophoresis or multilocus sequence typing, and largely focused on pathogens [37]. Such studies have involved mainly comparisons of cultivated strains, and not so much on genuine within or between population comparisons. Now, using genomics, it is possible to approach microbial population biology on its own terms — in real populations. For example, one recent study demonstrated that tremendous genomic variation can exist within a single population of microbes, which are virtually indistinguishable by rRNA sequence [34•]. The implication is that enormous allelic variation exists among free-living microbial species of the same population. This type of microbial population biology is as yet virtually unexplored. Microbial population genomics now promises to better define the relative importance of lateral transfer, recombination and genetic
drift on the diversification of microbial species and genomes within and between populations. Such studies should spur the development of new theory, and greatly improve our understanding of microbial evolution as it occurs in different organismal, ecological and population contexts.

The challenges
There are a number of challenges associated with the construction and analysis of large insert BAC libraries prepared from very-high-complexity natural microbial populations. Assessing the extent and fidelity of genome recovery is important, but difficult to quantify. Screening for phylogenetic representation within the library (via rRNA genes) is one approach to assess the diversity of genomes recovered \[24**, 27**\]. Statistical methods to estimate genome recovery in populations may also prove useful \[38\] but will require a fairly extensive sampling regime. The type of representation desired in the libraries also depends on the purpose of the study. For genome-oriented population biology studies, sample ‘normalization’ is not desirable, as quantitative information about genome representation is a major goal. Bioprospectors, on the other hand, may wish to maximize the diversity of recovered types and therefore use normalization methods to amplify representation of low-abundance genome types. In either case, both the ‘richness’ (species number) and ‘evenness’ (relative representation of each species) of microbial genomes in the original sample will greatly influence the end result. Another factor influencing genome recovery is the fact that, even with low-copy number BAC vectors \[39\], there will inevitably be differential recovery of different genomic sequences. These difficulties can arise from the presence of highly repetitive, tandem sequence motifs \[39\], or because gene expression of recombinant DNA in BAC clones does occur (especially if it is of bacterial origin) \[40\], and this may sometimes prove lethal.

Another concern that arises centers around the possibility of forming chimeric BACs from the DNA fragments of different organisms. Although a potential problem, the methods employed during library construction help reduce this risk, in particular, size selection of DNA fragments before cloning. Additionally, protocols that use lambda packaging to introduce BACs into \textit{E. coli} have an additional size selection during packaging that further reduces the risk of chimeric BAC recovery. Several methods can also be used to rule out the possibility that any given BAC is chimeric. BAC population screening helps tremendously: homologous but non-identical BACs can generally be identified in libraries, which represent closely related but non-identical strain variants in any given microbial population (see, for example, references \[17, 34**\]). Near-identical genomic structure in these BAC variants, including gene content, organization and synteny, in essence prove that the genomic structures in question are not artefacts (for example, not chimeric). In addition, there are high-resolution optical mapping methods becoming available \[41\] that might be used to test and verify the integrity of individual clones. Similarly, chromosome painting techniques \[42\] might also be used to verify the integrity and presence of paired-end BAC sequences in individual cells.

Conclusions
Zucherandl and Pauling \[43\] posited that macromolecules are documents of evolutionary history. A logical extension of this view is that naturally occurring genomes are documents of environmental, evolutionary and ecological history, on a global scale. Environmental genomic approaches are cultivation-independent, and so provide equal access to the collective genomes of abundant, yet-uncultivated microbial species. Genomic analyses of uncultured microbes can provide significant insight into the biological properties of individuals within microbial populations. Applied microbial population genomics is now facilitating predictions about phenotype from genotype, broadening our knowledge of natural environmental processes, and extending our ability to exploit natural microbial products and processes. A better understanding of genome evolution will undoubtedly result from microbial population genomic studies. Genomic windows into the natural microbial world now also provide unprecedented insight into microbial population structure, microdiversity and dynamics. Additionally, the genomic patterns that give rise to emergent properties of populations and their relationship to ecosystem function are bound to become increasingly more apparent.

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest


in soil-derived BACs was consistent with the known phylogenetic diversity populations. Two different BAC libraries were prepared from naturally occurring  

This is one of the first reports of a large insert library from soil microbial  

Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C

24. Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH,  

23. Richardson TH, Tan X, Frey G, Callen W, Cabell M, Lam D, Macomber J,  

22. Cottrell MT, Moore JA, Kirchman DL: Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on  


This is one of the first reports of a large insert library from soil microbial populations. Two different BAC libraries were prepared from naturally occurring soil microbes, one with an average insert size of 27 kb, and the other averaging 44.5 kb. The phylogenic composition of ribosomal RNA genes in soil-derived BACs was consistent with the known phylogenetic diversity in soil. Several functional genes were shown to be expressed from the BACs by phenotypic screening.


This report documents that extremely large insert BAC libraries can be constructed from mixed microbial populations. BAC libraries with an average insert size of about 20 kb, with some clones approaching 200 kb in size, were constructed from the genomic DNA of planktonic microbial populations in Monterey Bay. One clone containing a genomic fragment from an uncultivated, planktonic archaean (kingdom Euryarchaeota) was fully sequenced and its genes annotated.


A 130 kb BAC clone from an un cultivated marine planktonic γ-proteobacterium (SAR86 group) was sequenced. One of the genes found in the SAR86 genome coded for the first rhodopsin (proteorhodopsin) ever found in the domain Bacteria. The proteorhodopsin was expressed in E. coli and, in the presence of retinal, was shown to be a light-driven proton pump, similar in its properties to halorarchael bacteriorhodopsin. The results suggested that an unsuspected type of rhodopsin-based phototrophy is common in ocean surface waters.


BAC clones from a natural microbial population containing identical or near-identical rRNA genes (from planktonic archaea) were characterized. Among rRNA-containing BAC clones that were virtually indistinguishable by rRNA sequence, there was tremendous diversity in rRNA-linked, enzyme-encoding gene sequences. The data indicated the existence of extremely high, within-population genomic diversity in sympatric species that have identical rRNA sequence.


This is a preliminary report on the comparative genomics of two closely related Prochlorococcus strains, one of the ocean’s most abundant phototrophs. The high-light and low-light variants differed greatly in genome size, GC content and gene content. The composition and variation in photosystem genes was consistent with their physiophotology and ecological distribution.


