MiniReview

Environmental genomics, the big picture?

Francisco Rodríguez-Valera *

Evolutionary Genomics Group, División de Microbiología, Universidad Miguel Hernández, Campus de San Juan, 03550 San Juan de Alicante, Spain

Received 24 October 2003; received in revised form 19 December 2003; accepted 19 December 2003

First published online 23 January 2004

Abstract

The enormous sequencing capabilities of our times might be reaching the point of overflowing the possibilities to analyse data and allow for a feedback on where to focus the available resources. We have now a foreseeable future in which most bacterial species will have an annotated genome. However, we know also that most prokaryotic diversity would not be included there. On the one hand, there is the problem of many groups not being easily amenable to culture and hence not represented in culture-centred microbial taxonomy. On the other hand, the gene pools present in one species can be orders of magnitude larger that the genome of one strain (selected for genome sequencing). Contrasting with eukaryotic genomes, the repertoire of genes present in one prokaryotic cell genome does not correlate stringently with its taxonomic identity. Hence gene catalogues from one environment might provide more meaningful information than the classical species catalogues. Metagenomics or microbial environmental genomics provide a different tool that gravitates around the habitat rather than the species. Such tool could be just the right way to complement ‘organismal genomics’. Its potential to advance our understanding of microbial ecology and prokaryotic diversity and evolution is discussed.

W© 2004 Federation of European Microbiological Societies. Published by Elsevier B.V. All rights reserved.

Keywords: Biodiversity; Metagenome; Environmental genomics; Genome evolution

1. Introduction: Culture is not enough

Prokaryotic micro-organisms represent by far the largest reservoir of genetic diversity on Earth. They are much older than protists, metazoans and plants (which largely derive from the exploitation of prokaryotic endosymbionts) and occupy much larger sections of the biosphere (including the Earth crust). They outnumber all other organisms, have a larger biomass and make the planet liveable by managing its biogeochemistry, cycling nutrients and breaking down wastes, natural or anthropogenic. Besides, microbes possess the highest potential for production of bioactive products, enzymes, polymers and most of the tools used in biotechnology. Yet prokaryotic diversity is one of the greatest knowledge gaps in the biological sciences and remains largely unexplored and unexploited. Furthermore, there are even conflicting views about what is the real extent of this diversity. It is not only that there are very different estimates about the real number of bacterial species [1–4]. There are even more serious misgivings as to whether there are species at all, and if they exist what they really represent in terms of the diversity they contain or what evolutionary principles govern their origin and change [5–8].

Bacteriology developed around the technique of pure culture and the study of pure cultures is still the axis from taxonomy to genomics. However, we know now with a high level of certainty that culturability of bacteria in most environments is very poor, most often in the range of 1% or less. What is more important, we know that many groups of prokaryotic microbes are very reluctant to grow in the laboratory, and will have to be studied by alternative means if we want to achieve a more realistic description of the microbial world. These facts have become common knowledge. Less often considered, but probably equally important, is the fact that even within the boundaries of a well-known and easy-to-cultivate species such as Escherichia coli may lie a vast gene pool that is not accessible by studying one single strain. The actual diversity of the genes within a bacterial species is another important indeterminate in the study of prokaryotic diversity [9]. Different strains that maintain a nearly identical genomic ‘backbone’ may have genomes that differ dramatically in size and gene content, and this ‘within-species’
diversity can have dramatic impact in the physiological and ecological diversity (see below).

There is little doubt that the culture avenue should be revitalised since the possibilities of culture are yet very far from exhausted. For example, just the use of high-throughput sequencing of identity markers such as the ribosomal RNA genes and spacers, combined with innovative culture methods, could lead to retrieving many new groups that can then be studied by conventional methods (including organismal genomics) [10,11]. However, it is doubtful whether the culture approach will allow a very rapid improvement of our knowledge of microbial diversity. Microbes are just too diverse and the parameters to consider too many. Furthermore, there is certainly the possibility that very important groups are impossible to get in the high population densities required to study cultures or that they require co-culture with other microbes. Therefore, even though culture will stay a very useful tool for the study of microbial diversity, and should probably be included in research projects more often than our polymeric chain reaction (PCR)-dominated laboratories tend to do, there is a need for alternatives.

2. Environmental genomics or metagenomics

Cloning microbial genes directly from the environment is really a rather old idea. The cloning strategy has been used for retrieving 16S rRNA genes from environmental samples since the mid-1980s, before the advent of PCR. However, the always decreasing cost of sequencing and high-throughput methods to clone, store and screen environmental libraries have enormously enlarged the possibilities. Environmental genetic libraries in vectors that carry large segments of DNA (40–150 kb) can now be constructed and relevant clones sequenced or screened for the expression of gene products (good methodological sources are references [12,13]). This novel approach, in which larger fragments of environmental DNA are the object of study, has received different denominations, from environmental genomics to ecogenomics [14,15], although probably the most common name derives from the paper by Rondon et al. [16] where the term metagenome was used to describe the compound genome of the whole microbiota, specifically in soil. Hence the term ‘metagenomics’ or the study of the metagenome has become the most popular name. The idea of the genomes of all microbes present in an specific habitat as an entity carries a certain conceptual load. The integrated functioning of the microbial community in the ecosystem, particularly the osmotrophic compartment (mostly prokaryotes, although in some environments the contribution of protists would be significant), is not an unrealistic proposition. The pool of soluble compounds found in the milieu will be the result of the combined activities of enzymes produced by the whole community, and syntrophic and antagonistic relationships, coupled and combined with cell to cell communication, could make the community act as a kind of meta-organism, not very differently from how the different tissues and organs of a multicellular organism would work. Hence the parallelism between a multicellular genome (animal or plant) and the microbial metagenome is not devoid of a certain logic. Even in terms of size, the metagenome of many microbial habitats will be within the range of, for example, a metazoa genome. These ideas lend conceptual support to the very serious attempts, that are already under way, at sequencing the complete microbial metagenome of certain environments.

At a more directed level, the vast potential of the thousands of genomes of micro-organisms present in environments such as soil, waters or sediments can be exploited for biotechnology using this approach. This was appreciated very early, as witness the large number of companies that were set up specifically for this end, plus countless programmes in more diversified biotech outfits. Actually the private sector has been much more aware of the possibilities than the funding agencies, illustrating once more that it needs little enhancement from public funding when a winning research line is in sight. In principle, the advantages of this approach for exploiting the putative horn of plenty of the not-so-easy-to-grow microbes are obvious. Particularly when an easy screen for the desired activity is available, as is the case for many enzymes and bioactive compounds, positive clones can be easily detected and even selected for. Naturally, the main drawback of this approach lies in the requirement for the gene(s) to be expressed in what would often be a widely different genomic environment. Although the alternative of screening by sequence, i.e. not requiring gene expression (PCR, hybridisation, etc.), is always available, there are obviously many possibilities of improving the process by development of better vectors and surrogate hosts. The latter is particularly promising and relatively unexplored. In this sense, easily transformable microbes with well-known genetics and displaying similar activities would be the most appropriate. Environmental genetic libraries constructed in vectors that carry large segments of DNA offer nearly as many possibilities of exploitation as the cultures of the organisms themselves. Potentially even more, since the surrogate host used is always easy to handle in the laboratory. The biotechnological applications of environmental genomics are already very advanced and I will not develop this point since very good reviews are already available (for a recent review see [17]). Hereafter I will concentrate on the potential of environmental genomics to help in the understanding of the ecology and evolution of prokaryotes.

3. Describing prokaryotic ecology and diversity

The study of the diversity of bacteria has been revolu-
tionised by the PCR amplification of genes that allow some kind of preliminary identification (mostly of the 16S rRNA gene) directly from environmental samples. The most relevant conclusion is that a vast section of prokaryotic diversity is not represented in culture collections and hence remains basically unknown. The main catch of the methodology is that, most often, there is no way to correlate the 16S rRNA sequence of a specific microbe with its ecophysiology. Not even with high matches and within well-defined phylogenetic groups such as cyanobacteria does the situation improve much. Although some hints about the biology of these unknown groups can be revealed from semi-quantitative assays such as FISH or ‘in situ’ physiological studies such as MAR-FISH, these techniques are cumbersome and strictly limited in their resolution power. The lack of finesse of the 16S rRNA gene catalogues is reflected by the relatively small contribution to a better understanding of ecosystem functioning. For example, one of the most common marine microbes, the marine Archaea of group I, has been known for 11 years and very little has been found out about what they actually do in the ocean [18,19]. It is true that the combination of 16S rRNA sequencing and FISH has permitted real advances in a few cases, but that has most often been only in very simple ecosystems such as endosymbionts or extreme environments [20]. The approach of using PCR for amplifying ecologically relevant genes such as *nif* and others is still far from exhausted and will no doubt provide very interesting information [21,22]. However, protein coding genes are much less conserved at the DNA sequence level due to the genetic code degeneracy that often complicates the experimental work. This also increases the risk of biasing or missing entirely important groups. A recent work on proteorhodopsin, a retinal protein that could provide an additional energy source for marine bacteria, showed that PCR could unveil new and different types of this gene, but miss others entirely [23,24].

Sequencing clones from a metagenomic library is potentially much more fruitful. The sequencing of a larger stretch makes it possible to detect the whole gene cluster or operon providing more refined information about the organismal physiology. Furthermore, the surrounding genomic environment often helps identify the carrier organism by similarity to known taxonomic (cultivated) groups. The presence of a cluster of genes in an environment is much more relevant ecologically than the presence of one specific gene and provides much more information about the organism carrying it. Besides, even though cloning is not free from biases, it is less prone to artificial results than PCR and the analysis of the sequence provides a more stringent quality control of the results. The limiting step for this kind of gene-targeted study is the screening of the library to select clones containing the gene(s) of interest. Typically an environmental genomic library for a system of medium complexity, such as the water column in an aquatic environment, will contain numbers of clones in the range from tens of thousands to hundreds of thousands. The selection of the clones that have to be sequenced is not a small endeavour. Of course, PCR screening based on conserved regions of the target gene is always an option. It will have similar limitations as any PCR-based approach, but is one of the most popular methods due to the widespread use of PCR at most microbiology laboratories. Protocols based on pooling of clones allow a reduction in the number of PCR reactions required. The potential of micro-array hybridisation here is very large, and this technique will probably win once it is widely available and less costly.

Obviously there is also the possibility of working in the other direction, i.e. searching for taxonomically informative genes (again the 16S rRNA genes are the most obvious candidates) belonging to target groups, with the hope that the surrounding genes will provide information about important physiological aspects about this operational taxonomic unit (OTU). The huge databases of rRNA genes provide an important advantage and very useful tool. There are methodologies to create libraries enriched in or composed exclusively of DNA pieces containing parts of the rRNA operon, for example, using homing endonucleases such as I-CeuI that cut only the 23S rRNA gene. For that matter it would not be difficult to devise cloning strategies that restrict libraries to clones containing specific genes, e.g. by targeting those genes by cassette mutagenesis with a selectable marker. In any case, bias towards genes more similar to those used for primer, probe or recombination target will be difficult to avoid.

### 4. Bulk metagenome sequencing, pros and cons

Sequencing of whole metagenomic libraries is presently becoming feasible. Initiatives to sequence the Sargasso Sea and the human gastrointestinal tract have been made public. Aside from the cost involved, there is little doubt that there are great advantages to this blind approach. Avoiding the screening step will decrease the risk of bias, while increasing the chances to get really new genes belonging to the actually predominant microbes ‘in situ’. The possibilities of assembling real genomes from this kind of information are probably small, although the growth of organismal genomes in databases will greatly facilitate the process of annotation and even assembly. Bioinformatic tools designed to detect specific codon adaptation indices, trinucleotide frequencies, etc. will also help. Furthermore, the catalogues of gene clusters generated might be as informative as individual genomes of cells, or at least complementary. Ecological and physiological diversity might lie as much in the combinations of different adaptive pathways within a single cell (or in different cells that can cooperate syntrophically) as in the taxonomic variation that
would be reflected by the backbone genome of the different species.

However, the approach is not devoid of pitfalls and drawbacks. Consider as an example the Sargasso Sea project. A decision has to be made about the sample, in the open ocean it is to be expected that it will be from the water column, but depth will be critical in determining the microbiota. If there is a deep chlorophyll maximum, as is often the case in oligotrophic waters, this would probably be the obvious choice. However, it would still represent only this specific situation in an ecosystem that extends 4 km in depth. Hopefully the sample will be devoid of macro-organisms and as free from bioturbation as possible, large numbers of fish or marine mammals will no doubt affect the results of the study. Pooling DNA samples retrieved at different times would help decrease the impact of local variation and the risk of sequencing an accidental situation. Tropical waters have little seasonal variation but weather patterns, currents and other factors could no doubt also affect the results. Another essential question would be the methodology used for biomass collection and DNA extraction. It is a well-known fact that the microbiota associated with particles differs largely from that of free-living planktonic cells. Excluding particles larger than 5 μm will bias the results towards the latter, but using all the biomass without pre-filtration could lead to massive proportions of algal and/or plastid DNA obliterating the prokaryotic contribution. DNA extraction might be another source of important unequal representation of different groups. Some bacteria have extremely resilient cell walls and they would be underrepresented in nearly any approach. Not to mention the ‘viral metagenome’ that might overlap with the prokaryotic one, but will remain untapped except for the highly lysogenic or defective phages. This little discussion is meant to illustrate how even the most apparently homogeneous and reproducible system can be subjected to sampling biases and lack of representation of the real diversity present. In more heterogeneous systems such as soil or sediment, establishing the boundaries of the sample to be characterised by metagenomics can be a major handicap to the rationalisation of the use of the sequencing resources.

In any case, the repetitive sequencing of backbone stretches from the predominant genomes would be the worst risk of wasting resources. Given the uneven distribution of OTUs in most environments, a few genomes will be largely overrepresented in the total DNA. Actually core genomes will always be somewhat repetitive while the most divergent genes are found among the easily exchangeable adaptive pools. For this, some strategies such as screening for mobile elements such as transposons or integrons might be helpful [25]. Subtractive hybridisation with pure cultures might help enrich the cloned DNA in the uncultured fraction of the community. A metaproteome could be used to detect the genes expressed most abundantly in the environments under different nutrient regimes or external forcing. All this could help to explore more efficiently the metagenome. Understanding how things are done in the ecosystem and who is in charge of the critical steps might be tricky, but presently metagenomics seems one of the most promising avenues to work it out.

5. Evolution, the Holy Grail

For the last two decades ambitious attempts to reconstruct the evolution of prokaryotic lineages have been fostered at many laboratories around the world. However, the representational biases mentioned above have rarely been considered. And they should certainly be appreciated as a major limitation. The population genetics of bacteria is still in its very infancy. However, some studies carried out with pathogenic bacteria and the comparative genomics of closely related species are starting to produce an image that is in obvious conflict with the simplified neo-Darwinian concepts most often taken for granted in the mentioned reconstructions [26,27]. Evolutionary innovation in eukaryotes, which have provided the experimental basis for those models, relies mostly on carrying different versions (alleles) of the same gene. On the other hand, the complement of genes is very conserved in eukaryotic lineages. For example, four sequenced genomes, belonging to the genus Saccharomyces (yeast), contain barely a dozen unique genes each, versus c. 6000 shared genes [28]. Even the genomes of metazoans sequenced to date reflect remarkably similar gene contents (Caenorhabditis, Drosophila and human share over 50% of genes), a fact celebrated for enormously facilitating the study of human genetic diseases [29]. Contrastingly, the pool of genes shared by prokaryotic strains decreases dramatically when the taxonomic distance increases, even within a single species. I already mentioned the strikingly different gene pools of different virotypes of E. coli. The comparison of the genomes of three E. coli strains (uropathogenic, saprophytic and enterohaemorrhagic) showed that only about 40% of the total number of genes identified were present in all three [30]. The differences are even much larger when wider taxonomic distances are considered. For example, while 1803 genes were shared by three Bacillus species (40–50% of the total gene content in each), the introduction of two other low-GC Gram-positives belonging to different genera (Clostridium acetobutylicum and Staphylococcus aureus) decreased the pool of shared genes to a meagre 354 [31]. The situation in the γ-Proteobacteria was even more dramatic, and in a recent survey of 13 genomes, from a total of more than 14 000 families of orthologous genes, only 275 were present in the 13 genomes while 8035 were present in only one genome [32]. This shows that, regardless of the taxonomic scale, the genetic diversity in prokaryotes is mostly due to the different repertoire of genes that each cell carries. The currently
hot debate about the relative contribution of vertical descent or horizontal transfer in prokaryotic evolution thus becomes largely irrelevant, since most genes present in a given species, or even a cluster of strains, are only found in their kin and are not amenable to intertaxon comparisons. Furthermore, in my view, this pool of genes, not necessarily associated with the taxonomically centred genomic core, become the real players in microbial evolution. How diverse are they? Where do they come from? How easily are they exchanged? At a different level, the genetic diversity reservoir represented by viruses might also be very large and critical for the evolution and diversification of prokaryotes [33]. The subject is beyond the scope of this review but it certainly shows that prokaryotic OTUs (species if you wish) are very difficult to describe at the level of genetic diversity. Presently there is no way to forecast the size of the gene pool within a well-defined species such as E. coli. The genomic backbones shared by bacterial groups such as the Enterobacteriaceae [32] might be only the frame in which the ever-changing seascape of bacterial evolution takes place. As was very graphically depicted in a recent review about the evolution of pathogenicity, the recipe 'add DNA—stir—reduce' seems to be a central motif [34]. It might be argued that pathogenic bacteria are a special case, particularly human pathogens that are still freshly adapted to the population explosion of their new host. However, information from some non-pathogenic bacteria reveals similar dynamics [35-37]. In any case, and in the absence of any other models, similar evolutionary patterns might be expected, although the time scale might be longer.

The previous discussion gives even more reasons to value the metagenomic approach. In a community of plants or animals a catalogue of species provides most of the information to understand the ecosystem. On the other hand, in prokaryotes catalogues of genes might be much more informative. Furthermore, the demonstrated trend of functionally related prokaryotic genes to cluster together [25,38,39] provides a golden opportunity to harvest the closest thing to evolutionary units. To advance in this field, the study of some metagenomes seems an appealing alternative, or at least complement, to the isolation of pure cultures and the sequencing of their genomes. Besides, the last will always be strongly biased towards rapidly growing laboratory-adapted isolates. Actually, the classical microbiology that involves study of pure cultures might deeply affect our present view of microbial evolution. Strains tend to cluster into species-like units that might just represent culture bias, i.e. clones related by their ability to grow well and adapt to the pure culture environment. One way or another, by random sequencing of environmental DNA this bias will not be acting.

This might be the real bonus of environmental genomics, opening a window to the fascinating world of microbial adaptation ‘in situ’. Unhindered by strain isolation, metagenomics could give us a snapshot of the real genome dynamics as it takes place in a microbial habitat. We might still be far from getting the big picture, but this is one of the most exciting and important conundrums in contemporary biology and we should not expect it to yield too easily. In spite of the present state of uncertainty, there is little doubt that the progress in this field would really change our view of microbiology.

Acknowledgements

The author is presently receiving funding from the European Commission Projects MIRACLE (EVK-CT-2002-00087) and GEMINI (QLK3-CT-2002-02056) and the Spanish Ministerio de Ciencia y Tecnologia, Grants BIO2000-0085-P-04 and PM99-0078. I am very grateful to Alex Mira (Universidad Miguel Hernández) for his advice and comments. Regular visits to the laboratory of W.F. Doolittle (Dalhousie University, Canada) have widened and updated the author’s views on prokaryotic evolution. Support from Genome Canada for one visit is gratefully acknowledged. The comments of two anonymous reviewers improved the manuscript.

References
