REVIEW ARTICLE

INTERNATIONAL MICROBIOLOGY (2010) 13:45-57
DOI: 10.2436/20.1501.01.110 ISSN: 1139-6709 **www.im.microbios.org**

INTERNATIONAL MICROBIOLOGY

The bacterial pan-genome: a new paradigm in microbiology

Alex Mira,[1¶] Ana B. Martín-Cuadrado,[2¶] Giuseppe D'Auria,[3,4] Francisco Rodríguez-Valera[2]

[1]Department of Health and Genomics, Center for Advanced Research in Public Health (CSISP), Valencia, Spain. [2]Evolutionary Genomics Group, Miguel Hernandez University, San Juan, Alicante, Spain. [3]Joint Unit of Research in Genomics and Health, Centre for Public Health Research (CSISP) and Cavanilles Institute for Biodiversity and Evolutionary Biology, University of Valencia, Valencia, Spain. [4]CIBER on Epidemiology and Public Health (CIBEResp).

Received 5 May 2010 · Accepted 31 May 2010

Summary. Bacterial strains belonging to the same species vary considerably in gene content. Thus, the genetic repertoire of a given species (its "pan-genome") is much larger than the gene content of individual strains. These variations in DNA material, together with differences in genomic structure and nucleotide polymorphisms among strains, confer upon prokaryotic species a phenomenal adaptability. Although the approach of sequencing multiple strains from a single species remains the main and often easiest way to study the pan-genome, feasible alternatives include those related to DNA hybridization. In other cases, the use of metagenomic sequences is already applicable by data mining from the growing metagenomic databases. Eventually, the single-cell genome approach might be the ideal solution. The pan-genome concept has important consequences for the way we understand bacterial evolution, adaptation, and population structure, as well as for more applied issues such as vaccine design or the identification of virulence genes. [**Int Microbiol** 2010; 13(2):45-57]

Keywords: bacterial pan-genome · bacterial species · bacterial evolution

Introduction

The idea that bacterial genomes within a single species can vary widely in DNA content is not new. Determination of genome size by pulse-field gel electrophoresis in the 1980s and 1990s showed that not only different representatives of the *Escherichia coli* ECOR collection had genomes ranging from 4.5 to 5.5 Mpb [3], but also that there was a certain association of size with the MLEE (iso-enzyme pattern) group. However, it was only with the advent of the genomic era that the phenomenon could be properly appreciated. Not only was the size of the genome different, but a significant portion of the genes present in different virotypes of *E. coli* were not even related, i.e., there were no homologous genes between one virotype and another. This is, in fact, not surprising considering the relatively unrestricted "sex life" of bacteria and the inherent metabolic and genomic flexibility that characterizes the prokaryotic cell. As is often the case, to

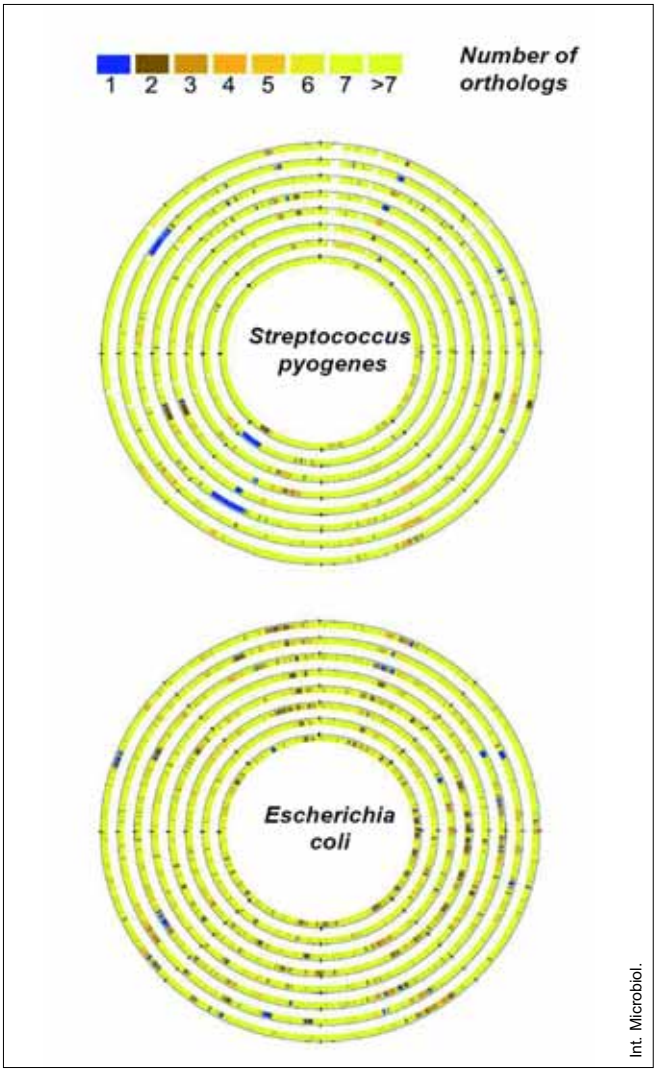
***Corresponding author:** A. Mira
Center for Advanced Research in Public Health (CSISP)
Av. Cataluña, 21
46020 Valencia, Spain
Tel. +34-961925925. Fax +34-961925703
E-mail: mira_ale@gva.es

¶Equal contributors.

This article contains supplementary information online, consisting of one table (Table S1), at the journal website [www.im.microbios.org].

*This article is based on the Closing Lecture of the 22nd National Congress of the Spanish Society for Microbiology (SEM), given by the author in Almería, Spain, on September 21, 2009. The first author was awarded the SEM's 2009 Jaime Ferrán Prize.

expect the opposite reflects the prejudices acquired through knowledge of the population genetics of classic model organisms such as eukaryotic multicellular plants and animals. Most eukaryotes are constrained by the forced homologous pairing of the meiotic chromosomes and the requirement that genetic changes appear in the germ line in order to be inherited. This restricts the acquisition of new features by import of radically new genes that would have to be integrated into two homologous chromosomes to be properly maintained through the sexual cycles. Indeed, eukaryotic genomes are extremely conserved in gene content and widely divergent species barely differ in gene repertoire. The prokaryotic cell, by contrast, is simple and does not require zygote formation in the exchange of genetic material. The only limitation to the acquisition of new genes lies in the total genome size, which has to be maintained within the limits manageable in a small cell volume. In addition, it is equally simple for prokaryotes to release genetic ballast when the environment allows, thus they are able to dispense with non-essential genes. But the apparent simplicity of prokaryotic cells is deceiving. As the nascent field of population genomics has developed, it has become clear that the gene pool of prokaryotic species is extremely large. Different lineages of bacteria contain different genomes (similarly to the way different tissues have different proteomes in a multicellular eukaryote), increasing enormously the metabolic and ecological capabilities of one bacterial species. As this species gene pool, or "pan-genome," can be very large, it is necessary to investigate patterns and to develop models if we are to expand our understanding of prokaryotes and, from our own human perspective, the clinical or biotechnological avenues that it opens. In this review, we describe the different approaches to study the pan-genome and discuss the impact that the pan-genome concept has on our understanding of bacterial evolution and population genomics, as well as its implications for more applied issues, such as vaccine design.



Core and adaptative genome

Once the genomes of different strains of *E. coli* and other well-known bacteria became available, it was observed that certain chromosomal regions were shared among them whereas other sections appeared to be highly variable. The shared, "core" genome accounted for around 40% of the total gene pool in *E. coli* and it is interrupted by multiple variable regions unique to individual (or a few) strains (Fig. 1). This clearly shows that the genomes of multiple, independent isolates are required to understand the global complexity of a bacterial species. Methods to evaluate species genetic diversity, such as complete genome hybridization or MLST, can explain only the presence, absence, and variability of the genetic loci that are already known, but they do not provide information on genes that are not present in the reference genome. The concept of bacterial species has dramatically changed during the last years [27] and bacterial species are more appropriately described nowadays by their pan-

Fig. 1. Core and accessory genones of *Escherichia coli* and *Staphylococcus aureus*. Each circle represents the genome of a given strain, and the color scale indicates the number of orthologs found for each gene across the sequenced strains. The circles correspond to the genomes of (from outside to inside): *E. coli* (536, APEC O1, CFT073, K12 MG1655, O157 H7 EDL933, UTI89, W3110, O157 H7 Sakai); *S. pyogenes* (M1GAS, MGAS10750, MGAS2096, MGAS10270, MGAS6180, MGAS5005, MGAS10394, SSI-1). To detect orthologs, all the coding sequences from each species were joined in one unique multi-FASTA file and the BLASTCLUST software used to group together genes sharing at least 70% similarity over 70% of length coverage. The GenomViz software was used to display the genomes.

genomes, which includes a core genome containing genes present in all strains and an accessory genome consisting of partially shared and strain-specific genes. If a bacterial species is more than a semantic term and has in fact a biological meaning, the core (or "backbone") genome is the essence of this phylogenetic unit and is thought to be representative at various taxonomic levels [43]. The accessory (or adaptive) genome, on the other hand, includes key genes to survive in a specific environment; it is commonly linked to virulence, capsular serotype, adaptation, and antibiotic resistance and might reflect the organisms´ predominant lifestyle [48].

What set of genes form the core and accessory genomes? Based on a strict classification of orthologous groups (70% similarity and 70% overlapping length) along eight genomes of *E. coli* (the commensal, laboratory K12 strain, the human uropathogenic strain CFT073, the two enterohemorrhagic strains EDL933 and SAKAI, and two diarrhea-associated *Shigella* 2a strains) and nine genomes from the opportunistic pathogen *Staphylococcus aureus* (methicillin-resistant strains, USA300 and MW2, subsp. *aureus* strain N315 and NCTC 8325, and the hypervirulent subsp. *aureus* strain 476), we carried out a functional classification of the backbone present in most of the strains, and of the accessory genes present in just one or two strains (Fig. 2).

At first examination it is possible to observe that the genomes of *E. coli* are more functionally heterogeneous than those of *S. aureus*, due to the higher presence of unique genes in more functional categories (Fig. 2, yellow-orange bars). In both cases, the majority of genes belonging to the accessory and core genomes fall into the "poorly characterized or unknown" group (Fig. 2, right). Nonetheless, some patterns are evident. The fraction of genes belonging to the translation informational category, for instance, is enriched in the core genome. The majority of genes belonging to the core group are related to housekeeping functions, the cell envelope, regulatory roles, and transport and binding proteins, whereas a gene fraction that appears enlarged in the adaptive genome corresponds to defense mechanisms. It may seem puzzling that genes involved in DNA replication are over-represented in strain-specific genes. However, this is due to the inclusion of transposases and other mobile elements within this category (Fig. 2B). The adaptive or accessory genome frequently represents a surprisingly large proportion of the total gene repertoire within a species [54]. To a large extent it is formed by hypothetical genes or genes of unknown function, as well as genes associated with mobile and extra-chromosomal elements, supporting the hypothesis that the majority of specific traits depend on lateral gene transfer events [35,53]. The core genome of *Legionella pneumophila* hosts all features necessary to infect, survive, and replicate in its natural hosts, amoebae and protozoa, as well as in macrophages while the accessory compartment of its pan-genome contains additional virulence factors mainly related to HGT acquired islands [14]. In *S. aureus* MRSA252, many of the unique genes are predicted to have metabolic and transport functions and may therefore increase the bacterium's metabolic repertoire.

The core genome has gained the scientific community's attention for several reasons, an important one being the identification of essential genes that might be used as antibiotic targets as well as genes universally present in all pathogenic strains that could be potential vaccine targets [47]. These shared genes at a larger scale are also interesting as sequences for phylogenetic inference. A second consequence of the core-adaptive split is the identification of dispensable and non-dispensable genes. A conveniently simple model, found in nature, to identify essential genes is given by obligate-host-related bacteria (symbionts in their majority) with a minimalist genome composition: These genomes have small sizes and are non-redundant. The term "minimal genome" has been used to describe the set of genes that are thought be essential for a self-sustainable cell and several proposals of this minimal gene set have been put forward [39].

Calm and choppy regions in a genomic sea

In the 1980s, the concept of adaptive islands was advanced, initially as pathogenicity islands because they were responsible for many of the virulence factors of specific virotypes [28], particularly in *E. coli*, a species extremely diverse regarding its pathogenic arsenal. Later on, the term was expanded to adaptive islands, reflecting a wide range of metabolic or ecological activities. In any case, the physical clustering of adaptive genes is quite general if not universal (see, for example, Fig. 1). Very often the general properties of the genome are obviously altered, starting with the G+C content and often extending to other, more subtle parameters such as codon usage, coding density, and GC skew. This is often interpreted as an indication of evolution in a different genomic environment or xenologous origins, but it might simply reflect the effect of relaxation of purifying selection on genes that are not involved in the optimization of cell processes but rather in increasing the range of environmental responses.

Phages are probably the main way to mobilize genes in or around adaptive islands, but there could be other, less obvious ways involving recombination mediated by homologous or transposable elements. Examination of the G+C content of prokaryotic genomes shows that they often contain "calm"

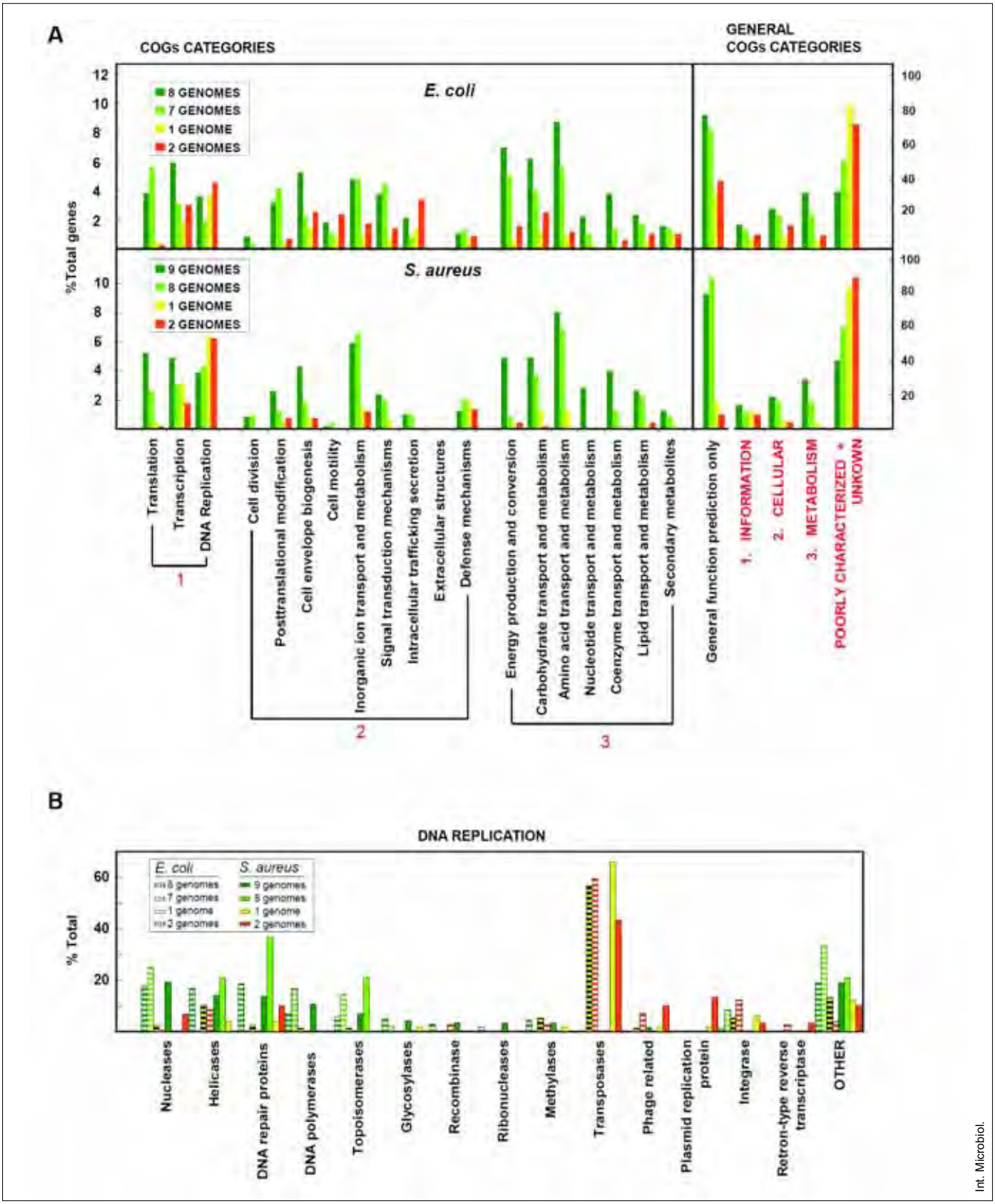


Fig. 2. Distribution of clusters of orthologous groups (COGs) among different sequenced genomes of *Escherichia coli* and *Staphylococcus aureus*. Genes present in one or two strains can be considered part of the accessory gene pool, whereas genes present in seven, eight or nine strains are members of the species backbone. The lower graph shows the detailed distribution of orthologs within the DNA replication category.

(core) and "choppy" (adaptive) regions (Fig. 5). Typically there are at least two or more such regions located at regular intervals along the replicon. Also, there is often some degree of symmetry in circular chromosomes, in which choppy regions are located symmetrically with respect to the origin of replication [30]. In *Strepcococcus agalactiae* [54], most of the strain-specific genes are in genomic islands. They are often flanked by insertion elements and display an atypical nucleotide composition, suggesting that their acquisition occurred through horizontal transfer. The specific location of choppy regions could be related to structural constraints of the chromosome and to the availability of different chromosomal sections to foreign elements.

Open and closed genomes

Comparative genomic explorations lend insight into the metabolic and ecological diversity of microbial taxa and provide us with the raw materials for inferring evolutionary history among lineages. The existence of open and closed genomes can be now performed in many genera as the result of a burgeoning increase in microbial genome sequences from different strains within the same species. A recent study of the genes shared between the 26 sequenced species within the genus *Streptococcus* [29], which contains between 1697 and 2376 coding genes, has indicated that the core genome reaches a plateau around 600 genes, whereas the *Streptococcus* pan-genome probably surpasses 6000 genes. This would exceed by at least three-fold the average genome size of a typical *Streptococcus* species. In a previous study with several strains of *Streptococcus agalactidae*, mathematical extrapolation of the data suggested that the available gene reservoir in the *S. agalactidae* pan-genome is vast and that unique genes will always continue to be identified even after hundreds of genomes have been sequenced [54]. These species in which new strains always appear to provide novel genes are said to be "open," in the sense that their pan-genome is, theoretically, infinite. Although this infinite gene pool is obviously a mere mathematical extrapolation from the available sequenced strains, it nevertheless makes clear the fact that some species exhibit extreme versatility in gene content. An unresolved question is therefore how widespread open genomes are in nature and whether there are also "closed" genomes, in which case multiple strains would not significantly vary the size of the pan-genome.

Rarefaction curves are a useful tool normally used by ecologists to graphically determine when further sampling would not increase the number of newly identified species. Using the same principle for estimating the pan-genome size after systematic sequencing, we can quantify the increase of novel genes with the availability of newly sequenced strains [14]. When such curves are plotted for several species, a gradient can be observed ranging from extremely open to more closed pan-genomes. In the latter, the gene pool is no longer expanded after two or three sequenced strains, as is clearly

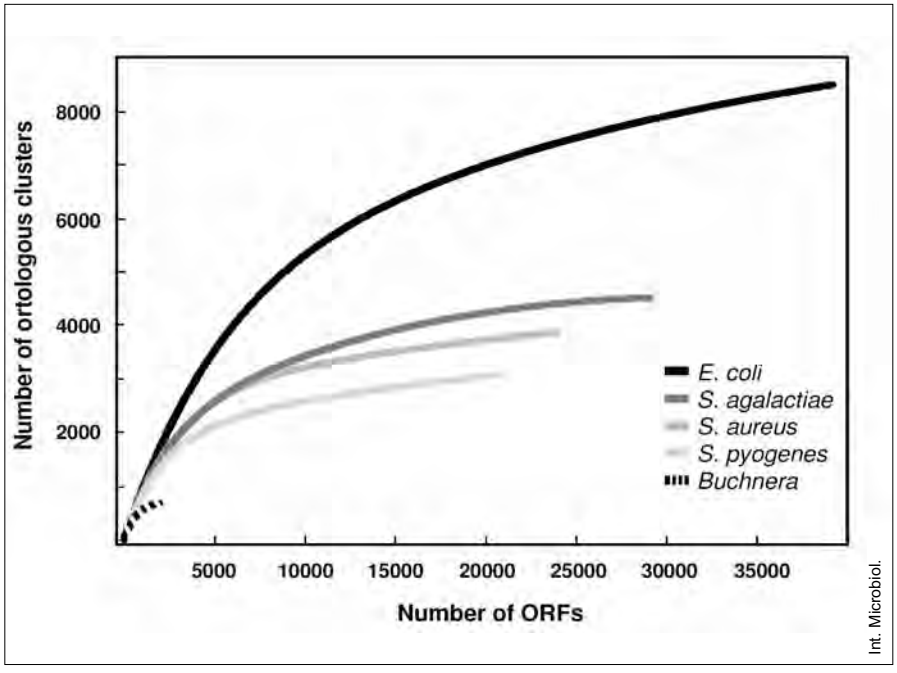


Fig. 3. Rarefaction curves for the pan-genomes of *Escherichia coli* (8 sequenced genomes), *Staphylococcus aureus* (9 genomes), *Buchnera aphidicola* (4 genomes), *Streptococcus agalactidae* and *Streptococcus pyogenes* (9 genomes). The plots indicate the novel genes obtained when sequencing additional strains. A gene is considered novel if it shares >70% protein similarity over 70% of the length (except for the highly divergent *Buchnera* strains, where 40% similarity was used). All species shown here, except *Buchnera*, have open pan-genomes with non-asymptotic curves.

visible in the intracellular symbiont *Buchnera aphidicola* (Fig. 3). Its four sequenced strains may have a closed pan-genome because they occupy an isolated and restricted niche that would hamper the capacity to acquire foreign genes, in addition to the lack of mechanisms for gene exchange and recombination. Another, non-intracellular example is given by *B. anthracis*, which can be fully described by four genomes [34]. In this case, *B. anthracis* could be considered not as a true genetic species on its own, but merely as a clone of *Bacillus cereus* with very distinctive phenotypic traits provided by the acquisition of a virulence plasmid coding for the anthrax toxin. It must be nevertheless kept in mind that a closed genome does not necessarily imply that all strains show an identical phenotype, because different nucleotide polymorphisms could confer unique features. It has recently been shown, for example, that a single nucleotide mutation in a promoter region of some *Buchnera* strains alters the thermal tolerance of both the bacteria and their insect hosts [18] and that a single amino-acid substitution in the NDK gene of some halobacteria confers different salt tolerances [38], confirming that strain flexibility can also be achieved by sequence polymorphism.

On the opposite extreme from closed pan-genomes, the model organism *E. coli* is an example of a well-studied open pan-genome. Based on the gene content of the initially sequenced eight strains, its core genome was predicted to be formed by about 2800 genes, implying that hundreds of genes can be unique to different *E. coli* strains. The number of shared and specific genes that would be found by additional genome sequences for *E. coli* genomes has been determined in rarefaction curves [27], and the genetic novelty contributed by each new strain predicted to be over 300 genes. This is well above the 27 and 33 unique genes obtained, respectively, from each new strain of group A and B *Streptococcus*. Thus, open pan-genomes are more pervasive than previously anticipated, as shown also by genomic analysis of 19 isolates of *Salmonella typhi* [25]. Given that most of these studies were done with strains isolated from different parts of the world, it was still possible that the pan-genome structure did not apply to local, co-inhabiting strains. However, recent work showed that two strains of *Salinibacter ruber* isolated from the same liter of water differed by more than 10% in gene content [45], demonstrating that local populations of a given species were not clonal.

The structural pan-genome

The existence of a pan-genome is not merely restricted to gene content but it is likely to extend to structural features such as variations that arise as a consequence of genomic rearrangements. Intraspecific differences in genome architecture may influence the bacterial phenotype even in the presence of the same gene repertoire. The reason for this is that gene location affects important physiological processes such as protein dosage or expression level [42,50]. Furthermore, changing the location of mobile elements may put other genes in contact with regulatory regions and activate gene expression. As a consequence of these emerging properties, genomic rearrangements may affect cell fitness, and different genomic arrangements probably have a biological meaning. Thus, the existence of a structural pan-genome is not inconsequential, as different genomic architecture variants can influence important aspects, such as growth rate or strain pathogenicity [50].

Mechanistically, genomic rearrangements can be caused by illegitimate recombination between repeats. Thus, repeated sequences within a genome, e.g., rRNA operons, IS elements, and other repetitive elements, generate apparent genome flexibility and genomes with high numbers of repeats are more shuffled than those with a low repeat number [35]. Other manifestations of a structural pan-genome are given by transient duplications. The replicated region can be advantageous for the cell under certain circumstances, as it increases the dosage of the duplicated genes. This phenomenon, known as transient amplification, has been shown to operate when bacteria encounter toxic substances or unusual concentrations of a nutrient [52]. It has been shown experimentally that if the selective pressure favoring an increase in gene dose is removed, the duplication quickly reverts to its initial, single-dose situation [49]. This phenomenon evolves extremely fast under laboratory conditions and it is highly likely that different isolates of the same bacterial species vary in their repertoire of duplicated regions.

Strains from multiple-replicon species are also predicted to vary with respect to gene location, although this is a little studied occurrence. The placing of a gene at a chromosome, a second chromosome, or a plasmid may have phenotypic consequences. For example, genes involved in the synthesis of essential amino acids in many variants of the symbiotic bacterium *Buchnera aphidicola* are preferentially located in plasmids, which increases their provisioning to its insect host [58], and certain functional categories, such as antibiotic resistance genes, are found at higher densities within plasmids [37]. Even the existence of a second chromosome in some bacterial species has been proposed to be an adaptation to regulate protein dosage by multiple replication forks in different replicons [9]. Analysis of the gene content of extrachromosomal replicons of related species shows that the same genes are present in different locations, such as in plasmids in one species and integrated within the main chromosome in another.

Can the variability of the bacterial pan-genome with respect to its genomic architecture be measured? Initial approaches used PCR amplification of regions that included repetitive elements, inferring rearrangements in different *Salmonella* serovars from the size of the bands obtained [1]. A traditional approach has involved the use of restriction enzymes in different natural isolates from the same species. Incongruent patterns in the length of the obtained DNA fragments, visualized in pulse-field electrophoresis gels, are indicative of genomic rearrangements. Using this approach, a systematic study of over 100 strains of *S. typhi* showed that genomic rearrangements were responsible for producing ribotype heterogeneity in this species [40]. The use of rare-cutting endonuclease analysis and PFGE revealed diverse genomic changes (including translocations, inversions, duplications, and point mutations) among isolates of *S. typhimurium*, even in archival collections [33]. The systematic sequencing of multiple strains provides an alternative approach to directly screen for genomic rearrangements. When whole-genome sequences are available, gene position plots between two strains should follow a diagonal line if all genes are in the same relative position (Supplementary Fig. 1). However, X-shaped patterns are frequently observed [20], which are indicative of genomic inversions. Most of these inversions appear to be symmetric with respect to the replication origin or terminus, suggesting that replication forks are hot-spots for recombination and that these inversions pivoting around *ori* and *ter* sites are favored by selection. Nevertheless, changes in gene position and orientation are also observed [12].

Another way of observing genomic inversions and translocations comes from looking at GC-skew plots from fully sequenced genomes. Due to mutational bias between leading and lagging strands, the DNA sequence shows an excess of G over C that reverts at the end of a replichore. If a recent genomic rearrangement changes the strand where a gene is located, the GC-skew undergoes a drastic change that can be easily visualized. Clear examples are observed in the genomes of *Yersinia pestis*, *Leptospira interrogans*, and other species with several sequenced strains [see, for example, http://insilico.ehu.es/oligoweb]. Finally, microarray-based comparative genome hybridizations among multiple strains frequently uncover tandem duplications when the intensity of a DNA segment is increased on the array relative to the control, as shown in the intracellular, small-genomed *Bartonella henselae* [31]. In conclusion, the observed genomic flexibility in the structural pan-genome is surprisingly high, even in intracellular species traditionally assumed to be fairly stable and these variations confer upon prokaryotic species an important degree of plasticity.

Mechanisms of genome expansion

What is the origin of the accessory component of the pan-genome? How can new functional capabilities be generated? One of the prominent advances of the genomics era has been not only to identify but also to quantify the importance of the different sources of new genes. A large proportion of novel genes arise by DNA duplication, followed by sequence divergence. In prokaryotes, the proportion of these paralogous genes can be very important, reaching up to 50% of the DNA in versatile species with large genomes [46]. In the last decade, a second mechanism has been supported by solid evidence of its overwhelming impact on bacterial evolution: lateral transfer of DNA sequences between different bacterial cells. This process has been shown both experimentally and by sequence analysis to occur frequently between different species (see, for example, [42]).

The existence of horizontal gene transfer (HGT) in prokaryotes was, of course, known for many years before the completion of bacterial genomic sequences. The transmission of antibiotic resistance among clinical isolates, for example, has long been identified as the outcome of plasmid transmission, and many bacterial species are naturally competent and can incorporate DNA from the environment. Apart from conjugation and transformation, a third route of HGT is achieved when bacterial DNA is packaged into bacteriophage capsids and transferred by the virus during the infection of a new host cell. Through this transduction process, bacteriophages are considered a major cause of HGT, especially regarding virulence factors and invasion-related functions [6], and fundamental for gene innovation in the microbial world [13]. Finally, although it may not seem intuitive, it has been shown that not only gene expansions but also deletions can change bacterial phenotypes. A clear case has been shown in *E. coli*, in which the elimination of certain genes activates other pathways that confer virulent capabilities. Experimental insertion of the gene restores the non-virulent phenotype. Thus, the different abilities of strains within a population must be seen not only as the consequence of novel functions coded by additional genes but also as the outcome of a regulatory network that can be altered by a smaller gene set and by different gene interactions (see, for example, [7]).

The contribution of horizontally transferred genes to the non-core, accessory genome fraction is vast, as indicated by the high proportion of mobile elements, phage-related genes, and pathogenicity islands in this section of the pan-genome. The term “accessory” genome is nevertheless misleading.

Whole-genome sequences can now be screened for unusual compositional features or phylogenetic incongruence, and different methods point to the major role played by exogenous genes in the lifestyle of bacteria. For example, this accessory (or rather, adaptive) pool includes, apart from the well-studied antibiotic resistance genes, others coding for bacteriocins, heavy-metal resistance, cell-wall components, nitrogen fixation, virulence, and many others, including metabolic genes [23], and HGT events have repeatedly induced a change in lifestyle for the recipient genome.

The presence within the accessory pan-genome of an intriguing group of genes unrelated to any other genes in databases deserves special mention. These "orphan genes" do not show any detectable similarity to other available sequences, including those of related species and strains. They are overrepresented in genomic islands [26] and can be an important fraction of the accessory component: in a microarray-based analysis of 15 clinical isolates of *Helicobacter pylori*, for example, up to 56% of the strain-specific genes were orphans and even higher values were obtained when comparing different sequenced *Rickettsia* genomes [44].

Orphan genes have drawn considerable attention because they could be involved in functions conferring a given species its host specificity or biological individuality. However, they are significantly shorter than their better-characterized counterparts [35] and many have an unusually high number of non-synonymous substitutions [41], suggesting that they could be non-functional. In addition, when examined in detail, some of the orphan genes in *Rickettsia* appeared to be fragments of longer, functional genes in related species [15]. This suggested that genomes are generally over-annotated. However, the analysis of lineage-specific genes in γ-proteobacteria revealed that they were not only shorter but also AT-rich and fast-evolving [13]. Furthermore, their nucleotide substitution patterns suggested that most were functional and they frequently formed part of clusters that contained foreign sequences. This has led to the suggestion that orphan genes are sequences that come from bacteriophages, many of which are retained if they confer a useful new function for the recipient cell. It is therefore probable that the small genomes of intracellular bacteria such as *Rickettsia* have many non-functional orphans as a consequence of an ongoing process of genome reduction and pseudogenization. In larger-size genomes of free-living organisms, by contrast, many orphan genes are probably functional and their lack of resemblance to any available sequences is a consequence of the little-surveyed viral genomes in current databases.

Approaches to study the microbial pan-genome

Multiple strain sequencing. An obvious approach to define and study a prokaryotic pan-genome is the sequencing of multiple strains from a given species. Methodologically, the existence of a sequenced bacterium facilitates the otherwise tedious assembly of subsequent relatives. The spectacular genomic differences reported between a saprophytic, a uropathogenic, and an enterohemorrhagic strain of *E. coli* [57] has encouraged successive sequencing of other *E. coli* strains, and similar approaches have been adopted in other bacterial species. Multiple strain sequencing was initially conducted by different research laboratories, but in the last few years single research projects are routinely sequencing multiple strains [54]. The time-consuming process of gap-closure can be obviated by assuming synteny among the strains. This, of course, constrains the analysis of the structural pan-genome but allows a complete scrutiny of differences in gene content and polymorphisms, as the gap regions are normally very short and typically over 98% of the genomic DNA is available with standard sequencing coverage levels. The new high-throughput sequencing techniques allow multiple-strain genome completion with an extraordinarily reduced budget and time schedules. Combined strategies using 454-pyrosequencing, Solexa, and traditional Sanger technology have proved to be very efficient for systematic sequencing of different microbes [24]. Thus, the next few years will surely provide microbiologists with a myriad of genomic variants for many microbes, and a new period in pan-genomic science will emerge in which bioinformatic analysis of these large datasets will be crucial.

Hybridization-based techniques. The evolution of bacterial strains under different ecological circumstances and their adaptation to specific niches should be reflected at the genomic level in terms of gene content. Microarray techniques have become a powerful approach to study these processes. If the genome of a species representative is sequenced, all ORFs are typically spotted on a microarray slide, serving as a probe against which a strain of unknown gene content is tested. Polymorphisms for gene deletions and insertions can be detected as a change in the ratio of fluorescence emitted when the two labeled genomic DNA samples are hybridized to the microarray [21]. DNA/DNA hybridization techniques allow the identification of genomic differences even between bacterial species [17,32]. Comparisons between strains within the same or related species, however,

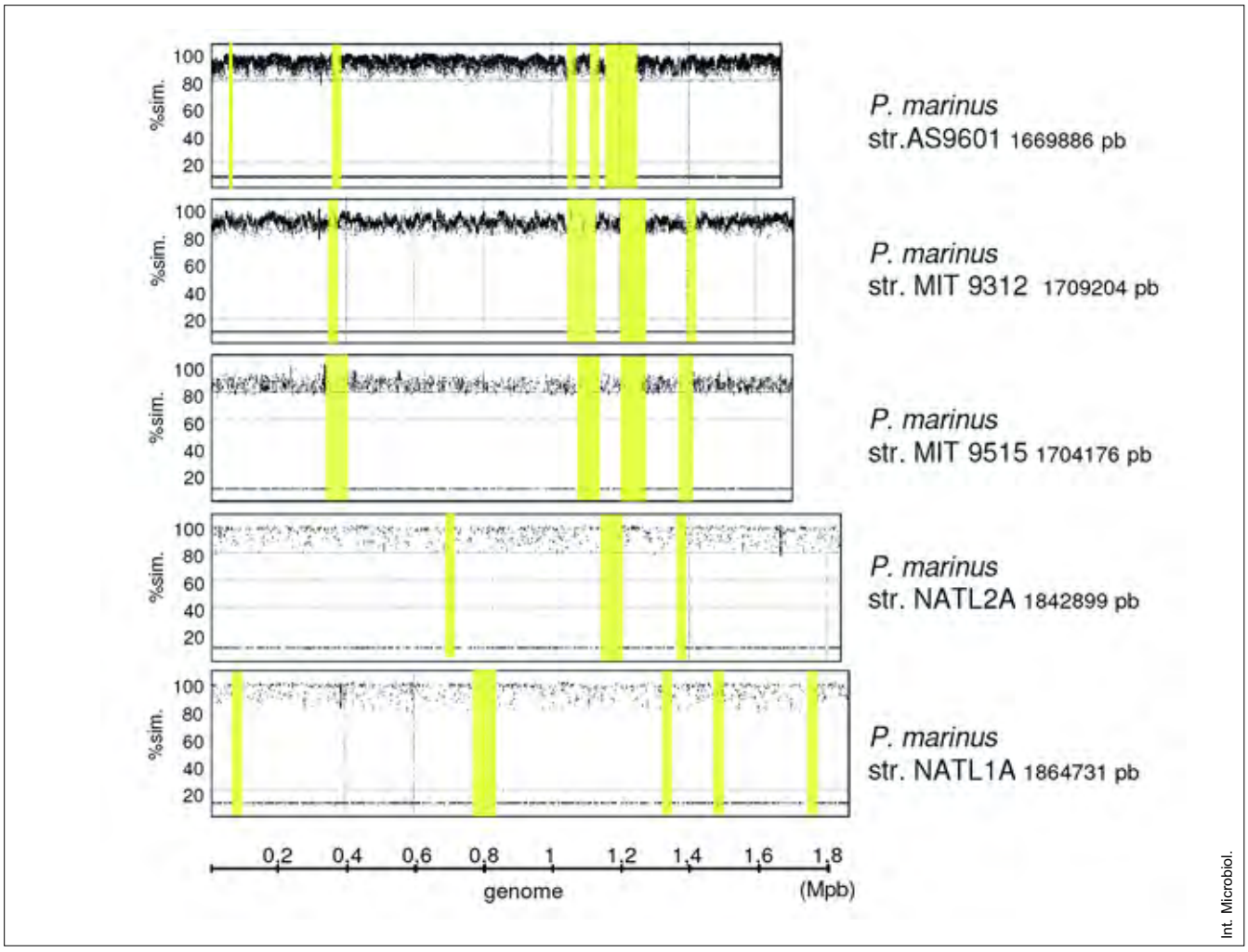


Fig. 4. Alignment of several *Prochlorococcus* genomes with environmental sequences from the Atlantic Ocean. Values on the y-axis indicate nucleotide percent identity between a given *Prochlorococcus* genomic region and the metagenomic sequences from the Sargasso Sea. Areas with unusually low representation in the metagenome are highlighted in yellow and described in the text as genomic islands. Whole-genome alignments were performed with NUCmer, from the MUMmer 3.10 package.

are more consistent and have been shown to be very informative, relating hybridization efficiency and therefore gene composition differences to virulence, host inflammatory responses, relatedness among genes within operons, endemic and pandemic disease isolates, or the symptoms of plants under microbial infection. Thus, the systematic use of comparative genome hybridization (CGH) by microarrays, also called genomotyping, is an effective tool to study intraspecific differences in gene content in bacteria. One of the most important limitations of this technology is that the presence or absence of the sequences is established always in relation to the reference genome and therefore unique genes on the tested strains cannot be identified. This problem can be circumvented by suppression subtractive hybridization (SSH), in which DNA from a test strain is depleted by hybridization to sequences from a reference strain. The remaining DNA is enriched in test-strain-specific genes, which are then cloned and/or sequenced [61]. While this technique has been used to a lesser extent by the scientific community, both genomotyping methods are very helpful approaches to determine the scope of the genomic backbone and to establish the functional characteristics of accessory genes.

Microarray techniques have become a powerful and fast method to quantify genomic variability among strains, and an increasing number of species from different taxonomic groups and lifestyles has been tested (Supplementary Table 1). This has shown that the percentage of the genome containing conserved ORFs (i.e., the core genes) is variable among species. Although part of this variation is probably due to the number of strains studied, there is also most likely an effect of

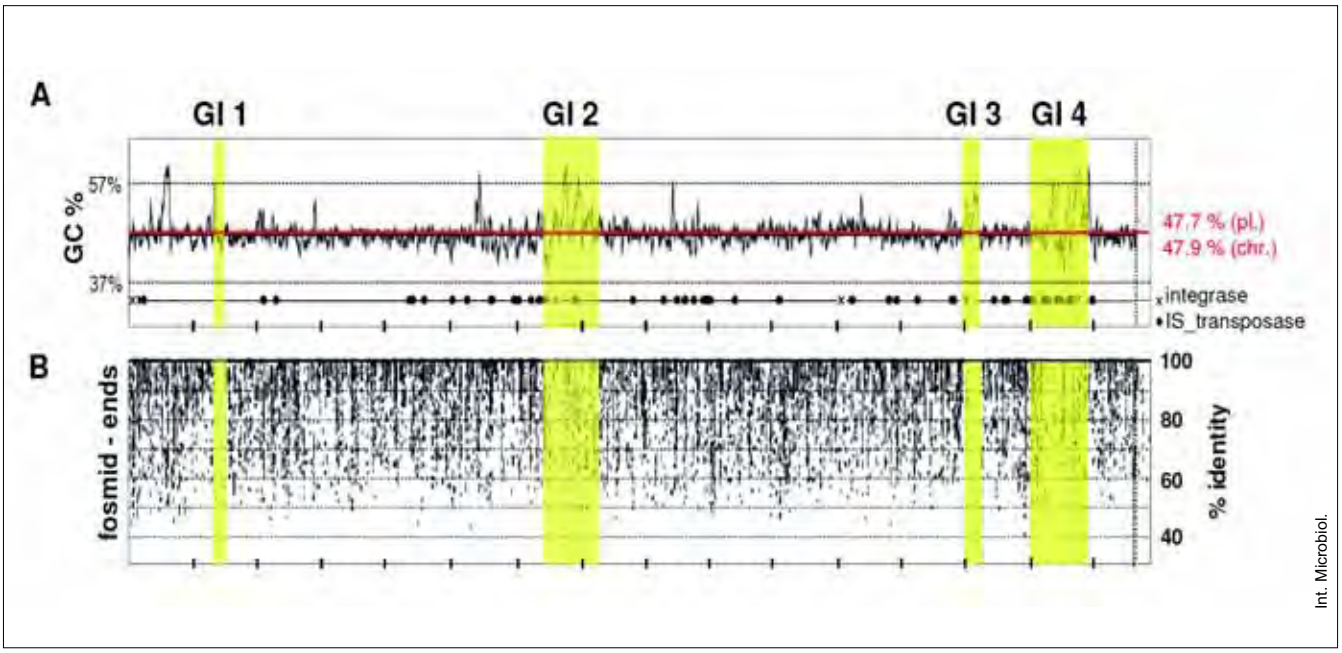


Fig. 5. (**A**) Genomic islands (GI) of *Haloquadratum walsbyi* DSM 16790, showing "choppy" and "calm" regions. The GC-content of *H. walsbyi* is plotted with a sliding window of 1000 bp. Location of integrases and IS transposases along the genome are indicated. (**B**) Fosmid-end coverage. Individual fosmid end sequences were aligned to the sequenced strain genome and the alignments' sequence conservation visualized in the form of percent identity plot. Each dot on the graph represents the amino acid similarity between fosmid-end sequences and their homologous regions in the *H. walsbyi* genome. Regions of unusually low representation in the metagenome are shaded.

lifestyle. For example, *Salmonella enterica* strains appear to be highly variable [11], in agreement with the different host niches for different serovars and the dramatic input of new sequences by lateral gene transfer. Other labile species with large genomes, such as *Bacillus*, also appear to have a large fraction of dispensable genes. *Bartonella henselae* strains would represent the opposite extreme; its only slight variability is consistent with its more enclosed lifestyle and static genome [31]. Interestingly, free-living species of similar or larger genome size than *B. henselae*, such as *Campylobacter jejuni* and *Vibrio cholera*, display low variation. Thus, intracellular niche does not seem to be the only reason for lack of variation; rather, the frequency of mobile and repetitive elements may be an important factor [36]. Some similarity can be found among species in the function of the missing, variable ORFs. Most of them appear to be hypothetical proteins or ORFs with unknown function, as also observed in multiple-strain comparisons of fully sequenced genomes. Other categories in which variability is frequently high are those related to pathogenicity, e.g., virulence genes, membrane proteins, or phage sequences.

Multiple genomes in a metagenomic survey. An alternative approach for studying the pan-genome is to use metagenomic sequence data obtained directly from an environment in which a species is well represented. Metagenomics enables us to study microorganisms by deciphering their genetic information from DNA that is extracted directly from environmental samples, thus sidestepping the need for culturing. From the sequences, it is possible to identify the total genetic repertoire of the species forming the microbial communities under study.

The approach was undertaken by Craig Venter to directly clone and sequence DNA from the microbes in open ocean waters [56], and it has demonstrated that the genetic diversity of the microbiota in natural samples is even larger than expected. Nowadays, there are several hundred ongoing metagenomic projects, ranging from those involving natural marine communities to those focused on the human microbiome. One of the strategies to localize variable regions in a genome is to align it with the thousands of short sequences normally derived from these metagenomic studies, inferring the presence of this species by the number and degree of similarity of the matched sequences (Fig. 4). Furthermore, it is possible to detect different nucleotide polymorphisms, small insertions, deletions, and hypervariable segments.

As a practical example, we focus on the case of the species *Prochlorococcus marinus* from the Sargasso Sea metagenome survey [56]. Prochlorococci are globally abundant and have compact genomes with sizes between 1.7 and

2.4 Mbp. Their populations consist of multiple coexisting ecotypes whose relative abundances vary markedly along gradients of light, temperature, and nutrients. Coleman and collaborators [8] aligned the thousands of sequences from the metagenome against each of the 11 completely sequenced strains, observing "metagenomic islands" (MGI) characterized by the absence of matches to the metagenome. A comparison between strains MIT 9312 (North Atlantic) and MED4 (Mediterranean) among themselves and against the Sargasso Sea metagenome revealed that some of these islands were located at the same position in the two genomes. This supports the idea that special areas function are hotspots for pan-genomic variability. In this specific case, the islands arose partly by phage-mediated HGT and it was possible to detect associations with tRNA genes (which are common integration sites for mobile elements). Analysis of the islands' genetic content identified genes similar to those of non-cyanobacterial organisms or with no detectable homologous genes in databases, reflecting a large, as yet unknown gene pool in oceans. In addition, MGIs contain phage-like sequences (integrases, DNA methylases, endonucleases, etc.) and, interestingly, proteins involved in cell surface modification, including the biosynthesis of lipopolysaccharide, which is a common phage receptor. These data underscore that phages are important agents of mortality in the oceans [51]. Other functions were related to physiological stress and nutrient uptake, characters that are a direct response to the environment. Thus, a gene within the accessory genome not present in all sequenced strains should not be literally regarded as dispensable.

Single-site species metagenomics. The case of *Haloquadratum walsbyi*.

The ideal situation for metagenomic analysis of a species pan-genome is the availability of a sequenced genome together with an environmental population of the same species obtained from the same site. Such data would enable evaluation of how representative an individual is within the population and would reveal the form and distribution of genetic variability in the core and the accessory gene pool. This perfect-case approach has been carried out in extremely simplified habitats such as the low-diversity saturated brines of solar salterns, where the square-shaped haloarchaea *Haloquadratum walsbyi* thrives [4, 5,30], or the acidic drainage of a mine where *Ferroplasma acidarmanus* is a major component [2,19].

In the first study, a crucial advantage was the possibility to enrich the recovered environmental sample in cells of *Haloquadratum* due to the particular shape and size of this postal-stamp-looking archaeon. Additionally, the previously sequenced *Haloquadratum walsbyi* DCM 16790, isolated from the very same site, showed a low G+C content, which is unusual in this environment and made it possible to easily discern which sequences from the total data set belong to *Haloquadratum*. The construction of a metagenomic library and the sequencing of a relatively small amount of DNA (< 3 Mpb) revealed a remarkable diversity of genes, showed evidence for MGIs, and established that the *H. walsbyi* total gene pool was, even in that relatively simple and constant environment, at least twice the genome size of the sequenced strain [30]. Thus, a picture of high genetic diversity was found, in contrast with other simplified extreme environments analyzed by similar approaches [55]. To complement this work, Cuadros-Orellana and collaborators completed the sequencing of several fosmids from the same metagenomic library [10]. Two of the metagenomic islands (MGI 2 and MGI 4) showed the usual hallmarks of these regions, i.e., atypical G+C content and a rich complement of mobile elements. The first island, as in the previous case of *P. marinus*, contains genes required to synthesize the rigid components of the cell envelope. Some of the cell-surface glycoproteins encoded within this island seem to be paralogous and might be involved in intragenomic recombination processes that would produce large cell-surface variability. The authors of the study hypothesized that this promotes variation at phage attachment sites, which could be a useful strategy for phage evasion by "competitive dominants" microdiversity [51]. The polysaccharide-related genes found in MGI 4 might contribute to a similar purpose. A large variability of new transporters and catabolic genes was also found, suggesting that different cells or lineages within *H. walsbyi* specialize in the exploitation of different organic compounds and coexist in what is a chemically diverse set of organic carbon sources. They do so by containing different gene pools that are largely associated with the MGIs. If this were indeed true, individual cells would be specialized to a specific set of nutritional requirements but would not directly compete with other members of the same species for the same resources. This is in striking contrast to the way species are typically conceived, in which intraspecies competition is for the same resources.

Future perspectives in a pan-genomic world

The observed genomic variability between multiple strains of the same species clearly demonstrates that the genome sequence of a bacterial isolate underestimates the biological properties of a species. The pan-genomic concept therefore corroborates the well-documented plasticity of prokaryotes

from the very core of these organisms: their genome sequence. This has important theoretical and applied consequences. Firstly, the large variations in gene content between bacterial strains challenge the value of a eukaryotic-based species concept for the microbial world [59]. Genomic data from multiple strains allow us to question whether species names are merely a semantic issue or whether they represent real biological entities with distinctive features. In this sense, the asymptotic 40% threshold of shared genes among *E. coli* strains might agglutinate the gene pool that gives this species its phylogenetic and ecological coherence. This would highly contrast with the genomic situation of higher organisms, with species as distant as humans and chimpanzees sharing over 99% of their gene repertoire. If bacteria differ from each other mainly in gene content, differences between eukaryotic species arise primarily from the regulation, as well as the genetic polymorphisms of those pre-existing shared genes. On the more applied side, it has become clear that the sequence of a single genome does not reflect how genetic variability drives pathogenesis within a bacterial species and also limits genome-wide screens for vaccine candidates or antimicrobial targets. Genetic plasticity entails the evolution of many virulent and drug-resistant strains, presenting a major and constantly changing clinical challenge. As pointed out for *N. meningitidis* serogroup B, a successful vaccine target must be present in all relevant strains against which immunization is intended [22]. In *S. aureus*, a comparative-genomic approach has been used to explore the mechanisms of evolution of clinically important strains and to identify regions affecting virulence and drug resistance, with similar approaches being taken in other human pathogens.

In the future, the technology that will revolutionize the analysis of microbial communities could be the ability to obtain a complete genome sequence from an individual bacterial cell [60]. At present, the majority of published genome sequences represent bacteria that can be grown in culture. But as most bacteria cannot be cultured by current methods and many live in contact with other organisms and cannot be propagated in pure culture, there are still serious limitations to studying the genomes of microbial species. Single-cell genomics will presumably allow the differentiation of bacterial lineages based on their nucleotide polymorphisms and will be applied to characterize seemingly homogeneous bacterial populations, including those made up of clonal descendants [16]. This transition from species and strains genomics to single-cell genomics will doubtlessly be accompanied by further decreases in the cost and time of DNA sequencing as well as by the need to implement computer software able to cope with the colossal amount of data thus produced. Even today, the pace at which genomic data accumulates is far greater than our capacity to analyze it, and the situation is practically guaranteed to become even more extreme in the near future, showing once again that microbes surpass our capability to explore and understand the natural world.

Acknowledgements. This article is dedicated to the memory of Dr Hillevi Lindroos. The work was supported by Consolider Grant CSD2009-00006 from the Spanish Ministry of Science and Innovation (MICINN). G.D. is funded by Project CP09/00049 "Miguel Servet" of ISCIII and A.B.M-C was supported by a Juan de la Cierva scholarship, both from the MICINN.

References

1. Alokam SS, Liu L, Said K, Sanderson KE (2002) Inversions over the terminus region in *Salmonella* and *Escherichia coli*: IS*200*s as the sites of homologous recombination inverting the chromosome of *Salmonella enterica* serovar typhi. J Bacteriol 184:6190-6197
2. Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, Banfield JF (2007) Genome dynamics in a natural archaeal population. Proc Natl Acad Sci USA 104:1883-1888
3. Bergthorsson U, Ochman H (1998) Distribution of chromosome length variation in natural isolates of *Escherichia coli*. Mol Biol Evol 15:6-16
4. Bolhuis HE, Poele M, Rodriguez-Valera F (2004) Isolation and cultivation of Walsby's square archaeon. Environ Microbiol 6:1287-1291
5. Bolhuis H, Palm P, Wende A, Falb M, Rampp M, Rodriguez-Valera F, Pfeiffer F, Oesterhelt D (2006) The genome of the square archaeon *Haloquadratum walsbyi*: life at the limits of water activity. BMC Genomics 7:169
6. Boyd EF, Brussow H (2002) Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. Trends Microbiol 10:521-529
7. Branger C, Zamfir O, Geoffroy S, Laurans G, Arlet G, Thien HV, Gouriou S, Picard B, Denamur E (2005) Genetic background of *Escherichia coli* and extended-spectrum beta-lactamase type. Emerg Infect Dis 11:54-61
8. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, Chisholm SW (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311:1768-1770
9. Couturier E, Rocha EP (2006) Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol 59:1506-1518
10. Cuadros-Orellana S, Martin-Cuadrado AB, Legault B, D'Auria G, Zhaxybayeva O, Papke RT, Rodriguez-Valera F (2007) Genomic plasticity in prokaryotes: the case of the square haloarchaeon. ISME J 1:235-245
11. Chan K, Baker S, Kim CC, Detweiler CS, Dougan G, Falkow S (2003) Genomic comparison of *Salmonella enterica* serovars and *Salmonella bongori* by use of an *S. enterica* serovar typhimurium DNA microarray. J Bacteriol 185:553-563
12. Dalevi DA, Eriksen N, Eriksson K, Andersson SG (2002) Measuring genome divergence in bacteria: a case study using chlamydian data. J Mol Evol 55:24-36
13. Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. Genome Res 14:1036-1042
14. D'Auria G, Jiménez-Hernández N, Peris-Bondia F, Moya A, Latorre A (2010) *Legionella pneumophila* pangenome reveals strain-specific virulence factors. BMC Genomics 11:181
15. Davids W, Amiri H, Andersson SG (2002) Small RNAs in *Rickettsia*: are they functional? Trends Genet 18:331-334

16. Dethlefsen L, Relman AD (2007) The importance of individuals and scale: moving towards single cell microbiology. Environ Microbiol 9:8-10
17. Dong F, Allawi HT, Anderson T, Neri BP, Lyamichev VI (2001) Secondary structure prediction and structure-specific sequence analysis of single-stranded DNA. Nucleic Acids Research 29:3248-3257
18. Dunbar HE, Wilson AC, Ferguson NR, Moran NA (2007) Aphid thermal tolerance is governed by a point mutation in bacterial symbionts. PLoS biology 5:e96
19. Edwards RA, Rodriguez-Brito B, Wegley L, et al. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7:57
20. Eisen JA (2001) Gastrogenomics. Nature 409:463-466
21. Gibson G (2002) Microarrays in ecology and evolution: a preview. Mol Ecol 11:17-24
22. Giuliani MM, Adu-Bobie J, Comanducci M et al. (2006) A universal vaccine for serogroup B meningococcus. Proc Natl Acad Sci USA 103:10834-10839
23. Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19:2226-2238
24. Goldberg SM, Johnson J, Busam D, et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci USA 103:11240-11245
25. Holt KE, Parkhill J, Mazzoni CJ, et al. (2008) High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nature Genetics 40:987-993
26. Hsiao WW, Ung K, Aeschliman D, Bryan J, Finlay BB, Brinkman FS (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. PLoS Genet 1:e62
27. Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci USA 102:2567-2572
28. Lee CA (1996) Pathogenicity islands and the evolution of bacterial pathogens. Infect Agents Dis 5:1-7
29. Lefebure T, Stanhope MJ (2007) Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome Biol 8:R71
30. Legault BA, Lopez-Lopez A, Alba-Casado JC, Doolittle WF, Bolhuis H, Rodriguez-Valera F, Papke TR (2006) Environmental genomics of *Haloquadratum walsbyi* in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. BMC Genomics 7:171
31. Lindroos H, Vinnere O, Mira A, Repsilber D, Näslund K, Andersson SG (2006) Genome rearrangements, deletions, and amplifications in the natural population of *Bartonella henselae*. J Bacteriol 188:7426-7439
32. Lindroos HL, Mira A, Repsilber D, et al. (2005) Characterization of the genome composition of *Bartonella koehlerae* by microarray comparative genomic hybridization profiling. J Bacteriol 187:6155-6165
33. Liu GR, Liu WQ, Johnston RN, Sanderson KE, Li SX, Liu SL (2006) Genome plasticity and ori-ter rebalancing in *Salmonella typhi*. Mol Biol Evol 23:365-371
34. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Op Genet Dev 15:589-594
35. Mira A, Klasson L, Andersson SG (2002) Microbial genome evolution: sources of variability. Curr Op in Microbiol 5:506-512
36. Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589-596
37. Mira A, Pushker R (2007) Evolution of genome architecture and the evolution of bacterial pathogens. In: Baquero F, et al. (eds) Introduction to evolutionary biology of bacterial and fungal pathogens. ASM Press, Washington, DC, pp 115-128
38. Mizuki T, Kamekura M, DasSarma S, Fukushima T, Usami R, Yoshida Y, Horikoshi K (2004) Ureases of extreme halophiles of the genus *Haloarcula* with a unique structure of gene cluster. Biosc Biotech Biochem 68:397-406
39. Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 93:10268-10273
40. Ng I, Liu SL, Sanderson KE (1999) Role of genomic rearrangements in producing new ribotypes of *Salmonella typhi*. J Bacteriol 181:3536-3541
41. Ochman H (2002) Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes. Trends Genet 18:335-337
42. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299-304
43. Ochman H, Lerat E, Daubin V (2005) Examining bacterial species under the specter of gene transfer and exchange. Proc Natl Acad Sci USA 102 Suppl 1:6595-6599
44. Ogata H,Audic S, Renesto-Audiffren P, et al. (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 293:2093-2098
45. Peña A, Teeling H, Huerta-Cepas J, et al. (2010) Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting *Salinibacter ruber* strains. ISME J 4:882-895
46. Pushker R, Mira A, Rodríguez-Valera F (2004) Comparative genomics of gene-family size in closely related bacteria. Genome Biol 5:R27
47. Rappuoli R (2001) Reverse vaccinology, a genome-based approach to vaccine development. Vaccine 19:2688-2691
48. Read TD, Ussery DW (2006) Opening the pan-genomics box. Curr Op Microbiol 9:496-498
49. Reams AB, Neidle EL (2003) Genome plasticity in *Acinetobacter*: new degradative capabilities acquired by the spontaneous amplification of large chromosomal segments. Mol Microb 47:1291-1304
50. Rocha EP (2004) The replication-related organization of bacterial genomes. Microbiology 150:1609-1627
51. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Paši? L, Thingstad TF, Rohwer F, Mira A (2009) Explaining microbial population genomics through phage predation. Nature Rev Microbiol 7:828-836
52. Romero D, Palacios R (1997) Gene amplification and genomic plasticity in prokaryotes. Annu Rev Genet 31:91-111
53. Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. Proc Natl Acad Sci USA 97:14668-14673
54. Tettelin H, Masignani V, Cieslewicz MJ, et al. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA 102:13950-13955
55. Tyson GW, Chapman J, Hugenholtz P, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37-43
56. Venter JC, Remington K, Heidelberg JF, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74
57. Welch RA, Burland V, Plunkett G 3rd, et al. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci USA 99:17020-17024
58. Wernegreen JJ, Moran NA (2001) Vertical transmission of biosynthetic plasmids in aphid endosymbionts (*Buchnera*). J Bacteriol 183:785-790
59. Whitaker RJ, Banfield JF (2006) Population genomics in natural microbial communities. Trends Ecol Evol 21:508-516
60. Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM (2006) Sequencing genomes from single cells by polymerase cloning. Nature Biotech 24:680-686
61. Zhang Q, Melcher U, Zhou L, Najar FZ, Roe BA, Fletcher J (2005) Genomic comparison of plant pathogenic and nonpathogenic *Serratia marcescens* strains by suppressive subtractive hybridization. Appl Environ Microbiol 71:7716-7723

INTERNATIONAL MICROBIOLOGY

MIRA M, ET AL. (2010) INT MICROBIOL 13:45-57

DOI: 10.2436/20.1501.01.110

Supplementary Table 1. Selected CGH studies in different prokaryotic species

Species	Tested Strains	Relevant features	Reference
Helicobacter pylori	15	56% of strain-specific genes are "ORFans"	[10]
Escherichia coli O157:H7	31	Even within a single serotype, 1751 ORFs were variable	[13]
Bartonella henselae	11	Genomic islands mediate genomic rearrangements	[8]
Streptococcus mutans	9	Accesory genome is 20%; half shows signs of HGT	[12]
Campylobacter jejuni	11	Largest fraction of acces. genes (19%) related to cell envelope	[4]
Salmonella enterica	25	Core genome was only 54%	[3]
Bacillus anthracis	19	Variation in strains ranges 8-34% of reference genome	[9]
Vibrio cholerae	9	Core genome was 97%	[5]
Streptococcus agalactidae	19	Extensive variation recently confirmed by sequencing	[11]
Escherichia coli - Shigella	22	*E. coli* backbone estimated at 2,800 ORFs	[6]
Streptococcus pneumoniae	20	Variability within strains < 2.1% Overall variability < 10%	[7]
Francisella tularensis	27	Regions specific to highly virulent strains were identified	[2]
Enterococcus faecalis	9	15-23% variable; transport and metabolic genes very conserved	[1]
Yersinia pestis	36	22 genomic regions absent in some of the strains	[14]

REFERENCES:

1. Aakra A, Nyquist OL, Snipen L, Reiersen TS, Nes IF (2007) Survey of genomic diversity among *Enterococcus faecalis* strains by microarray-based comparative genomic hybridization. Appl Environ Microbiol 73:2207-2217
2. Broekhuijsen M, Larsson P, Johansson A, Bystrom M, Ericsson U, Larsson E, Prior RG, Sjostedt A, Titball RW, Forsman M (2003) Genome-wide DNA microarray analysis of *Francisella tularensis* strains demonstrates extensive genetic conservation within the species but identifies regions that are unique to the highly virulent *F. tularensis* subsp. tularensis. J Clin Microbiol 41:2924-2931
3. Chan K, Baker S, Kim CC, Detweiler CS, Dougan G, Falkow S (2003) Genomic comparison of *Salmonella enterica* serovars and *Salmonella bongori* by use of an *S. enterica* serovar typhimurium DNA microarray. J Bacteriol 185:553-563
4. Dorrell N, Mangan JA, Laing KG, et al. (2001) Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. Genome Res 11:1706-1715

5. Dziejman M, Balon E, Boyd D, Fraser CM, Heidelberg JF, Mekalanos JJ (2002) Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. Proc Natl Acad Sci USA 99:1556-1561
6. Fukiya S, Mizoguchi H, Tobe T, Mori H (2004) Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray. J Bacteriol 186:3911-3921
7. Hakenbeck R, Balmelle N, Weber B, Gardes C, Keck W, de Saizieu A (2001) Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of *Streptococcus pneumoniae*. Infect Imm 69:2477-2486
8. Lindroos H, Vinnere O, Mira A, Repsilber D, Näslund K, Andersson SG (2006) Genome rearrangements, deletions, and amplifications in the natural population of *Bartonella henselae*. J Bacteriol 188:7426-7439
9. Read TD, Peterson SN, Tourasse N, et al. (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature 423:81-86
10. Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. Proc Natl Acad Sci USA 97:14668-14673
11. Tettelin H, Masignani V, Cieslewicz MJ, et al. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA 102:13950-13955
12. Waterhouse JC, Swan DC, Russell RR (2007) Comparative genome hybridization of *Streptococcus mutans* strains. Oral Microbiol Immunol 22:103-110
13. Zhang Y, Laing C, Steele M, Ziebell K, Johnson R, Benson AK, Taboada E, Gannon VP (2007) Genome evolution in major *Escherichia coli* O157:H7 lineages. BMC Genomics 8:121
14. Zhou D, Han Y, Song Y, et al. (2004) DNA microarray analysis of genome dynamics in *Yersinia pestis*: insights into bacterial genome microevolution and niche adaptation. J Bacteriol 186:5138-5146