The microbial pan-genome

Duccio Medini[1], Claudio Donati[1], Hervé Tettelin[2], Vega Masignani[1] and Rino Rappuoli[1]

A decade after the beginning of the genomic era, the question of how genomics can describe a bacterial species has not been fully addressed. Experimental data have shown that in some species new genes are discovered even after sequencing the genomes of several strains. Mathematical modeling predicts that new genes will be discovered even after sequencing hundreds of genomes per species. Therefore, a bacterial species can be described by its pan-genome, which is composed of a 'core genome' containing genes present in all strains, and a 'dispensable genome' containing genes present in two or more strains and genes unique to single strains. Given that the number of unique genes is vast, the pan-genome of a bacterial species might be orders of magnitude larger than any single genome.

Addresses
[1] Immunobiological Research Institute of Siena (IRIS), Chiron Vaccines, via Fiorentina 1, 53100 Siena, Italy
[2] Department of Microbial Genomics, The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA

Corresponding author: Rappuoli, Rino (rino_rappuoli@chiron.com)

Current Opinion in Genetics & Development 2005, **15**:589–594

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Available online 26th September 2005

0959-437X/$ – see front matter

DOI 10.1016/j.gde.2005.09.006

Introduction to the pan-genome

Ten years after the first sequence of a free-living organism was revealed, public databases contain 239 complete bacterial genomes. However, as shown in Table 1, in 83% and 8% of the cases, only one or two genomes per bacterial species have been sequenced, respectively. In a recent work [1••], eight genomes representative of the serogroup (see Glossary) diversity among group B *Streptococcus* (GBS) strains were analyzed to answer the question of how many genomes are needed to fully describe a bacterial species. Each GBS strain was found to contain an average of 1806 genes that are present in every strain (core genome [see Glossary]), plus 439 genes that are absent in one or more strains (dispensable genome [see Glossary]).

The dispensable genes are also divided into genes present in two or more but not all strains (18% of the genome) and genes unique to each strain (1.5% of the genome). Mathematical modeling based on the eight genomes showed that unique genes will continue to emerge even after hundreds or thousands of genomes are sequenced [1••]. Hence, core and dispensable genes represent the essence and the diversity of the species, respectively.

The surprising conclusion from the study is that, in theory, the bacterial species will never be fully described, because new genes will be added to the genome of the species with each new genomic sequence. Therefore, the best approximation to describe a species is to use the concept of the pan-genome ('pan' — 'παν' in Greek — means 'whole' [see Glossary]), which is made up of the sum of core and dispensable genomes (Figure 1). In the case of GBS, presently, the pan-genome contains 2713 genes, of which 1806 belong to the core genome, and 907 belong to the dispensable genome. The GBS pan-genome is predicted to grow by an average of 33 new genes every time a new strain is sequenced (Figure 1). Similar analysis [1••] carried out on five strains of *Streptococcus pyogenes* revealed a similar genomic diversity, indicating an asymptotic value of 27 specific genes for each new genome added, leading, again, to an 'open' pan-genome. A different behavior was observed in the study of eight independent *Bacillus anthracis* isolates. In this case, the number of specific genes added to the pan-genome was found to rapidly converge to zero after the addition of only a fourth genome [1••]. Hence, the *B. anthracis* species has a 'closed' pan genome, and four genome sequences are sufficient to completely characterize this species.

In this review, we discuss how the concept of the pan-genome might fit with the available data and consider which experiments need to be done to address the questions raised by this concept.

A large microbial gene pool driving evolution

Much indirect evidence had already hinted at the concept of the pan-genome, even before it was properly defined by mathematical quantification [1••]. Several studies of subtractive hybridization and comparative genome hybridization (CGH) using multiple isolates of the same species had shown that bacterial species such as *Helicobacter pylori*, *Staphylococcus aureus* and *Escherichia coli* display an extensive genetic diversity, with an average of 20–35% of genes being specific for a single strain [2–4].

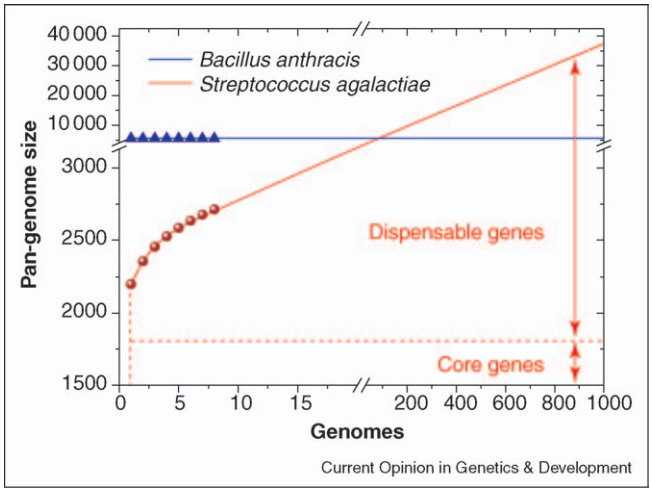
The presence of so many strain-specific genes in each of these species suggests that — as in the case of GBS — they could also display an open pan-genome. This raises

Glossary
Core genome: The pool of genes shared by all the strains of the same bacterial species.
Dispensable genome: The pool of genes present in some — but not all — strains of the same bacterial species.
Lateral gene transfer: Mechanism by which an individual of one species transfers genetic material (i.e. DNA) to an individual of a different species.
Pan-genome: The global gene repertoire of a bacterial species: core genome + dispensable genome.
Serogroup: Group of related bacterial strains characterized by the same composition of the capsular polysaccharide.

the question of whether the microbial world contains enough genes to fit the prediction of such a vast gene pool generated by the pan-genome.

For instance, it has been shown that a single environmental sample of DNA from marine water encodes more than 1.2 million previously unknown genes from 1800 predicted genomic species [5••]. Similar results have also been obtained for a totally different ecosystem, the human gastrointestinal tract. In this case, almost 400 different bacterial phylotypes were identified, of which 244 were novel [6•]. Even more impressive is the recent estimate of 10^7 distinct bacterial species in a 10 gram soil sample containing a total of approximately 10^{10} cells [7••], a species diversity two orders of magnitude larger than previous estimates [8], showing that further quantitative and rational explorations of microbial ecology are strongly needed [9•]. Furthermore, a great heterogeneity was also identified when looking at a single species within a well-defined natural bacterial population: 16S RNA sequencing and pulse-field gel electrophoresis (PFGE) analysis were performed to study the diversity associated with the species *Vibrio splendidus* within coastal bacterioplankton, revealing in this same species the presence of as many as

Figure 1



The set of genes pertaining to a species, or species pan-genome, depends on the number of available genome sequences. In this figure, the size of *S. agalactiae* (red dots) and *B. anthracis* (blue triangles) pan-genomes are shown as a function of the number of sequenced strains. The curves represent a mathematical extrapolation of the data to a large number of strains. The size of a species pan-genome can grow with the number of sequenced strains, or quickly saturate to a limiting value. The *S. agalactiae* pan-genome is 'open'; the *B. anthracis* one is 'closed'. After sequencing a large number of strains, the number of dispensable genes in an open pan-genome is orders of magnitude larger than the size of the core genome, forcing us to reconsider the definition of a bacterial species.

1287 distinct genotypes, most of which are differentiated by the insertion or deletion of large genomic elements [10••]. Although new genes can originate through duplication of existing sequences, followed by diversification, the most common way to acquire new functions is by the

Table 1

Number of genomes sequenced in different bacterial species.

Species with sequenced genome(s)	Number of species (% of the total)	Number of genomes sequences per species
Streptococcus agalactiae, *Bacillus anthracis*, *Burkholderia mallei*	3 (1.2%)	8
Burkholderia pseudomallei	1 (0.4%)	7
Staphylococcus aureus, *Streptococcus pyogenes*	2 (0.8%)	6
Salmonella enterica, *Escherichia coli*, *Bacillus cereus*, *Chlamydophila pneumoniae*, *Haemophilus influenzae*, *Listeria monocytogenes*, *Xylella fastidiosa*	7 (2.8%)	5
Prochlorococcus marinus, *Buchnera aphidicola*, *Burkholderia cenocepacia*, *Ehrlichia ruminantium*, *Legionella pneumophila*, *Pseudomonas syringae*, *Streptococcus thermophilus*, *Yersinia pestis*	8 (3.2%)	3
Streptococcus pneumoniae, *Mycobacterium tuberculosis*, *Neisseria meningitidis*, *Bacillus licheniformis*, *Bifidobacterium longum*, *Campylobacter jejuni*, *Chlorobium phaeobacteroides*, *Corynebacterium glutamicum*, *Haemophilus somnus*, *Helicobacter pylori*, *Lactococcus lactis*, *Leptospira interrogans*, *Mycoplasma genitalium*, *Pseudomonas aeruginosa*, *Shigella flexneri*, *Staphylococcus epidermidis*, *Synechococcus elongates*, *Thermus thermophilus*, *Tropherymaa whipplei*, *Vibrio vulnificus*, *Xanthomonas campestris*	21 (8.3%)	2
Various species	211 (83.3%)	1

Bacterial species for which multiple sequenced strains are available represent a small fraction of the species for which only one strain has been sequenced to date. In this table, we report the number of strains for these species, and the percentage that they represent over the total number of sequenced bacterial species.

transfer of genetic material from unrelated organisms. The importance of the mechanisms of lateral gene transfer (see Glossary) in evolutionary processes has been hotly debated in recent years [11–16,17•,18], but it is now generally accepted that it represents an evolutionary 'fast route', which enables an organism to quickly adapt to a changing environment.

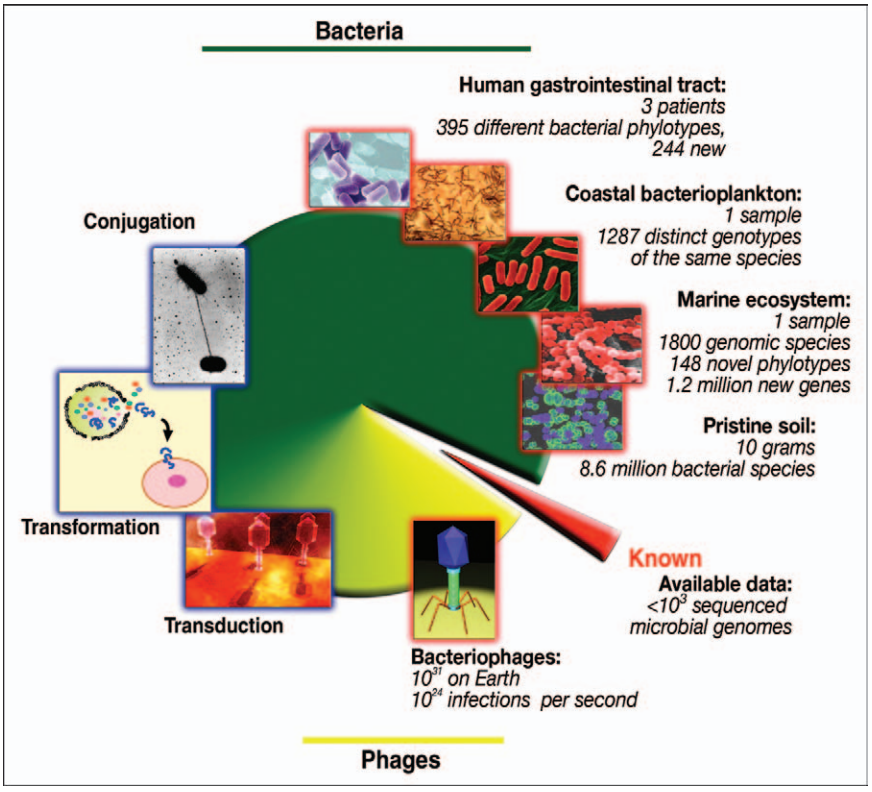
Genes from this large pool are continuously exchanged within and between bacterial species by three main processes: (i) by transformation, when genetic material can be taken up from the environment; (ii) by transduction, when the DNA is delivered by a virus; and (iii) by conjugation, when DNA is directly exchanged between cells. Transformation and conjugation require that the source and target organisms live in close contact, and bacteriophages might enable bacterial species populating different environments to exchange genetic material, which often contains genes that are crucially important for pathogenesis [19•]. Considering that the global population of phages has been estimated to be in the range of 10^{31} and that they are responsible for an average of 10^{23} infections per second [20], it is easy to conclude that the global pool of genes present in the microbial world is likely to exceed by several orders of magnitude any estimate that has been made to date, and that the presence of billions of genes is no longer unexpected (Figure 2).

The surprisingly large gene pool described here suggests that, during evolution, the vast majority of novel functions were probably generated in the microbial world and not in large animals, such as humans, which have only 25 000–35 000 genes. The consequence of this would be that microbes and large animals might have totally different roles in evolution. In fact, under this theory, microbes would generate new genes and functional modules, whereas large animals would evolve by first taking up modules generated by microbes and then by rearranging them in many different ways within the genome itself and by alternative splicing of the mRNAs.

Core and dispensable genes

In general, the core genome includes all genes responsible for the basic aspects of the biology of a species and its major phenotypic traits. By contrast, dispensable genes contribute to the species diversity and might encode

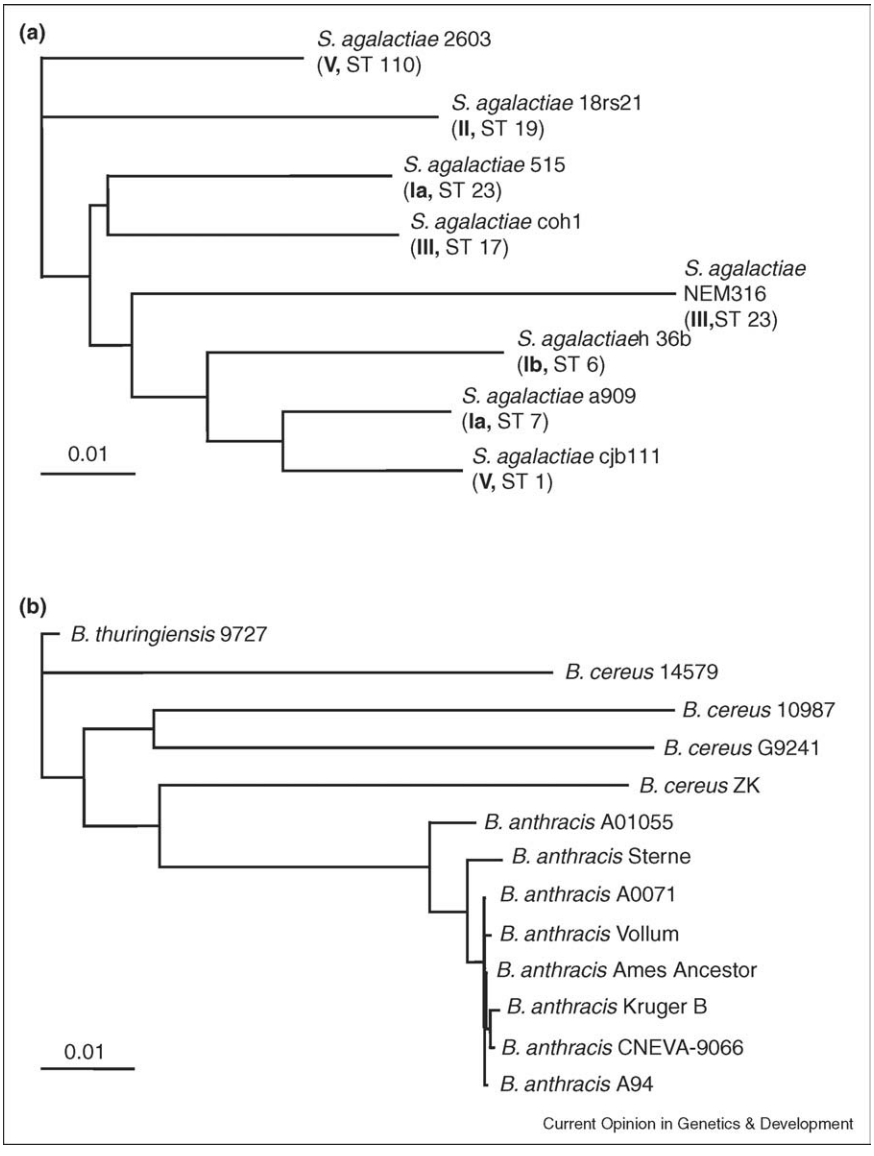
Figure 2



Diversity of the microbial universe. Recent experimental findings have shown that the genetic diversity within bacterial populations is much higher than expected [5••,6•,10••,20], raising the question of what is the origin of this genetic diversity. Bacteria can acquire genes from the environment by conjugation, transformation and phage infection (transduction), and experimental studies on environmental samples have shown that the amount of genetic material available in single ecosystems is large enough to constitute a virtually infinite reservoir of new genes. Estimates of the contribution of mechanisms of lateral gene transfer to the innovation rate show that genes acquired through this route give an essential contribution to species diversification [18,19•,20,21]. To date, of all this diversity, less than 1000 microbial genomes have been completely sequenced.

supplementary biochemical pathways and functions that are not essential for bacterial growth but which confer selective advantages, such as adaptation to different niches, antibiotic resistance, or colonization of a new host. Such genes are generally clustered on large genomic islands that are typically flanked by short repeated DNA sequences and are characterized by an abnormal G + C content. Investigation and functional annotation of dispensable genes reveals that hypothetical, phage- and transposon-related genes account for the vast majority of findings, whereas in a typical genome this type of gene represents much smaller percentages [21]. The fact that these genes are mostly associated with a limited number of strains indicates a weak positive selection for these functions and shows that mobile elements contribute poorly to the overall fitness and differentiation of the species, although sometimes they can carry important genes [19•,22••]. Given that these genes are not necessary for

Figure 3



Dendrograms of the eight *Streptococcus agalactiae* **(a)** and thirteen *B. cereus* group **(b)** genomes. The fraction of genes of one strain that is not shared with other strains was used to define a distance matrix. The matrix was then used to build a dendrogram with the neighbor-joining method, as implemented in the NEIGHBOR program of the PHYLIP suite (http://evolution.genetics.washington.edu/phylip.html). Tree branches are proportional to the fraction of the gene content not shared between the different isolates, and the ruler shows the length corresponding to 1% difference in gene content.

For each *S. agalactiae* strain, serogroup (bold, roman letters) and sequence type (ST) are reported in brackets. The *S. agalactiae* and *B. cereus* group genomes are available for downloading at http://www.ncbi.nlm.nih.gov. From the figure, it is evident that the distance between two *S. agalactiae* strains is comparable to the distance between *B. anthracis* stains and other *B. cereus* group species, making the definition of *B. anthracis* as an autonomous species questionable.

survival or maintenance of the species, they can also be deleted from the genome; however, in pathogenic species, this loss is often accompanied by a parallel reduction in virulence. For example, a spontaneous loss of the genes coding for fimbriae, hair-like projections thought to have an important role in colonization, has been observed in successive passages of *in vitro* cultures of *Haemophilus influenzae* and *E. coli*. Similarly, GBS was recently found to encode a pilus-like structure, which is not ubiquitous in all strains, and the presence or absence of which could be related to either gene acquisition or loss [23].

Serotypes and sequence types do not correlate with genomic diversity

Classical methods to catalogue bacterial species are based on knowledge convenient phenotypic traits. The most popular is the agglutination of bacterial cells by specific antisera against the capsular polysaccharide surrounding many pathogens. For a variety of encapsulated bacteria, this method has been widely used for epidemiology studies and vaccine design, assuming that all strains belonging to the same serogroup are similar. More recently, techniques such as multilocus enzyme electrophoresis (MLEE) and multilocus sequence typing (MLST), which are based on the detection of variability associated with housekeeping genes, were applied to several bacterial species and led to the classification of strains into 'clonal complexes' and sequence types, respectively.

However, comparison of the whole genome sequences of GBS strains has shown that the genomic diversity does not segregate with serotypes or MLST sequence-types (Figure 3a). In fact, the analysis revealed that, often, isolates belonging to different serogroups are more closely related than are isolates of the same serogroup, and that strains of the same sequence type can be genetically very distant (Figure 3a). The reason for the absence of correlation between serotypes and genetic diversity is likely to reside in the fact that capsular specificity genes are present in the dispensable genome, which is exchanged freely between strains with different genetic background. By contrast, the genes used to determine the MLST type belong to the core genome, and they do not pick up similarities present in the dispensable genome, which often are linked to pathogenic features.

Challenging the concept of species

Species can have an open or a closed pan-genome. An open pan-genome is typical of those species that colonize multiple environments and have multiple ways of exchanging genetic material. *Streptococci*, *Meningococci*, *H. pylori*, *Salmonellae* and *E. coli* have these properties and are likely to have an open pan-genome. By contrast, other species such as *B. anthracis*, *Mycobacterium tuberculosis* and *Chlamydia trachomatis*, which are known to be more conserved, live in isolated niches with limited access to the global microbial gene pool. Such species, with a low capacity to acquire foreign genes, have a closed pan-genome. An extreme example is represented by *Buchnera aphidicola*, an endosymbiont of aphids, the genome of which has undergone no chromosome rearrangements, duplications or horizontal gene transfer in the past 50 million years, thus demonstrating the most extreme genome stability observed to date [24].

A closer look at the structures of the genetic trees of open pan-genomic species (such as GBS; see Figure 3a) and closed pan-genomic species (such as *B. anthracis*) shows that the latter species resembles a clone of a *B. cereus* species rather than being a true independent species (Figure 3b). *B. anthracis* is, in fact, genetically very closely related to other members of the *B. cereus* group (*B. cereus* and *Bacillus thuringiensis* species), and the main feature that distinguishes these organisms is the acquisition of two virulence plasmids, one of which codes for anthrax toxin [25]. Although this feature is extremely important in justifying the classification of *B. anthracis* as an independent species, genetically, this is just a phenotypic trait encoded by the dispensable genome of the *B. cereus* group. This example shows that the criteria used to define microbial species might be inconsistent with the genetic information. In the future, we will need to consider how to handle these inconsistencies.

Conclusions and practical implications

The need to sequence multiple genomes from each species to better understand the diversity of bacterial species is not just a theoretical exercise. Recently, it has been shown that the design of a universal vaccine against GBS was only possible using dispensable genes [26•]. In addition, sequencing of multiple genomes was instrumental in discovering the presence of the pilus in GBS, group A *Streptococcus*, and *Pneumococcus*, an essential virulence factor that had been missed by all conventional technologies for a whole century [23]. It is very likely that the study of the bacterial pan-genome will continue to surprise us with fascinating discoveries that cannot be predicted with the conventional methods used to date in microbiology.

Acknowledgements

The authors would like to thank Michael Cieslewicz, Antonello Covacci and John Telford for their contribution to the pan-genome concept. They are also grateful to the IRIS Bioinformatic group, to the TIGR information technology and database server groups led by Vadim Sapiro and Michael Heaney, respectively, and to Giorgio Corsi for artwork.

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

1. •• Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones A, Durkin AS *et al.*: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'**. *Proc Natl Acad Sci USA* 2005, in press.

This study reports the first quantification of the diversity of a single prokaryotic species on the basis of genomic sequences of multiple strains. The authors introduce the pan-genome concept: the gene pool pertaining to a single species, which can be orders of magnitude larger than the genome of any single isolate.

2. Fitzgerald JR, Sturdevant DE, Mackie SM, Gill SR, Musser JM: **Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic**. *Proc Natl Acad Sci USA* 2001, **98**:8821-8826.

3. Dorrell N, Mangan JA, Laing KG, Hinds J, Linton D, Al-Ghusein H, Barrell BG, Parkhill J, Stoker NG, Karlyshev AV *et al.*: **Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity**. *Genome Res* 2001, **11**:1706-1715.

4. Fukiya S, Mizoguchi H, Tobe T, Mori H: **Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray**. *J Bacteriol* 2004, **186**:3911-3921.

5. •• Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W *et al.*: **Environmental genome shotgun sequencing of the Sargasso sea**. *Science* 2004, **304**:66-74.
The first work in which the 'whole-genome shotgun sequencing technique' was applied to an environmental sample. It greatly improves our understanding of how complex and variable is the world of uncultured bacteria.

6. • Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA: **Diversity of the human intestinal microbial flora**. *Science* 2005, **308**:1635-1638.
A study showing that intestinal flora contains a much higher number of species than expected, with relevant variations both between different sites and between different subjects.

7. •• Gans J, Wolinsky M, Dunbar J: **Computational improvements reveal great bacterial diversity and high metal toxicity in soil**. *Science* 2005, **309**:1387-1390.
This excellent study introduces innovative computational methods for measuring species abundance in environmental samples, leading to an increase of more than two orders of magnitude in current estimates of species diversity.

8. Sandaa R, Torsvik V, Enger Ø, Daae FL, Castberg T, Hahn D: **Analysis of bacterial communities in heavy metal-contaminated soils at different levels of resolution**. *FEMS Microbiol Ecol* 1999, **30**:237-251.

9. • Curtis TP, Sloan WT: **Exploring microbial diversity — a vast below**. *Science* 2005, **309**:1332-1333.
Illuminating perspective on the new frontiers opened in microbial ecology by the use of advanced mathematical and computational approaches to the quantification of prokaryotic diversity.

10. •• Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, Sarma-Rupavtarm R, Distel DL, Polz MF: **Genotypic diversity within a natural coastal bacterioplankton population**. *Science* 2005, **307**:1311-1313.
This study presents illuminating data on the genetic diversity of a bacterial population, clearly showing that macroscopic differences in genetic content are present also within seemingly homogeneous populations from a single ecosystem.

11. Kurland CG, Canback B, Berg OG: **Horizontal gene transfer: a critical view**. *Proc Natl Acad Sci USA* 2003, **100**:9658-9662.

12. Lawrence JG, Hendrickson H: **Lateral gene transfer: when will adolescence end?** *Mol Microbiol* 2003, **50**:739-749.

13. Koonin EV: **Horizontal gene transfer: the path to maturity**. *Mol Microbiol* 2003, **50**:725-727.

14. Ochman H, Lerat E, Daubin V: **Examining bacterial species under the specter of gene transfer and exchange**. *Proc Natl Acad Sci USA* 2005, **102(Suppl 1)**:6595-6599.

15. Simonson AB, Servin JA, Skophammer RG, Herbold CW, Rivera MC, Lake JA: **Decoding the genomic tree of life**. *Proc Natl Acad Sci USA* 2005, **102(Suppl 1)**:6608-6613.

16. Hooper SD, Berg OG: **Duplication is more common among laterally transferred genes than among indigenous genes**. *Genome Biol* 2003, **4**:R48.

17. • Dobrindt U, Hochhut B, Hentschel U, Hacker J: **Genomic islands in pathogenic and environmental microorganisms**. *Nat Rev Microbiol* 2004, **2**:414-424.
Comprehensive review presenting the results on lateral gene transfer obtained for pathogenicity islands in pathogenic bacteria. It shows that lateral gene transfer is a universal mechanism of evolution.

18. Kazazian HH Jr: **Mobile elements: drivers of genome evolution**. *Science* 2004, **303**:1626-1632.

19. • Brussow H, Canchaya C, Hardt WD: **Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion**. *Microbiol Mol Biol Rev* 2004, **68**:560-602.
A clear review highlighting the importance of phages as a source of novel genes in bacterial evolution.

20. Hendrix RW: **Bacteriophage genomics**. *Curr Opin Microbiol* 2003, **6**:506-511.

21. Daubin V, Ochman H: **Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli***. *Genome Res* 2004, **14**:1036-1042.

22. •• Feil EJ: **Small change: keeping pace with microevolution**. *Nat Rev Microbiol* 2004, **2**:483-495.
Outstanding review, detailing the latest experimental results for characterizing species variability, with a special focus on genome sequencing, microarray and MLST data. The consequences on the concept of species and on evolutionary mechanisms in bacteria are thoroughly discussed.

23. Lauer P, Rinaudo CD, Soriani M, Margarit I, Maione D, Rosini R, Taddei AR, Mora M, Rappuoli R, Grandi G *et al.*: **Genome analysis reveals pili in Group B *Streptococcus***. *Science* 2005, **309**:105.

24. Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG: **50 million years of genomic stasis in endosymbiotic bacteria**. *Science* 2002, **296**:2376-2379.

25. Rasko DA, Altherr MR, Han CS, Ravel J: **Genomics of the *Bacillus cereus* group of organisms**. *FEMS Microbiol Rev* 2005, **29**:303-329.

26. • Maione D, Margarit I, Rinaudo CD, Masignani V, Mora M, Scarselli M, Tettelin H, Brettoni C, Iacobini ET, Rosini R *et al.*: **Identification of a universal group B *Streptococcus* vaccine by multiple genome screen**. *Science* 2005, **309**:148-150.
A clear demonstration of the practical importance of sequencing multiple strains of a single pathogen for the formulation of an effective vaccine.