ELSEVIER

FEMS Microbiology Reviews 29 (2005) 147–167

FEMS MICROBIOLOGY Reviews

www.fems-microbiology.org

Towards a prokaryotic genomic taxonomy ☆

Tom Coenye [a,*,1], Dirk Gevers [a,b,1], Yves Van de Peer [b], Peter Vandamme [a], Jean Swings [a,c]

[a] *Laboratory of Microbiology, Ghent University, Ledeganckstraat 35, B-9000 Ghent, Belgium*
[b] *Bioinformatics and Evolutionary Genomics, Ghent University/Flanders Interuniversity Institute for Biotechnology (VIB), Technologiepark 927, B-9052 Ghent, Belgium*
[c] *BCCM/LMG Bacteria Collection, Ghent University, Ledeganckstraat 35, B-9000 Ghent, Belgium*

Received 24 March 2004; received in revised form 9 September 2004; accepted 7 November 2004

First published online 7 December 2004

Abstract

One of the most interesting developments in the field of modern-day microbiology is the ever increasing number of whole-genome sequences that is publicly available. There is an increasing interest in the use of these genome sequences to assess evolutionary relationships among microbial taxa, as it is anticipated that much additional taxonomic information can be extracted from these sequences. In a first part of the present review, mechanisms that are responsible for the evolution of genomes will be discussed. Subsequently, we will give an overview of approaches that are presently available to assess the taxonomic relationships between prokaryotic species based on complete genome sequences, followed by a brief discussion of the potential implications of these novel approaches for bacterial taxonomy in general and our thinking about the bacterial species concept in particular.

Keywords: Microbial taxonomy; Whole-genome sequences; Comparative genomics; Microarrays; Multilocus sequence typing; Prokaryotic species concept

Contents

1. Introduction . . . 148
2. The bacterial species genome . . . 148
3. Shaping the prokaryotic genome . . . 149
   3.1. Gene duplications . . . 149
   3.2. Horizontal gene transfer . . . 150
   3.3. Gene loss . . . 150
   3.4. Chromosomal rearrangements . . . 151
   3.5. Impact on taxonomy . . . 151
4. Novel approaches for assessing taxonomic relationships among prokaryotes based on whole-genome sequences . . . 154
   4.1. Gene content . . . 154
   4.2. Gene order . . . 154
   4.3. Comparative sequence analysis of conserved macromolecules . . . 154

☆ Edited by Annick Wilmotte.

* Corresponding author. Tel.: +32 9 264 5128; fax: +32 9 264 5092.
E-mail address: tom.coenye@ugent.be (T. Coenye).

[1] Both authors contributed equally.

0168-6445/$22.00
doi:10.1016/j.femsre.2004.11.004

4.4. Genome blast distance phylogeny . . . 156
4.5. Presence/absence analysis . . . 157
4.6. Nucleotide composition . . . 157
4.7. Metabolic pathway reaction content . . . 157
4.8. Examples and discussion . . . 157
5. Comparative microbial genomics using DNA microarrays and subtractive hybridisation . . . 159
6. Detection of intraspecies diversity . . . 159
7. Implications for the prokaryotic species concept . . . 160
8. Some concluding remarks . . . 161
Acknowledgements . . . 162
References . . . 162

1. Introduction

Polyphasic taxonomy is aiming at the integration of different kinds of data and information (including phenotypic, genotypic and phylogenetic information) that allow classification of biological entities [1]. Since the 1970s, physicochemical DNA–DNA hybridisation methods have been used to delineate the prokaryotic species, and at present a bacterial species is "a category that circumscribes a (preferably) genomically coherent group of individual isolates/strains sharing a high degree of similarity in (many) independent features, comparatively tested under highly standardised conditions" [2]. Practically, bacterial species are considered to be groups of strains that are characterised by a certain degree of phenotypic consistency, by a significant degree (50–70%) of DNA hybridisation and over 97% of 16S ribosomal RNA (rRNA) gene sequence identity [1,3]. As stated by the ad-hoc committee for the re-evaluation of the species definition in bacteriology [2], the introduction of innovative methods is providing new opportunities for prokaryotic systematics. One of the particularly interesting developments includes the analysis of complete genome sequences. The steady increase in the number of completely sequenced prokaryotic genomes witnessed over the last decade has been associated with advances in sequencing technology, a boom in bioinformatics, and sustained funding [4,5]. Recently, there has been an increasing interest in the use of these genome sequence data to assess evolutionary relationships among prokaryotic species. Although 16S rRNA gene sequences and DNA–DNA hybridisation continue to be considered as molecular criteria for species delineation, it is anticipated that much additional taxonomic information can be extracted from complete genome sequences.

In a first section of this paper, we will discuss the mechanisms that are responsible for the evolution of genomes. Subsequently, we will give an overview of approaches that are presently available to assess the taxonomic relationships between prokaryotic species based on complete genome sequences. In a final section, we will briefly discuss the potential implications of these novel approaches to bacterial taxonomy and the effect they may have on our thinking about the bacterial species concept.

2. The bacterial species genome

There are now a number of conspecific genomes whose gene content can be compared in silico. From these comparisons, and from results obtained with advanced molecular tools like microarrays, it can be concluded that the genome sequence of one single strain does not offer us a complete picture of the genetic diversity of a species. For example, as shown in Table 1, the gene content of *Escherichia coli* genomes can be as different as 29.25%. According to Cohan [6], it should not come as a surprise that species contain this magnitude of genetic and ecological diversity. For decades, taxonomists have known that there can be considerable variation in metabolic traits within named species. In this regard, Lan and Reeves [7] suggested that we define a 'species genome' as comprising all the genes present in the characterised strains of a species, in order to get an idea of the variation in the species. This species genome consists of two components, a core of genes present in most of the strains, and an 'auxiliary set'. In Table 1, the strict species core (genes present in all conspecific strains) for some species for which a triplet (or quartet) of complete genome sequences is available is shown. Although such results might be biased by the strains included (e.g., inconsistent phylogenetic sampling), a correlation between the species' lifestyle and the magnitude of the species core is reflected in the results. Obligate intracellular parasites, such as *Chlamydophila pneumoniae*, are closely adapted to the physiologically stable environments of their host cells and tend to contain fewer auxiliary genes (see also [8]). This extreme stability contrasts with the high genomic variability observed in their free-living relatives, such as *E. coli*, known to have a large flexible gene pool providing the properties to adapt to special environmental conditions.

Table 1
The species core (genes present in all conspecific strains) and gene content similarity for some species for which a triplet/quartet of complete genome sequences is available

Species and strain designation	No. (%) ORFs in species core[a]	Gene content similarity matrix[b]			
Chlamydia pneumoniae (949 species core families)					
CWL029	1034 (98.01%)	–			
AR39	1031 (92.63%)	95.94	–		
J138	1032 (96.54%)	98.62	94.42	–	
Escherichia coli (2616 species core families)					
K12	3710 (86.06%)	–			
EDL933	3879 (72.23%)	83.42	–		
RIMD0509952	3864 (71.78%)	83.01	93.02	–	
CFT073	3793 (70.24%)	77.04	72.52	71.75	–
Streptococcus pyogenes (1257 species core families)					
SF370	1541 (90.22%)	–			
MGAS8232	1578 (84.98%)	89.46	–		
MGAS315	1565 (83.38%)	87.41	90.73	–	
SSI-1	1556 (83.16%)	87.27	90.24	96.41	–

[a] For each species we determined the species core families, i.e., the gene families that have at least one member in all strains of the species, and calculated the number of genes in these families for each of the strains separately.
[b] Gene content similarity was calculated from the number of common gene families divided by the weighted genome size [56].

3. Shaping the prokaryotic genome

The availability of complete genome sequences of closely related organisms presents an opportunity to reconstruct events of genome evolution. It has become clear that, in addition to nucleotide substitutions, other genetic forces shape the genome. By comparing related genomes and inferring ancestral ones, we can identify events such as specific chromosomal rearrangements, gene acquisition, and deletions, which have led to a vast diversity in genome content and organisation. Recently, efforts have been made to quantify the relative contribution of these processes through comparative analysis of multiple prokaryotic genomes [9–11]. Despite these efforts, and because of inconsistent results, evolutionary models that accurately describe the phylogenetic relationships of micro-organisms based on their genome are not yet described [12]. Here, we address the evolutionary dynamics of prokaryotic genomes by presenting an overview of the principal genetic forces that shape prokaryotic genomes, including gene duplication, horizontal gene transfer, gene loss and chromosomal rearrangements, in the light of a comparative genomics approach, and discuss their impact on prokaryotic taxonomy.

3.1. Gene duplications

Gene duplication is considered an important mechanistic antecedent of gene innovation, and consequently of genetic novelty, that has facilitated adaptation to changing environments and exploitation of new niches [13,14]. The role of gene duplications in evolution has been studied extensively [9,10,15,14]. With the availability of numerous complete genome sequences, the study of gene duplication has moved on from specific genes or gene families to the level of complete genomes. A wealth of data can now be examined to assess various aspects of the role of gene duplication in prokaryotic genome evolution. Whereas in the pre-genomic era, it was suggested that bacterial genomes may have evolved from small to large genomes by several genome duplications [16], the analysis of the complete genomes of *Haemophilus influenzae* and *Mycoplasma genitalium* did not support this idea [17]. A recent analysis of the prevalence and genomic organisation of paralogs in bacterial genomes, showed that most of the duplicated genes in *Bacteria* seem to have been created by small gene duplication events [18]. Evidence for large-scale gene duplications such as those observed in eukaryotic genomes [19–21] has not been detected in any of the bacterial genomes investigated so far. Nevertheless, paralogous genes comprise a significant fraction (up to 44%) of the bacterial genome coding capacity [15,14,18,22]; a fraction that is found to have a strong linear correlation with the genome size. Regarding the organisation of paralogous genes within the genomes, it was found that 15% of the paranome (collection of paralogous genes) of 106 bacterial genomes investigated consists of tandem duplicates, 9.5% is located in block duplicated segments, and the majority (75.5%) lost its organisation as a consequence of genome dynamics [18]. Most of the block duplications resemble the typical bacterial operon size (3–4 genes), thereby indicating a putative history of operon duplication and its retention. Jordan et al. [15] were specifically concerned with the contribution of gene duplication to the genomic differences between genera of prokaryotes, and studied lineage-specific gene expansions, i.e., groups of paralogous genes generated

following the divergence of specific prokaryotic lineages as inferred from comparative 16S rRNA gene sequence analysis. It was found that these paralogous gene families are likely to contribute substantially to the phenotypic differences between bacterial lineages. Recently, a clear correlation between the number of retained duplicates and the functional class to which they belong was reported [18]. For example, in the mycobacterial paranome the functional class of genes encoding proteins involved in fatty acid metabolism is represented by an excess of retained duplicated genes, which is in agreement with the complex nature of the mycobacterial cell wall, and might reflect adaptive evolution of the bacterial cell wall [18,23]. This shows that particularly genes involved in the adaptation to a constantly changing environment seem to have been preserved after duplication, demonstrating the importance of gene duplication for biological evolution.

3.2. Horizontal gene transfer

Besides gene duplication and subsequent functional divergence, prokaryotes have an alternative mechanism for genetic adaptation to their environment. The introduction of novel genes or alleles by horizontal gene transfer (HGT) allows for niche-specific adaptation, which eventually might lead to bacterial diversification and speciation [24,25]. Although HGT was known to be involved in, for example the transmission of antibiotic resistance genes, it is only through recent comparative analysis of multiple prokaryotic genomes that the scale of HGT became apparent [26–29]. This finding has even been suggested as the most fundamental change in our perception of general aspects of prokaryotic biology brought about by massive genome sequencing [30]. According to Woese [31], evolution as we know it seems impossible without the interplay between vertically generated and horizontally acquired variation. In an extreme view, it has been suggested that two taxa are more similar than a third one, not because they share a more recent common ancestor, but because they exchange genes more frequently [32]. The awareness of the potential impact of HGT has been translated into a theoretical concept of three categories of genes: (i) a 'hard core' of genes recalcitrant to HGT, (ii) a 'soft core' with genes that are rarely transferred and (iii) the 'shell' genes which are susceptible to HGT [33]. Practically speaking, HGT is probably one of the most controversial topics in genomics. The essence of the problem is that any phylogenetic evidence of HGT can also be explained via a combination of gene duplication and lineage-specific gene loss events and high-throughput detection of HGT is therefore far from straightforward [30,33]. Although there are various studies in which HGT has been observed (and quantified) directly (see for example [34]), a great deal of work is thus yet required to unambiguously quantify the rates at which different processes occur [33]. Simplistic estimations of the frequency of HGT at the genome level in a phylogenetic context seem to indicate that it is rather low (less than 10% of the genes) [9,10,35]. Nevertheless, HGT has a significant biological impact as many pathogenic properties are encoded on plasmids, phages or 'pathogenicity islands' and the transfer of such genetic elements is widely thought to be associated with the origin of pathogenic clones [36].

3.3. Gene loss

Bacterial genomes are not growing ever larger in size, they are sampling rather than accumulating sequences, consequently gene acquisition (both by duplication and HGT) must be counter balanced by gene loss. Because bacterial genomes can only protect a finite amount of information against mutation and loss, chromosomal deletions will serve to eliminate genes that fail to provide a meaningful function, that is, the bulk of acquired DNA as well as superfluous ancestral sequences [37,38]. In some cases, the loss of gene function may provide a selective advantage, as for example in the succession of genetic events contributing to virulence in *Shigella* [39]. As *Shigella* spp. evolved from *E. coli* to become pathogens, they not only acquired virulence genes on a plasmid but also shed genes via deletion of a large genomic segment. The formation of these 'black holes', i.e., deletions of genes that are detrimental to a pathogenic lifestyle, provides an evolutionary pathway that enables a pathogen to enhance virulence. Extreme genome reduction has been documented in several bacterial groups with a host-associated lifestyle, including mycoplasmas, chlamydiae, spirochetes, buchneras and rickettsias. Host tissues provide a constant supply of many metabolic intermediates, eliminating the pressure to maintain many biosynthetic genes. Adaptation of the intracellular organisms to the physiologically stable environments of their host cells causes selection to be less effective in protecting against gene inactivation and loss, even of beneficial but non-essential genes, and consequently these intracellular organisms experience a deletional bias [38,40–42]. This occurs because host-associated bacteria have small genetic population sizes compared to free-living relatives, resulting in higher levels of fixation of slightly deleterious mutations, i.e., a population bottleneck [43]. In the pre-genomic era, it was thought that reduced genomes converged on a set of universal genes that underlie the core processes of cellular growth and replication, with each genome also containing some specific loci tuned to that species' ecology or relationship with its host [44]. This is contradicted by the analysis of complete genome sequences. It seems that each lineage has taken a different evolutionary route to minimalism, whereby the same

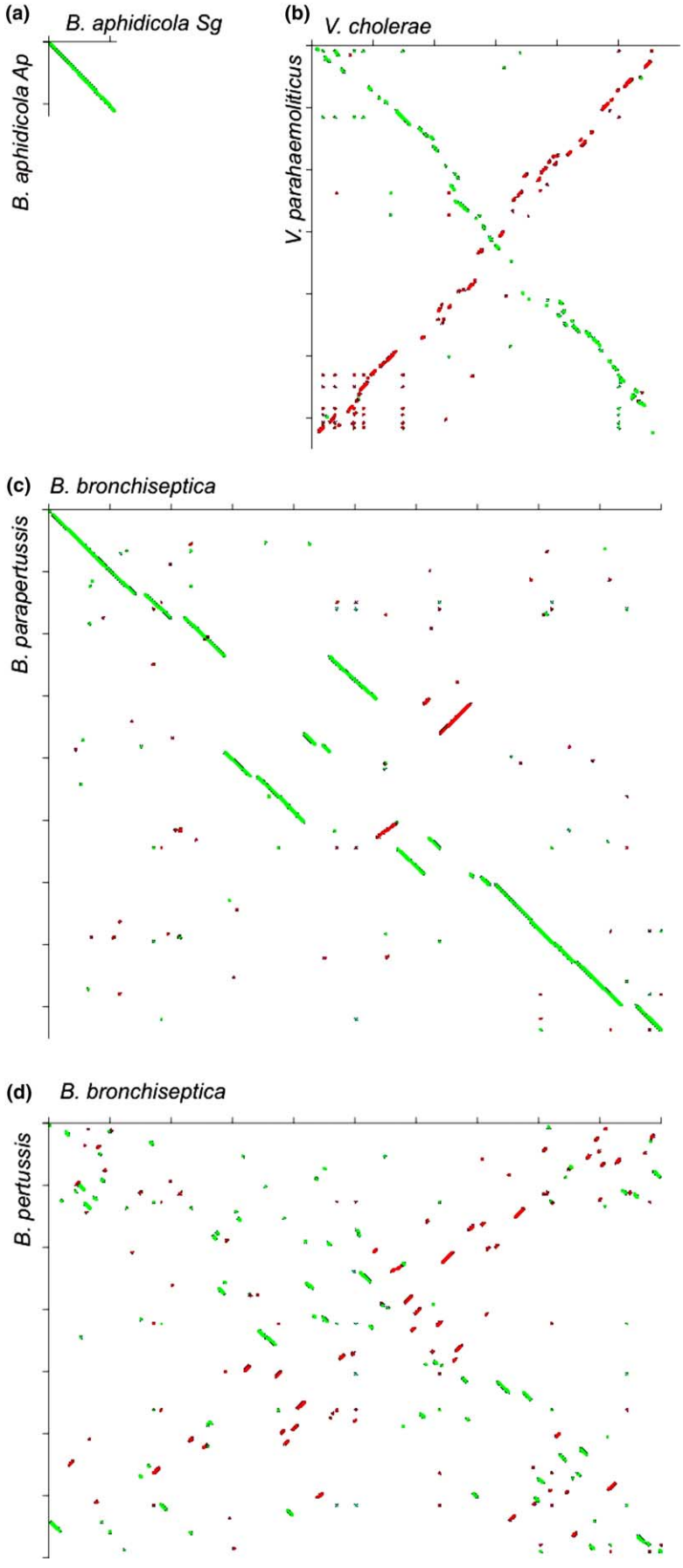
functions can be achieved by retention of non-homologous genes [45]. Knowing that: (i) the bacterial core set of genes consists presumably of less than 100 genes [30] and (ii) that most obligate intracellular bacteria are phylogenetically distantly related, it is less surprising that different minimal genomes do not share a large number of genes. Genome reduction is thought to be a main force behind the evolution of parasitic and/or intracellular bacteria [42,46], and is referred to as evolution by reduction. *Buchnera* spp. and *Wigglesworthia* spp. posses almost no genes that are not present in close relatives, suggesting that the shift to obligate associations with hosts does not necessarily require acquisition of novel genes [47]. Gene loss of individual loci or operons is the only source of divergence in the gene inventories of *Buchnera* strains [48,49]. Whereas the process of genome reduction is accompanied by chromosomal rearrangements, the genomes in a more advanced stage of reduction and specialisation to their hosts are often associated with an exceptional level of genome stability, as observed within the last 150 million years of evolution of *Buchnera* (Fig. 1(a)) [48,49]. Also other endosymbiontic bacteria like the rickettsia's have highly conserved gene content and order [50]. A snapshot of genome degradation in progress is provided by the complete sequence of the genome of *Mycobacterium leprae* [51]. Comparing *Mb. leprae* to its relative *Mycobacterium tuberculosis* indicates that the *Mb. leprae* lineage has discarded more than 2000 genes. DNA corresponding to more than 1000 of these genes is still present as partial copies or as non-functional pseudogenes, the largest proportion of non-coding DNA of any fully sequenced bacterial genome [44].

3.4. Chromosomal rearrangements

Besides the phenomena that influence variation in gene content, described above, genome rearrangement is a driving force behind a constantly evolving genome organisation. Basic questions on bacterial genome organisation and evolution could not be addressed before the complete sequencing of two bacterial genomes. From an analysis of the first two sequenced genomes, *H. influenzae* and *M. genitalium*, it was found, rather unexpectedly, that gene order was not conserved throughout bacterial evolution [52]. As more genomes were sequenced, it was found that the level of conservation can be high when organisms are phylogenetically closely related, mostly because rearrangements not yet had time to occur, but that conservation is lost when phylogenetic distance increases [53]. Gene order in prokaryotes, even at the operon level, is conserved to a much lesser extent than average protein sequence similarity [54]. These observations appear to indicate that the relative gene position is not substantially essential for gene function. Suyama and Bork [55] evaluated the conservation of gene order in relation to an evolutionary timescale, and concluded that not only the number of amino acid substitutions, but also the degree of genome rearrangement constantly increases along the time of divergence. Gene order conservation could therefore be used as a phylogenetic measure to study relationships between species [56]. The degree of genome flexibility is dependent on the content of repeated and mobile sequences such as insertion sequence elements (ISE), (conjugative) transposons, plasmids and bacteriophages [17]. Chromosomes containing a higher repeat density have higher rates of rearrangements leading to accelerated loss of gene order [57]. ISE are particularly important as mediators of single gene rearrangements by offering multiple similar sequences among which recombination can be initiated. For example, at some point in its recent evolution, *Bordetella pertussis* seems to have undergone a massive expansion of one specific family of ISE (*IS481*), compared to its close relatives *Bordetella parapertussis* and *Bordetella bronchiseptica*. Subsequent recombinations between these perfect DNA repeats can explain the large amount of rearrangement and deletion in the chromosome, probably contributing to species diversification (Fig. 1(c) and (d)) [58]. In addition, genome evolution is influenced by large-scale chromosomal rearrangements, especially large inversions, leading to the phenomenon of x-shaped patterns in alignments of whole genomes (Fig. 1(b)) [55,59,60]. These large segment inversions are caused by homologous recombination between inverted repeats. Remarkably, inversions that get fixed in natural populations tend to be symmetrical around the replication origin, thereby minimizing the disruption of the chromosomal structure relative to replication. Non-symmetrical large inversions are relatively rare in natural strains, suggesting that they lead to a reduced fitness. A factor that could lead to the biased genome rearrangements is the distance of a gene from the origin of replication, as this determines the relative copy number of the gene in each cell of fast-growing cultures of bacteria [61]. Furthermore, it has been suggested that these replication-directed translocation processes can explain not only genome-wide rearrangements, but also local rearrangements and possibly even gene gain and loss in the evolutionary divergence of genomes, thereby reserving a major role for replication in directing genome evolution [59].

3.5. Impact on taxonomy

Increased insights in the different mechanisms discussed above should allow us to accurately describe evolutionary history explaining the phylogenetic relationships of micro-organisms. This would allow us to develop a natural classification that is truly representing evolutionary history.


(a)
B. aphidicola Sg
B. aphidicola Ap
(b)
V. cholerae
V. parahaemoliticus
(c)
B. bronchiseptica
B. parapertussis
(d)
B. bronchiseptica
B. pertussis

Recently, the α-*Proteobacteria* were used as a model system to analyse the contribution of different processes (duplication, HGT, loss) in the genome evolution of this bacterial lineage [62]. A computational inference of ancestral genomes provided a scenario for the expansion and contraction in the genomic repertoire (reflecting population size and bacterial lifestyle) and an insight in the functional categories mostly affected by these alterations. In addition to a better understanding of the evolutionary history and phylogenetic relationships, the study demonstrated the importance of gene duplication for expansion and diversification and a previously underestimated number of paralogous proteins. The authors suggest that the continuous generation of novel paralogs may provide one explanation for the difficulty to obtain congruent single gene trees in phylogenomic surveys.

Genome variability as a consequence of the actions of mobile genetic elements has phenotypic consequences [63]. To what extent this variability also results in novel microbial taxa to be generated is largely an open question. Adaptation of micro-organisms to a new niche promotes the evolution of new species. Some niche invasions are accessible through mutations of existing genes [64], whereas the majority of the bacteria evolve into new niches by acquiring new gene loci from other species [65]. It has been argued that a classification scheme that is based on the analysis of a core set of genes (which is thus ignoring the majority of the information content of the entities to be classified) can have little claim of being natural [66]. HGT contributed to evolutionary history and should thus be considered when defining relationships between organisms.

The classical genotypic methods used in bacterial taxonomy (determination of the G + C content, DNA–DNA hybridisation studies and 16S rRNA gene sequencing) will not significantly be influenced by either of the forces described above. The G + C content of a genome is not significantly affected by chromosomal rearrangements, gene los or gene duplications. While HGT could have an effect on the G + C content, the 'amelioration' of horizontally transferred genes (since introgressed genes are subject to the same mutational processes affecting other genes in the recipient genome, over time they will start to reflect the DNA composition of the acceptor genome) will prevent them from having a lasting effect on the G + C content [67]. During DNA–DNA hybridisations, the overall ability of fragments of two genomic sequences to form heteroduplexes is measured [68], and it is expected that this measure will be influenced mostly by sequence divergence and only to a lesser extent by rearrangements, gene loss or duplications, and HGT. DNA-based typing methods (in which a DNA banding pattern is generated) include the separation of macrorestriction fragments by pulsed-field gel electrophoresis (PFGE) and various PCR-based methods (including BOX, RAPD and AFLP fingerprinting) [1,69,70]. Although most of these typing methods were originally developed for subdividing species into a number of distinct types, they are often also used to derive relationships between species [71]. The banding patterns obtained by these methods can be altered in various ways. Chromosomal rearrangements (including large insertions and/or deletions) can have a tremendous effect on banding patterns generated by restriction enzyme analysis based methods like PFGE [72], as well as on patterns obtained with several PCR-based methodologies [73]. The loss and/or gain of restriction sites and/or primer-binding sites can also result in altered patterns. Now that many genome sequences are available it is possible to link differences observed in fingerprints to the actual differences in the genome; this way these genomic typing methods will learn us more about the actual evolution of the bacterial genome.

Phenotypic and chemotaxonomic characteristics are influenced indirectly by changes to the genome. The effect of gene duplications on phenotypic and chemotaxonomic characteristics will depend on the evolutionary fate of the duplicated gene. Neofunctionalisation (i.e., the duplicated gene will evolve a new beneficial function) can lead to the acquisition of novel biochemical characteristics, while subfunctionalisation (i.e., each copy will adopt parts of the tasks of their parental gene) will have little effect on these characteristics [74]. Similarly, the acquisition of genes or alleles by HGT can result in novel traits (if the acquired genes encode functions not present in the acceptor genome), while gene inactivation or gene loss will usually be accompanied by the loss of specific traits. However, the example of *Bordetella* sp., in which it was shown that many individual traits of these organisms could be due to independent mechanisms of

Fig. 1. Comparison of gene order conservation between 4 genome pairs, illustrating different forms of bacterial genome evolution. All genome sequences were obtained from EBI, and plots were generated using the ADHoRe software [211]. X and Y axes represent the linearised chromosomes. The axes are graduated in 500 genes. Green and red dots indicate pairs of homologous genes that are in the same orientation in both genomes, or in inverted orientation in one relative to the other, respectively. (a) *Buchnera aphidicola* Sg versus *B. aphidicola* Ap gives an example of exceptional genome stability. Their only source of divergence over the last 50 million years is gene loss [48]. (b) *Vibrio cholerae* chromosome I versus *Vibrio parahaemolyticus* chromosome I gives an example of symmetrical genome rearrangements (mainly large inversions) around the axis defined by putative origin and termination of replication, resulting in a X-shaped pattern. (c) *Bordetella bronchiseptica* versus *Bordetella parapertussis* and (d) *B. bronchiseptica* versus *Bordetella pertussis* show recombinations between ISE leading to genome rearrangements. The large expansion of the *IS481* family in *B. pertussis* compared to the presence of ISE in the 2 other *Bordetella* spp. resulted in an accelerated loss of gene order in this lineage [58].

gene inactivation and gene loss [58], demonstrates that the phenotypic outcome of the events that shape the prokaryotic genome are not always as straightforward as expected.

We believe that the ongoing genome sequencing and additional evolutionary studies will allow us to gain more knowledge of how prokaryotic genomes have evolved. This will be the way to provide a definitive natural classification.

4. Novel approaches for assessing taxonomic relationships among prokaryotes based on whole- genome sequences

Although DNA–DNA reassociation methods and sequencing of the 16S rRNA gene are still the cornerstones of present-day bacterial taxonomy, they have several short-comings. Besides problems associated with reproducibility and workability of DNA–DNA hybridisation experiments, DNA reassociation values do not represent actual sequence identity or gene content differences since DNA heteroduplexes will only form between strands that show at least 80% sequence complementarity; therefore a difference of 20% of sequence identity may be spread out between 0% and 100% DNA reassociation [68]. The resolution of 16S rRNA gene sequence analysis between closely related species is generally low, and while organisms with less than 97% 16S rRNA gene sequence identity will generally not give DNA association values of more than 60%, there is no threshold value of 16S rRNA gene sequence identity for species recognition, and there are many examples of cases where two taxa show high 16S rRNA gene sequence identity but low DNA–DNA binding values [75,76]. In addition, considerable within-species 16S rRNA gene sequence diversity has been reported for members of certain bacterial groups (see for example [77]) and there is concern that single-gene trees may not adequately reflect phylogenetic relationships, because of the possibility of HGT and differences in mutation rate. Although a recent study [78] showed that the intragenomic heterogeneity between multiple 16S rRNA operons in sequenced bacterial genomes is rather limited (suggesting limited HGT of 16S rRNA genes), there are several examples in which HGT of (parts of) the 16S rRNA gene has been demonstrated [79–82]. Considerable differences in mutation rates between different lineages have also been observed for 16S rRNA gene sequences, which might lead to tree construction artefacts [83]. Below we give an overview of a number of novel approaches for assessing taxonomic relationships based on whole-genome sequences.

4.1. Gene content

A first approach is to consider the genomes of two organisms as a 'bag of genes' and compare the content of both 'bags' [84]. The identification of orthologous genes is pivotal in this approach and largely depends on the definition of orthology [84,85]. In most studies a minimal definition is used in which putative orthologs are defined as those homologous genes that show the largest identity of several possibilities above a certain threshold [86,87]. Several studies have highlighted the utility of comparing gene content for assessing phylogenetic relationships among prokaryotes [84,86,88–90] and some general trends have been observed. First of all, large genomes have many genes in common and this suggests that it could be useful to normalise the data (i.e., correct for differences in genome size) before in silico analysis. Secondly, the number of genes two genomes have in common depends on their evolutionary distance (Fig. 2) and trees based on gene content generally correspond well with 16S rRNA gene trees (see for example Fig. 3). When using gene content to determine relationships between organisms, it should be taken into account: (i) that the fraction of shared orthologous sequences decreases rapidly in evolution, and (ii) that the evolution of gene content of organisms can have non-tree like aspects (i.e., phylogenetically closely related species do not necessarily share orthologous genes that either of them shares with a phylogenetically more distant species) [84]. Interestingly, it was recently reported that HGT or parallel gene loss does not cause systematic biases in gene content trees [90]. It will be particularly interesting to correlate the fraction of shared genes with measurements of DNA–DNA hybridisation level, but so far such studies have not been performed.

4.2. Gene order

The conservation of gene order can also be used to deduce relationships between organisms. The conservation of local gene order has been documented in several studies, but overall there is very little conservation of overall gene order when the average protein-sequence identity shared by orthologs in two genomes is smaller than 50% and it is clear that the order of orthologous genes is less preserved than their presence [55,84,91]. Due to the high rate of intragenomic rearrangements that break up associations between genes, gene order trees may be especially suitable to resolve the phylogeny of closely related species but in general this method offers a poor resolution on intermediate phylogenetic distances [55,85].

4.3. Comparative sequence analysis of conserved macromolecules

The comparison of sequences of conserved macromolecules like the 16S rRNA gene [92] or housekeeping genes like *recA*, *gyrB* or *rpoD* (see for example [93,94])

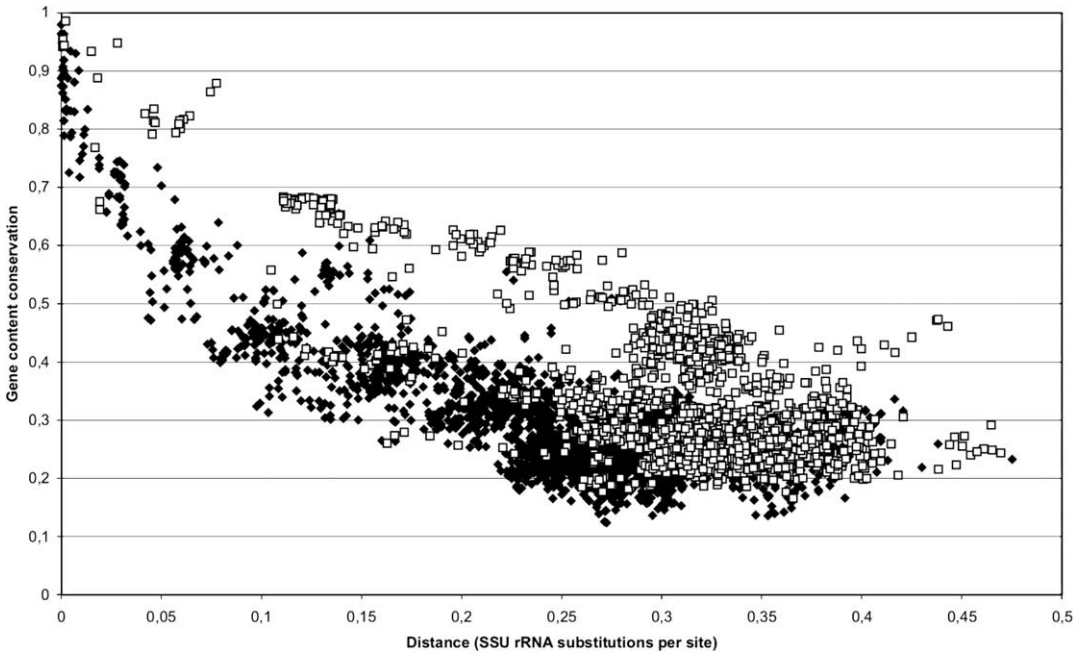


Fig. 2. Gene content conservation versus evolutionary distance for 106 genomes. Each point represents a non-redundant pair of strains. The horizontal axis denotes the average number of nucleic acid substitutions per site for the 16S rRNA and is therefore a representation of evolutionary distance. The vertical axes indicates the fraction of shared genes, calculated by dividing the number of genes two genomes have in common (excluding duplicated genes) by the weighted average genome size, and represents thus a measure for gene content conservation. Genomes included are the same as in [18]. Pairs in which one (or both) genome(s) is an obligate intracellular organism are white, other pairs are black.

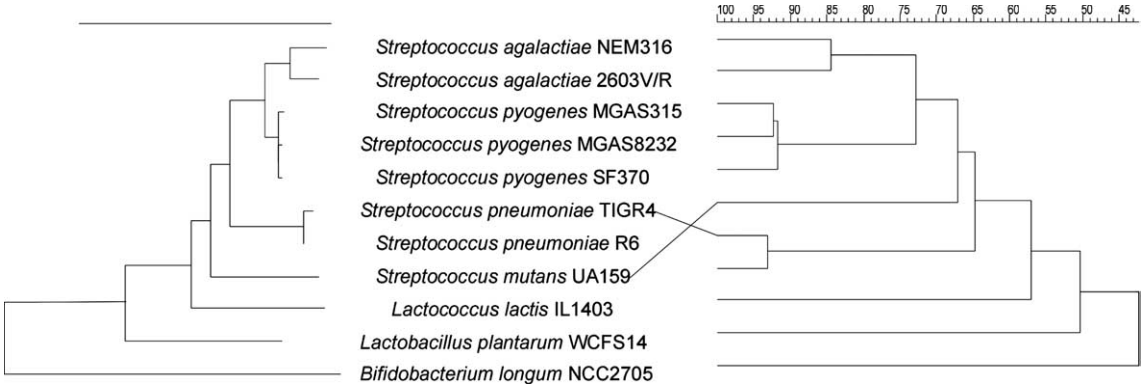


Fig. 3. Comparison of a phylogenetic tree based on 16S rRNA gene sequence similarities (left) and a tree based on the fraction of shared putative orthologs between the genomes of a number of lactic acid bacteria (right). The scale-bar in the tree based on 16S rRNA gene sequence similarities indicates 10% dissimilarity. Data adapted from [87].

have become standard practice in prokaryotic taxonomy during the last few decades. Nevertheless, as mentioned higher, several studies raised concern that single-gene trees may not adequately reflect phylogenetic relationships, because of the possibility of HGT, variable mutation rates and variable rates of recombination. Trees based on large combined alignments of conserved orthologous proteins (so-called 'supertrees' [95]) are mostly highly robust [96] (see Fig. 4 for an example) [97–99]. Yet, based on the alignment of 29 protein families, Teichman and Mitchison [100] found very little phylogenetic signal and concluded that it may be impossible to reconstruct bacterial phylogeny using models based on nucleotide substitution. Philippe and Douady [33] recently have proposed a consensus view, in which genes are subdivided into three categories (see higher). The hard-core of genes that have undergone no detectable HGT can be used to infer the 'true' phylogeny, the soft-core genes (which have undergone few detectable HGTs during their history) can be used to infer phylogeny at small and intermediate evolutionary scale, while the evolutionary history of the shell genes cannot be represented in a bifurcating tree but rather has a network-like appearance. It appears that, although HGT

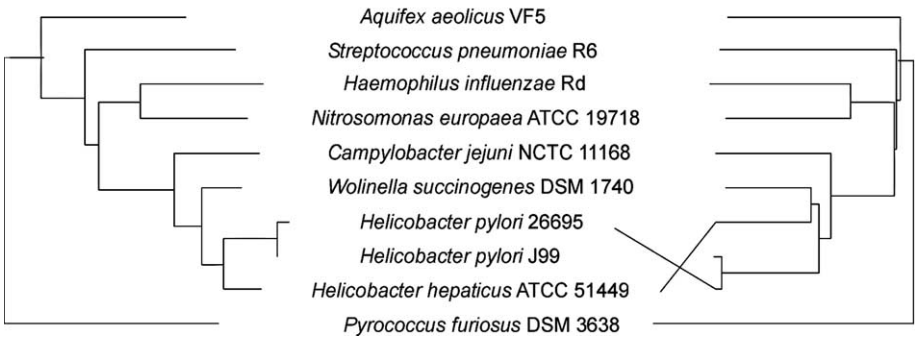


Fig. 4. Comparison of a phylogenetic tree based on 16S rRNA gene sequence similarities (left) and a phylogenetic "supertree" based on the combined amino acid sequences of 20 proteins (right) for a number of prokaryotic taxa. The scale-bar in both trees indicates 10% dissimilarity. Data adapted from [121].

must have played a role in genome diversification, sufficient phylogenetic signal remains present to allow the construction of universal trees. The ad-hoc committee for the re-evaluation of the species definition in bacteriology proposed that a minimum of five housekeeping genes should be sequenced to achieve an adequate informative level of phylogenetic data [2]. Recent studies [101,102] confirmed that sequences from housekeeping genes can accurately predict genome relatedness and can be used for species level identification. Zeigler [102] even suggested that less than five genes might be sufficient to equal or surpass the power of DNA–DNA hybridisations. The selection of candidate genes to be included is of course of extreme importance. Zeigler [102] put forward four a priori criteria: (i) the genes must be widely distributed among genomes,(ii) the genes must be unique within a given genome, (iii) the individual gene sequences must be long enough to contain sufficient information but short enough to allow sequencing in a convenient way (900 – 2250 nt), and (iv) the sequences must predict whole-genome relationships with acceptable precision and accuracy. The latter is based on the finding that not all genes fulfilling the three other criteria contain enough phylogenetic information (i.e., are too conserved), or on the other hand are too variable to be useful for measuring phylogenetic distances between taxa. Applying these criteria to the set of genes from 44 whole genome sequences resulted in 32 useful genes [102]. As additional whole genome sequences became available, this set of genes was re-evaluated for a larger taxonomical sampling, revealing that the majority of these genes do not longer fulfil all the above criteria (Gevers D., unpublished results). The majority of these genes were either not found in most (>95%) genomes, or were found in multiple copies in one or more genomes. Hence, the development of an universal approach (as we know it for 16S rRNA gene sequencing) may be difficult and the development of taxon-specific approaches will be necessary. Similar results were recently reported by Santos and Ochman [103].

Although recombination has long been recognised as a possibly confounding factor for estimating taxonomic relationships among prokaryotes, few studies have addressed this issue in a systematic way. By using statistical tests, Feil et al. [104] showed that there is often extreme noncongruence between different gene trees in frequently recombining species like *Neisseria meningitidis* and *Streptococcus pneumoniae*. By applying phylogenetic reconstruction methods to simulate recombinant sequence alignments, Posada and Crandall [105] showed that the effect of recombination ultimately depends on the relatedness of the sequences involved and the relative size of the regions with different phylogenetic histories: ancient recombinations or recombinations between closely related taxa generally did not significantly influence the recovery of the phylogeny, while recent recombinations that divide the alignment in two regions of similar length resulted in the estimation of a phylogeny different from any of the true phylogenies underlying the data.

4.4. Genome blast distance phylogeny

Henz et al. [106] recently used a novel strategy to derive phylogenies based on the whole genomic information of organisms. This approach, called "genome blast distance phylogeny" (GBDP), starts with an all-against-all pairwise comparison of genomes. Subsequently a distance matrix is calculated from the resulting high scoring pairs. That distance matrix is then processed by a distance-based phylogenetic method to produce a tree or a network. The phylogenetic trees constructed using this approach resembled 16S rRNA gene-based phylogenies, albeit that there were some noteworthy exceptions (e.g., with regard to the position of some deep-branching taxa like *Thermotogales* and *Aquificales*).

4.5. Presence/absence analysis

Relationships between taxa can also be derived from observing presence and/or absence of specific molecular features in their genomes. The presence and absence of families of protein-encoding genes in sequenced genomes have been used to reconstruct the relationships between a number of organisms [107,108]. In this approach, proteins are grouped together in families if their pairwise similarity is greater than a preset value, thereby eliminating the need for the identification of putative orthologs in each genome and the need for specific alignments. Initial results were very similar to results obtained by 16S rRNA gene sequence analysis, although so far only a limited number of taxa have been investigated. Wolf et al. [109] and Lin and Gerstein [110] used the presence or absence of protein folds (protein families that share the same basic molecular shape but not necessarily sequence similarity) to build genome trees. Protein folds are by some considered to be ideal characteristics for building phylogenetic trees as they represent fundamental molecular units used by organisms [110]. So far only few genomes have been investigated this way, but the preliminary data indicate that trees based on presence/absence of protein folds agree fairly well with 16S rRNA gene trees, although some discrepancies observed remain at present unexplained. Another group of molecular features that can be used in presence/absence analysis are conserved insertions and deletions (also called 'indels' or 'signature sequences') in proteins: based on the presence or absence of shared insertions and deletions, taxa can be divided into distinct groups [111,112]. The phylogenetic placement of prokaryotes in different groups based on presence/absence of indels appears to correlate relatively well with 16S rRNA sequence data [111], although some classifications deduced by this approach remain highly controversial.

4.6. Nucleotide composition

While it is well-known that determination of the genomic G + C content lacks resolution for determining taxonomic relationships between taxa, several other biases in nucleotide composition of genomes have been used to determine the relationships of genomes. Dinucleotide relative abundance values are thought to be constant within a genome and it has been hypothesised that this is due to selective pressures that work on them which are constant throughout the genome [113]. Advantages of the analysis of dinucleotide relative abundance values include that it does not depend on finding putative orthologous genes and does not require prior alignment of sequences [113]. In a recent study the relationship between dinucleotide relative dissimilarity (δ^*), 16S rRNA gene sequence identity and levels of DNA–DNA hybridisation was investigated [114]. It was shown that the correlation between the genomic signature and DNA–DNA hybridisation values was high and taxa that showed less than 30% DNA–DNA binding will in general not have δ^* values below 40. On the other hand, taxa showing more than 50% DNA–DNA binding will not have δ^* values higher than 17. The overall correlation between genomic signature and 16S rRNA gene sequence identity appeared to be low, except for closely related organisms (16S rRNA gene identity >94%). Similarly, it was recently shown that relative tetranucleotide frequencies were more similar between closely related organisms than between more distantly related organisms, and that trees based on relative tetranucleotide frequencies were as congruent with 16S rRNA trees as were RpoA and RecA-based trees [115]. Other measures of compositional bias are the effective number of codons used in a genome (N_c) and the G + C content of the synonymous third codon positions ($GC3_s$) [116,117], although recent studies have shown that there appears to be little phylogenetic signal in codon usage [87,118].

4.7. Metabolic pathway reaction content

To increase our understanding of evolutionary relationships between the major domains of life, Podani et al. [119] evaluated the biochemical reaction pathways of 43 species. Comparison of the information transfer pathways of *Archaea* and eukaryotes indicated a close relationship between both domains. More recently, a phylogenetic tree was constructed based on genome-scale metabolic pathway reaction content [11]. Although this approach allowed the separation of *Archaea* from *Bacteria*, the meaning and evolutionary relevance of the subdivisions observed within the *Bacteria* appeared less clear. Phylogenetic trees based on the enzyme and reaction contents of metabolic networks reconstructed from annotated genome information. In more than 80 prokaryotic genomes showed that, although different functional subsystems (i.e., metabolism, cellular processes, and information storage and processing) give different pictures of phylogenetic relationships between organisms, the major results of metabolic network-based phylogenetic trees were in good agreement with trees based on 16S rRNA gene sequences and gene content trees [120].

4.8. Examples and discussion

A number of methods (including gene content, sequences of conserved macromolecules, gene order, dinucleotide relative abundance values and codon usage) were recently used to study phylogenetic relationships between bacterial taxa.

In a first study the taxonomic structure of the lactic acid bacteria was evaluated and it was shown that, for lactic acid bacteria, trees based on information derived from whole-genome sequences were mainly in agreement with trees derived more traditional methods like 16S rRNA gene sequencing [87] (see also Fig. 3). The second study, investigating the relationship of *Aquifex aeolicus* with the ε-*Proteobacteria* showed that it is not straightforward to come to a consistent picture of the phylogenetic position of the *Aquificales*, as different methods often resulted in discordant phylogenies [121]. Teeling et al. [122] used a supertree based on concatenated amino acid sequences of ribosomal proteins and DNA-directed RNA polymerase subunit, and gene content to evaluate the phylogenetic position of the planctomycete *Rhodopirellula baltica*. While most previous studies agreed on the phylogenetic distinctness of the planctomycetes, the exact position of this phylum remained unclear: a distant relationship to the *Chlamydiae* had been proposed based on 16S rRNA gene phylogenies, while a more distinct position as one of the deepest branching phyla within the bacterial domain was suggested on the basis of the comparison of other conserved macromolecules. All trees derived from concatenated protein sequences consistently placed *Rhodopirellula baltica* near the *Chlamydiae*. However, genome trees based on normalised BLASTP scores supported neither a position close to the *Chlamydiae* nor a deep branching position (although it should be noted that this could be due to tree construction artefacts caused by the availability of only a single planctomycete genome sequence at that time). Additional support for a relationship of the *Planctomycetes* and *Chlamydiae* comes from protein signatures [111], trees derived from concatenated sequences of subunits of the F_1F_0-ATPase operon, the presence of proteinaceous cell walls cross-linked via disulphide bonds, the presence of complex cell cycles and an unknown mode of cell division (as both taxa lack *ftsZ*).

It should be noted that most studies assessing taxonomic relationships based on whole-genome sequences have focussed on one or a few methods, and/or on a limited number of taxa, and large-scale systematic comparisons of the various novel methods have not been carried out. However, these preliminary results indicate that there is a stronger phylogenetic signal in some datasets than in others, and that different parameters have different taxonomic information levels [123]. In Fig. 5, we show the taxonomic resolution of some of the techniques mentioned above. It should be kept in mind that more large-scale studies are required to confirm these taxonomic resolutions and that the latter may vary along different phylogenetic lineages. While it may be too soon to thoroughly evaluate the existing methods, and although it is likely that more effective methods will be developed in the future, there appears to be a consensus that whole-genome information will help to confirm or enhance the phylogenetic signal derived from more traditional methods. As stated by Wolf et al. [85], it is clear that genomics did not only bring an extra layer of complexity to molecular evolution but also provided the information that is required to generate a new (and more complete!) picture of the tree of life.

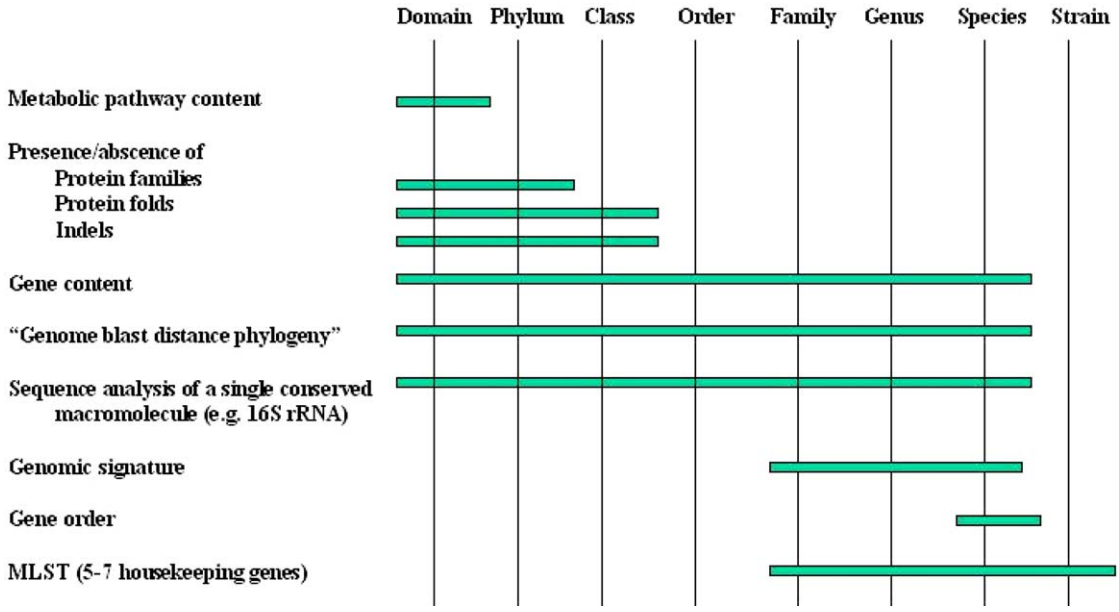


Fig. 5. Taxonomic resolution of some of the novel approaches for assessing taxonomic relationships based on whole-genome sequences. Taxonomic categories listed are domain (e.g., *Bacteria*), phylum (e.g., *Proteobacteria*), class (e.g., β-*Proteobacteria*), order (e.g., *Burkholderiales*), family (e.g., *Burkholderiaceaea*), genus (e.g., *Burkholderia*), species (e.g., *Burkholderia cepacia*) and strain. The taxonomic resolution was derived from the data presented in the references cited in the text and may therefore be considered as slightly speculative as further confirmation of these resolutions is required.

5. Comparative microbial genomics using DNA microarrays and subtractive hybridisation

Besides by directly comparing the full genome sequences, the differences between related micro-organisms can also be studied using DNA microarrays or subtractive hybridisation. These methods have the advantage that they do not require the availability of a whole-genome sequence of all organisms being studied. Hybridisation of DNA to whole-genome microarrays have been used to study the genetic diversity of a wide range of bacteria (see [124,125] for recent reviews). The microarray technology has also been used for a number of specialised applications. Cho and Tiedje [126] proposed a new approach to identify bacteria based on genomic DNA–DNA similarity employing microarray technology. This method (so far only evaluated with four *Pseudomonas* species) does not require laborious cross-hybridisations and the resulting hybridisation profiles can be used in statistical procedures to identify test strains and can be stored in a database. The authors stated that they expected that up to 20% of size differences between two genomes (like in *E. coli*) would still result in meaningful results, although the obtained hybridisation values would most likely be slightly higher than the values obtained from conventional whole-genome DNA–DNA hybridisation. Microarrays for the identification of specific (pathogenic) bacteria were also developed. These include arrays for the identification of *Staphylococcus* sp. [127], *Listeria* sp. [128], *E. coli*, *Shigella* sp. and *Salmonella* sp. [129], *Mycobacterium* sp. [130] and *Campylobacter* sp. [131]. Microarrays have also been developed to study specific bacterial populations and consortia, like sulfate-reducing bacteria [132] and toluene- and ethylbenzene-degrading consortia [133]. van Leeuwen et al. [134] used high-density DNA arrays for multilocus sequence typing (MLST) of *Staphylococcus aureus*. PCR-based subtractive hybridisation [135] also allows the identification of polymorphic DNA fragments specific for certain bacterial strains or species. An overview of applications of subtractive hybridisations is given by Winstanley [136] and includes identification of insertion sequences, transposable elements and genomic islands implicated in virulence, and identification of strain- or species-specific sequences for the development of diagnostic probes.

6. Detection of intraspecies diversity

Although the complete genome sequence of multiple strains of several species have been determined (see for example Table 1), it is unlikely that multiple genome sequences will become available for many species in the near future. Nevertheless, there is an urgent need to assess the intraspecies diversity from a genomic point of view and classify many strains per species based on genomic criteria. At present, two methods seem particularly suitable to assess this intraspecies diversity: MLST [6] and microarray hybridisations.

The indexing of sequence variation in multiple housekeeping genes by means of MLST has become increasingly popular during the last five years. In MLST the proven concepts of multilocus enzyme electrophoresis (MLEE) [137] were adapted to allow the identification of different alleles directly from nucleotide sequence of internal fragments of PCR-amplified housekeeping genes [138,139]. However, direct sequencing of housekeeping genes allows the detection of much more variation than was possible using MLEE. In addition, the portability of sequence data allows the construction of central databases (see for example [140,141]) in which data can be stored and accessed through the Internet. Since the introduction of MLST, numerous studies on a wide range of organisms have been published, including *N. meningitidis* [138], *Helicobacter pylori* [142], *S. pneumoniae* [143], *Yersinia pestis* [144], *Vibrio cholerae* [145], *S. aureus* [146], *Streptococcus pyogenes* [147], *Campylobacter jejuni* [148], *Enterococcus faecalis* [149], *Haemophilus influenzae* [150], *Bacillus cereus* [151], *Burkholderia mallei* and *Burkholderia pseudomallei* [152], and *Listeria monocytogenes* [153]. In addition, several alternatives to MLST, including multilocus restriction typing [154–157] and microarray-based MLST [134] have recently been developed as well. Besides the indexing of intraspecies variation, MLST allows to quantify the extent of genetic recombination between clones in a bacterial population and thus allows to make inferences regarding the structure of the population [158–161]. Recent studies have shown that bacterial population structures range from the extremes of strictly clonal (no recombination) to panmictic, with most populations occupying a middle ground where recombination is significant in the evolution of the population but not frequent enough to prevent the emergence of epidemic clonal lineages (for reviews see [158,160,162]). Although there are now a multitude of studies that show that in several species (including *N. meningitidis* and *S. pneumoniae* [163–165]) evolutionary change at housekeeping loci is more likely to occur by recombination than by mutation, the precise mechanisms by which recombination occurs and the underlying reasons for the variation in recombination rates are still largely unclear [105].

Other MLST applications include the study of differences in host preference or virulence potential between different clones or clonal complexes, monitoring the effects of vaccination programs and revealing evidence concerning the emergence and spread of antibiotic resistant clones [161].

The microarray hybridisation technology provides a second novel tool to assess intraspecies diversity. As is obvious from Table 1, significant differences in gene

content and genome size occur between multiple strains from the same species, and, as stated higher, the microarray technology can be used to detect these differences, without the need for sequencing additional genomes. This approach was already successfully applied to detect genomic diversity among multiple strains of *C. jejuni* [166], *H. pylori* [167], *Salmonella* sp. [168], *V. cholerae* [169], *S. pneumoniae* [170], *Mb. tuberculosis* [171] and *E. coli* [172]. The results of these and other studies confirm that there are significant differences in intraspecies diversity. Species like *C. jejuni*, *H. pylori*, *S. pneumoniae* and *E. coli* appear to have high intraspecies diversity (with 21%, 22%, 10% and 10% of the genes, respectively, being altered or deleted in one or more strains compared to the sequenced reference strain). Strain-specific genes often include putative virulence factors or factors involved in interaction with the host and mobile genetic elements, while the conserved genes encode most metabolic and cellular processes. Other species, like *V. cholerae*, revealed a high degree of conservation among the strains tested, with only approximately 1% difference between the genomes of most test strains and the sequenced strain. While these results are of course dependent on the strains included, they clearly show that a good picture of intraspecies diversity and intraspecies differences in gene content can be obtained without sequencing multiple strains of the same organism.

Besides microarray hybridisations and MLST, other methods relying heavily on information derived from whole-genome sequences have become available. In certain genetically homogeneous species like *Bacillus anthracis* or *Mb. tuberculosis*, the low level of nucleotide variation prevents the use of classical MLST schemes [161]. To assess the intraspecies diversity of these homogeneous species, alternative genomic methods have been developed. In variable number tandem repeat (VNTR) analysis, the variation in copy number of short nucleotide sequences that are repeated multiple times (resulting in length polymorphisms), is detected by PCR [173–175]. The resolution of this technique can even further be improved by simultaneously studying multiple loci (multiple-locus VNTR analysis, MLVA) [175–177]. It should be noted that this method has also been used to type strains of more variable species, like *Pseudomonas aeruginosa* [178]. A second method is based on the identification of single nucleotide polymorphism (SNP) markers for assessing intraspecies diversity. This method has proven to be particularly useful for studying the evolution and epidemiology of *Mb. tuberculosis* [123,179]. Recently, query software that allows identification of highly informative sets of SNP markers in entire MLST databases was developed [180].

As a final comment, we would like to point out that the more conventional methods for assessing bacterial intraspecies diversity (including various restriction-based and/or PCR-based fingerprint techniques) have also taken advantage of the increased availability of whole-genome sequences, as several software applications have been developed that allow to simulate these experiments in silico first [181,182]. This way a more optimal selection of the subsequent experimental conditions becomes possible.

7. Implications for the prokaryotic species concept

The species concept for prokaryotes is fundamentally different from the species concept for eukaryotes. The most popular eukaryotic species concept, the biological species concept (pioneered by Ernst Mayr), considered species to be groups of actually or potentially interbreeding natural populations which are reproductively isolated from other such groups but there now appears to be a consensus that, for prokaryotes, this concept should be abandoned in favour of an evolutionary or at least a phylogenetic species concept [68,183]. As mentioned higher, a prokaryotic species is "a category that circumscribes a (preferably) genomically coherent group of individual isolates/strains sharing a high degree of similarity in (many) independent features, comparatively tested under highly standardised conditions" [2] However, alternative prokaryotic species concepts have been proposed and the circumscriptions of species may vary depending on the species definition adopted [68,184]. It is generally accepted that prokaryotic species should be delineated after analysis and comparison of many parameters in what is known to be polyphasic taxonomy [1]. Although this practical approach to the prokaryotic species definition has been the subject of considerable debate and prokaryotic taxonomists have been urged to adopt a more natural species concept (see for example [185,186]), it has been found acceptable by many. Nevertheless, a more natural species concept would be welcomed and could be based on approaches now commonly used in studies on bacterial population structure, including the indexing of sequence variation in multiple housekeeping genes by means of MLST [6,25,138,139,187]. Although the patterns and amounts of genetic exchange are considerably different between eukaryotes and prokaryotes, their ecological diversity is organised into the same way, as individual organisms fall into more or less discrete clusters on the basis of a number of characteristics (e.g., DNA sequences) [25]. The prokaryotic counterpart of these eukaryote species seem to correspond to so-called 'ecotypes' and named species appear to contain many of these ecotypes [6,25]. However, using the ecotype as fundamental unit in a prokaryotic species concept would have far reaching consequences. As stated by Rosello-Mora [184], "the reductionist approach of giving genetics precedence in taxonomy conflicts with the purpose of classification, i.e., that of being a predictive and widely applicable

system''. The implementation of a prokaryotic taxonomy oriented more towards genomics should by no means implicate an orientation towards a less-applicable taxonomy.

It is at present difficult to assess the future impact of the availability of a large number of complete genome sequences on bacterial taxonomy and the prokaryotic species concept but it can be anticipated that novel insights into the current classification scheme will arise as the increasing number of available genome sequences will allow the bacterial taxonomists to evaluate the potential of many of the above-described approaches to deduce relationships between taxa. For example, it can be expected that the availability of a large number of completely-sequenced genomes will highly facilitate the development of more universal multilocus sequence analysis schemes, which in turn could allow us to adopt a more natural species concept. In addition, it will be possible to measure the extent of 'prokaryotic sex', as HGT has been called, and to better assess its potential implications for the prokaryotic species concept. However, as pointed out by Rosselo-Mora and Amann [68], it should be clear that the practical prokaryotic species concept as currently conceived has not much to fear from the problem of HGT, as it is based on whole genome similarities: genomic rearrangements (including HGT) will not affect the primary structure of the majority of genes and changing the physical map will not markedly influence DNA–DNA hybridisation levels.

8. Some concluding remarks

Several issues regarding sequencing of complete bacterial genomes remain at present unresolved. While some of these may appear as of minor importance to most, they may be crucial for taxonomists. A first important issue is that the organisms that have been sequenced are by no means representative for the total prokaryotic diversity, and, so far, have been biased towards organisms of medical or biotechnological importance. In addition, it is estimated that more than 99% of all micro-organisms observable in nature can not be cultivated using standard techniques [188,189] and this makes it very hard—if not impossible—to include them in standard polyphasic-taxonomic studies. However, the advances in methodology for isolating and sequencing DNA now allow the sequencing of the genomes of many of these fastidious organisms (see for example [190]) and it can be expected that much will be learned regarding their taxonomy and evolutionary history. The availability of enormous sequencing capabilities also makes it possible to embark on environmental genomics (metagenomics) studies, in which large fragments of environmental DNA are cloned and subsequently sequenced [191]. Shotgun sequencing of cloned environmental DNA fragments (including soil [192] and sea water [193]) has confirmed and extended previous observations [188,189] that the vast majority of microbial organisms has not been cultured and characterised yet. In a recent study, Venter et al. [193] sequenced DNA from microbial populations collected in the Sargasso Sea. A total of 1.045 billion base pairs of non-redundant sequence was generated, and subsequent analyses showed that the sequences were derived from at least 1800 genomic species, including 148 novel bacterial phylotypes. This clearly shows that, in order to get a better picture of microbial diversity, taxonomy and evolution, more attention needs to be devoted to the study of these uncultured organisms. It should be noted that genomic information can also provide new insights that can be used to design novel axenic media for fastidious and so-far uncultured pathogens, providing the opportunity to study them in more detail *in vitro* [194,195].

Our understanding of bacterial intraspecies diversity and population structure has benefited significantly from the introduction of mathematical methods like split decomposition [196] and eBURST [197], that allow to visualise relationships between taxa as networks rather than as bifurcating trees (see for example [198–201]). Other novel mathematical approaches to visualise the phylogenetic content of genomes include self-organising maps (SOMs) [202], Neighbor-Net [203] and dekapentagonal maps [204]. It can be anticipated that the application of these and other novel mathematical methods to data derived from whole-genome sequences will lead to novel insights regarding phylogenetic relationships, intraspecies diversity, population structures and frequency of recombination and HGT. Similarly, advances in methodology of combining many smaller, overlapping phylogenetic trees into a single 'supertree' will allow to construct more comprehensive phylogenetic hypotheses [95,205,206].

Another issue is that of the value of fully finished genome sequences as opposed to sequences in various levels of draft. While it is obvious that the costs of draft sequencing are lower than those of complete genome sequencing, the extent of resources that could be saved by switching to draft sequencing and the amount of information that would be lost by doing so is the subject of considerable debate [207,208]. It remains to be determined what approach will appeal most to the majority of bacterial taxonomists and funding agencies. A final issue that would need to be resolved is the availability of sequences and sequenced strains. Given the amount of efforts and money spent, all sequenced strains should be saved from extinction by depositing them in internationally recognised culture collections [209,210]. Similarly, the scientific community can only benefit from the sequencing efforts if sequences are deposited in publicly available databases within a reasonable timeframe following completion.

It is still too early to speculate on how the different new genomic data will be used in the developing genomic taxonomy. It is clear that a universal, portable, taxonomic system for culturable and non-culturable micro-organisms will need to be developed further in the coming years. But at least, the roadmap towards a genomic taxonomy of prokaryotes is now under construction.

Acknowledgements

T.C., P.V. and J.S. are indebted to the Fund for Scientific Research – Flanders (Belgium) for a position as postdoctoral fellow and research grants, respectively. T.C. also acknowledges the support from the Belgian Federal Government (Federal Office for Scientific, Technical and Cultural Affairs). The BOF (Bijzonder Onderzoeksfonds – UGent) is acknowledged by D.G. for support in the framework of project no. 01110803.

References

[1] Vandamme, P., Pot, B., Gillis, M., Devos, P., Kersters, K. and Swings, J. (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol. Rev. 60, 407–438.
[2] Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A., Kampfer, P., Maiden, M.C., Nesme, X., Rossello-Mora, R., Swings, J., Truper, H.G., Vauterin, L., Ward, A.C. and Whitman, W.B. (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 52, 1043–1047.
[3] Ursing, J.B., Rossellomora, R.A., Garciavaldes, E. and Lalucat, J. (1995) Taxonomic note – a pragmatic approach to the nomenclature of phenotypically similar genomic groups. Int. J. Syst. Bacteriol. 45, 604.
[4] Overbeek, R. (2000) Genomics: what is realistically achievable. Genome Biol. 1, 2.
[5] Janssen, P., Audit, B., Cases, I., Darzentas, N., Goldovsky, L., Kunin, V., Lopez-Bigas, N., Peregrin-Alvarez, J.M., Pereira-Leal, J.B., Tsoka, S. and Ouzounis, C.A. (2003) Beyond 100 genomes. Genome Biol. 4, 402.
[6] Cohan, F.M. (2002) What are bacterial species. Annu. Rev. Microbiol. 56, 457–487.
[7] Lan, R. and Reeves, P.R. (2000) Intraspecies variation in bacterial genomes: the need for a species genome concept. Trends Microbiol. 8, 396–401.
[8] Dobrindt, U. and Hacker, J. (2001) Whole genome plasticity in pathogenic bacteria. Curr. Opin. Microbiol. 4, 550–557.
[9] Snel, B., Bork, P. and Huynen, M.A. (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res. 12, 17–25.
[10] Kunin, V. and Ouzounis, C.A. (2003) The balance of driving forces during genome evolution in prokaryotes. Genome Res. 13, 1589–1594.
[11] Hong, S.H., Kim, T.Y. and Lee, S.Y. (2004) Phylogenetic analysis based on genome-scale metabolic pathway reaction content. Appl. Microbiol. Biotechnol. 65, 203–210.
[12] Andersson, S.G.E. (2000) The genomics gamble. Nat. Genet. 26, 134–135.
[13] Ohno, S. (1970) Evolution by Genome Duplication. Springer, Berlin.
[14] Hooper, S.D. and Berg, O.G. (2003) On the nature of gene innovation: duplication patterns in microbial genomes. Mol. Biol. Evol. 20, 945–954.
[15] Jordan, I.K., Makarova, K.S., Spouge, J.L., Wolf, Y.I. and Koonin, E.V. (2001) Lineage-specific gene expansions in Bacterial and Archaeal genomes. Genome Res. 11, 555–565.
[16] Kunisawa, T. (1995) Identification and chromosomal distribution of DNA sequence segments conserved since divergence of *Escherichia coli* and *Bacillus subtilis*. J. Mol. Evol. 40, 585–593.
[17] Kolsto, A.B. (1997) Dynamic bacterial genome organization. Mol. Microbiol. 24, 241–248.
[18] Gevers, D., Vandepoele, K., Simillion, C. and Van de Peer, Y. (2004) Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. 12, 148–155.
[19] Wolfe, K.H. (2001) Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. 2, 333–341.
[20] Simillion, C., Vandepoele, K., Van Montagu, M.C.E., Zabeau, M. and Van De Peer, Y. (2002) The hidden duplication past of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA 99, 13627–13632.
[21] Vandepoele, K., De Vos, W., Taylor, J.S., Meyer, A. and Van de Peer, Y. (2004) Major events in the genome evolution of vertebrates: paranome age and size differs considerably between fishes and land vertebrates. Proc. Natl. Acad. Sci. USA 101, 1638–1643.
[22] Tekaia, F. and Dujon, B. (1999) Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. J. Mol. Evol. 49, 591–600.
[23] Tekaia, F., Gordon, S.V., Garnier, T., Brosch, R., Barrell, B.G. and Cole, S.T. (1999) Analysis of the proteome of *Mycobacterium tuberculosis* in silico. Tuber Lung Dis. 79, 329–342.
[24] Ochman, H., Lawrence, J.G. and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405, 299–304.
[25] Cohan, F.M. (2001) Bacterial species and speciation. Syst. Biol. 50, 513–524.
[26] Lawrence, J.G. and Ochman, H. (1998) Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. USA 95, 9413–9417.
[27] Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., McDonald, L., Utterback, T.R., Malek, J.A., Linher, K.D., Garrett, M.M., Stewart, A.M., Cotton, M.D., Pratt, M.S., Phillips, C.A., Richardson, D., Heidelberg, J., Sutton, G.G., Fleischmann, R.D., Eisen, J.A. and Fraser, C.M., et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399, 323–329.
[28] Garcia-Vallve, S., Romeu, A. and Palau, J. (2000) Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. Mol. Biol. Evol. 17, 352–361.
[29] Pennisi, E. (2004). Science 305, 334–335.
[30] Koonin, E.V. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat. Rev. Microbiol. 1, 127–136.
[31] Woese, C.R. (2000) Interpreting the universal phylogenetic tree. Proc. Natl. Acad. Sci. USA 97, 8392–8396.
[32] Gogarten, J.P., Doolittle, W.F. and Lawrence, J.G. (2002) Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. 19, 2226–2238.
[33] Philippe, H. and Douady, C.J. (2003) Horizontal gene transfer and phylogenetics. Curr. Opin. Microbiol. 6, 1–8.
[34] Sorensen, S.J., Sorensen, A.H., Hansen, L.H., Oregaard, G. and Veal, D. (2003) Direct detection and quantification of horizontal gene transfer by using flow cytometry and gfp as a reporter gene. Curr. Microbiol. 47, 129–133.

[35] Mirkin, B.G., Fenner, T.I., Galperin, M.Y. and Koonin, E.V. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol. Biol. 3, 2.
[36] Hacker, J. and Kaper, J.B. (2000) Pathogenicity islands and the evolution of microbes. Annu. Rev. Microbiol. 54, 641–679.
[37] Lawrence, J.G., Hendrix, R.W. and Casjens, S. (2001) Where are the pseudogenes in bacterial genomes. Trends Microbiol. 9, 535–540.
[38] Mira, A., Ochman, H. and Moran, N.A. (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet. 17, 589–596.
[39] Maurelli, A.T., Fernandez, R.E., Bloch, C.A., Rode, C.K. and Fasano, A. (1998) Black holes and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. Proc. Natl. Acad. Sci. USA 95, 3943–3948.
[40] Andersson, J.O. and Andersson, S.G. (2001) Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. Mol. Biol. Evol. 18, 829–839.
[41] Frank, A.C., Amiri, H. and Andersson, S.G. (2002) Genome deterioration: loss of repeated sequences and accumulation of junk DNA. Genetica 115, 1–12.
[42] Klasson, L. and Andersson, S.G.E. (2004) Evolution of minimal-gene-sets in host-dependent bacteria. Trends Microbiol. 12, 37–43.
[43] Andersson, S.G. and Kurland, C.G. (1998) Reductive evolution of resident genomes. Trends Microbiol. 6, 263–268.
[44] Moran, N.A. (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108, 583–586.
[45] Koonin, E.V. (2003) Horizontal gene transfer: the path to maturity. Mol. Microbiol. 50, 725–727.
[46] Andersson, J.O. and Andersson, S.G. (1999) Insights into the evolutionary process of genome degradation. Curr. Opin. Genet. Dev. 9, 664–671.
[47] Moran, N.A. (2003) Tracing the evolution of gene loss in obligate bacterial symbionts. Curr. Opin. Microbiol. 6, 512–518.
[48] Tamas, I., Klasson, L., Canback, B., Naslund, A.K., Eriksson, A.S., Wernegreen, J.J., Sandstrom, J.P., Moran, N.A. and Andersson, S.G. (2002) 50 million years of genomic stasis in endosymbiotic bacteria. Science 296, 2376–2379.
[49] van Ham, R.C., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernandez, J.M., Jimenez, L., Postigo, M., Silva, F.J., Tamames, J., Viguera, E., Latorre, A., Valencia, A., Moran, F. and Moya, A. (2003) Reductive genome evolution in *Buchnera aphidicola*. Proc. Natl. Acad. Sci. USA 100, 581–586.
[50] McLeod, M.P., Qin, X., Karpathy, S.E., Gioia, J., Highlander, S.K., Fox, G.E., McNeill, T.Z., Jiang, H., Muzny, D., Jacob, L.S., Hawes, A.C., Sodergren, E., Gill, R., Hume, J., Morgan, M., Fan, G., Amin, A.G., Gibbs, R.A., Hong, C., Yu, X., Walker, D.H. and Weinstock, G.M. (2004) The complete genome sequence of *Rickettsia typhi* and comparison with other rickettsiae. J. Bacteriol. 186, 5842–5855.
[51] Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honore, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R.M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quail, M.A., Rajandream, M.A., Rutherford, K.M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J.R. and Barrell, B.G. (2001) Massive gene decay in the leprosy bacillus. Nature 409, 1007–1011.
[52] Mushegian, A.R. and Koonin, E.V. (1996) Gene order is not conserved in bacterial evolution. Trends Genet. 12, 289–290.
[53] Tamames, J. (2001) Evolution of gene order conservation in prokaryotes. Genome Biol. 2, RESEARCH0020.
[54] Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S. and Koonin, E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res. 11, 356–372.
[55] Suyama, M. and Bork, P. (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. 17, 10–13.
[56] Korbel, J.O., Snel, B., Huynen, M.A. and Bork, P. (2002) SHOT: a web server for the construction of genome phylogenies. Trends Genet. 18, 158–162.
[57] Rocha, E.P. (2003) DNA repeats lead to the accelerated loss of gene order in bacteria. Trends Genet. 19, 600–603.
[58] Parkhill, J., Sebaihia, M., Preston, A., Murphy, L.D., Thomson, N., Harris, D.E., Holden, M.T., Churcher, C.M., Bentley, S.D., Mungall, K.L., Cerdeno-Tarraga, A.M., Temple, L., James, K., Harris, B., Quail, M.A., Achtman, M., Atkin, R., Baker, S., Basham, D., Bason, N., Cherevach, I., Chillingworth, T., Collins, M., Cronin, A., Davis, P., Doggett, J., Feltwell, T., Goble, A., Hamlin, N., Hauser, H., Holroyd, S., Jagels, K., Leather, S., Moule, S., Norberczak, H., O'Neil, S., Ormond, D., Price, C., Rabbinowitsch, E., Rutter, S., Sanders, M., Saunders, D., Seeger, K., Sharp, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Unwin, L., Whitehead, S., Barrell, B.G. and Maskell, D.J. (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat. Genet. 35, 32–40.
[59] Tillier, E.R.M. and Collins, R.A. (2000) Genome rearrangement by replication-directed translocation. Nat. Genet. 26, 195–197.
[60] Eisen, J.A., Heidelberg, J.F., White, O. and Salzberg, S.L. (2000) Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol. 1, RESEARCH0011.
[61] Mackiewicz, P., Mackiewicz, D., Kowalczuk, M. and Cebrat, S. (2001) Flip-flop around the origin and terminus of replication in prokaryotic genomes. Genome Biol. 2, INTERACTIONS1004.
[62] Boussau, B., Karlberg, E.O., Frank, A.C., Legault, B.A. and Andersson, S.G. (2004) Computational inference of scenarios for alpha-proteobacterial genome evolution. Proc. Natl. Acad. Sci. USA 101, 9722–9727.
[63] Arber, W. (2000) Genetic variation: molecular mechanisms and impact on microbial evolution. FEMS Microbiol. Rev. 24, 1–7.
[64] Hall, B.G. and Malik, H.S. (1998) Determining the evolutionary potential of a gene. Mol. Biol. Evol. 15, 1055–1061.
[65] Lawrence, J.G. (1999) Gene transfer, speciation, and the evolution of bacterial genomes. Curr. Opin. Microbiol. 2, 519–523.
[66] Bapteste, E., Boucher, Y., Leigh, J. and Doolittle, W.F. (2004) Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol. 12, 406–411.
[67] Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. 44, 383–397.
[68] Rossello-Mora, R. and Amann, R. (2001) The species concept for prokaryotes. FEMS Microbiol. Rev. 25, 39–67.
[69] Olive, D.M. and Bean, P. (1999) Principles and applications of methods for DNA-based typing of microbial organisms. J. Clin. Microbiol. 37, 1661–1669.
[70] Van Belkum, A. (2003) High-throughput epidemiologic typing in clinical microbiology. Clin. Microbiol. Infect. 9, 86–100.
[71] Van Belkum, A., Struelens, M., De Visser, A., Verbrugh, H. and Tibayrenc, M. (2001) Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. Clin. Microbiol. Rev. 14, 547–560.

[72] Tenover, F.C., Arbeit, R.D., Goering, R.V., Mickelsen, P.A., Murray, B.E., Persing, D.H. and Swaminathan, B. (1995) Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel-electrophoresis – criteria for bacterial strain typing. J. Clin. Microbiol. 33, 2233–2239.
[73] Gurtler, V. and Mayall, B.C. (2001) Genomic approaches to typing, taxonomy and evolution of bacterial isolates. Int. J. Syst. Evol. Microbiol. 51, 3–16.
[74] Lynch, M., O'hely, M., Walsh, B. and Force, A. (2001) The probability of preservation of a newly arisen gene duplicate. Genetics 159, 1789–1804.
[75] Fox, G.E., Wisotzkey, J.D. and Jurtshuk, P. (1992) How close is close – 16S ribosomal-RNA sequence identity may not be sufficient to guarantee species identity. Int. J. Syst. Bacteriol. 42, 166–170.
[76] Stackebrandt, E. and Goebel, B.M. (1994) A place for DNA–DNA reassociation and 16S ribosomal-RNA sequence-analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. 44, 846–849.
[77] Harrington, C.S. and On, S.L.W. (1999) Extensive 16S rRNA gene sequence diversity in *Campylobacter hyointestinalis* strains: taxonomic and applied implications. Int. J. Syst. Bacteriol. 49, 1171–1175.
[78] Coenye, T. and Vandamme, P. (2003) Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. FEMS Microbiol. Lett. 228, 45–49.
[79] Ueda, K., Seki, T., Kudo, T., Yoshida, T. and Kataoka, M. (1999) Two distinct mechanisms cause heterogeneity of 16S rRNA. J. Bacteriol. 181, 78–82.
[80] Yap, W.H., Zhang, Z.S. and Wang, Y. (1999) Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. J. Bacteriol. 181, 5201–5209.
[81] Van Berkum, P., Terefework, Z., Paulin, L., Suomalainen, S., Lindstrom, K. and Eardly, B.D. (2003) Discordant phylogenies within the rrn loci of *Rhizobia*. J. Bacteriol. 185, 2988–2998.
[82] Schouls, L.M., Schot, C.S. and Jacobs, J.A. (2003) Horizontal transfer of segments of the 16S rRNA genes between species of the *Streptococcus anginosus* group. J. Bacteriol. 185, 7241–7246.
[83] Woese, C.R., Stackebrandt, E. and Ludwig, W. (1985) What are mycoplasmas – the relationship of tempo and mode in bacterial evolution. J. Mol. Evol. 21, 305–316.
[84] Huynen, M.A. and Bork, P. (1998) Measuring genome evolution. Proc. Natl. Acad. Sci. USA 95, 5849–5856.
[85] Wolf, Y.I., Rogozin, I.B., Grishin, N.V. and Koonin, E.V. (2002) Genome trees and the tree of life. Trends Genet. 18, 472–479.
[86] Bansal, A.K. and Meyer, T.E. (2002) Evolutionary analysis by whole genome comparisons. J. Bacteriol. 184, 2260–2272.
[87] Coenye, T. and Vandamme, P. (2003) Extracting phylogenetic information from whole-genome sequencing projects: the lactic acid bacteria as a test case. Microbiology-Sgm 149, 3507–3517.
[88] Tekaia, F., Lazcano, A. and Dujon, B. (1999) The genomic tree as revealed from whole proteome comparisons. Genome Res. 9, 550–557.
[89] Snel, B., Bork, P. and Huynen, M.A. (1999) Genome phylogeny based on gene content. Nat. Genet. 21, 108–110.
[90] Dutilh, B.E., Huynen, M.A., Bruno, W.J. and Snel, B. (2004) The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J. Mol. Evol. 58, 527–539.
[91] Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. 23, 324–328.
[92] Garrity, G.M. and Holt, J.G. (2001) The road map to the manual (Boone, D.R. and Castenholz, R.W., Eds.), Bergey's Manual of Systematic Bacteriology, Vol. 1. Springer, New York.
[93] Eisen, J.A. (1995) The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16s rRNAs from the same species. J. Mol. Evol. 41, 1105–1123.
[94] Yamamoto, S., Kasai, H., Arnold, D.L., Jackson, R.W., Vivian, A. and Harayama, S. (2000) Phylogeny of the genus pseudomonas: intrageneric structure reconstructed from the nucleotide sequences of gyrB and rpoD genes. Microbiology-Sgm 146, 2385–2394.
[95] Bininda-Emonds, O.R.P. (2004) The evolution of supertrees. Trends Ecol. Evol. 19, 315–322.
[96] Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E. and Stanhope, M.J. (2001) Universal trees based on large combined protein sequence data sets. Nat. Genet. 28, 281–285.
[97] Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Tatusov, R.L. and Koonin, E.V. (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol. Biol. 1, 8.
[98] Daubin, V., Gout, M. and Perriere, G. (2001) Bacterial molecular phylogeny using supertree approach. Genome Inf. 12, 155–164.
[99] Lerat, E., Alves, L.M. and Campanharo, J.C. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-*Proteobacteria*. PLoS Biol. 1, E19.
[100] Teichmann, S.A. and Mitchison, G. (1999) Is there a phylogenetic signal in prokaryote proteins. J. Mol. Evol. 49, 98–107.
[101] Wertz, J.E., Goldstone, C., Gordon, D.M. and Riley, M.A. (2003) A molecular phylogeny of enteric bacteria and implications for a bacterial species concept. J. Evol. Biol. 16, 1236–1248.
[102] Zeigler, D.R. (2003) Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int. J. Syst. Evol. Microbiol. 53, 1893–1900.
[103] Santos, S.R. and Ochman, H. (2004) Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. Environ. Microbiol. 6, 754–759.
[104] Feil, E.J., Holmes, E.C., Bessen, D.E., Chan, M.S., Day, N.P.J., Enright, M.C., Goldstein, R., Hood, D.W., Kalla, A., Moore, C.E., Zhou, J.J. and Spratt, B.G. (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. USA 98, 182–187.
[105] Posada, D. and Crandall, K.A. (2002) The effect of recombination on the accuracy of phylogeny estimation. J. Mol. Evol. 54, 396–402.
[106] Henz, S.R., Huson, D.H., Auch, A.F., Nieselt-Struwe, K. and Schuster, S.C. (2004) Whole-genome prokaryotic phylogeny. Bioinformatics epub.
[107] Fitz-Gibbon, S.T. and House, C.H. (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. 27, 4218–4222.
[108] House, C.H. and Fitz-Gibbon, S.T. (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. J. Mol. Evol. 54, 539–547.
[109] Wolf, Y.I., Brenner, S.E., Bash, P.A. and Koonin, E.V. (1999) Distribution of protein folds in the three superkingdoms of life. Genome Res. 9, 17–26.
[110] Lin, J. and Gerstein, M. (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res. 10, 808–818.
[111] Gupta, R.S. (2001) The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. Int. Microbiol. 4, 187–202.
[112] Gupta, R.S. and Griffiths, E. (2002) Critical issues in bacterial phylogeny. Theor. Popul. Biol. 61, 423–434.

[113] Karlin, S., Mrazek, J. and Campbell, A.M. (1997) Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. 179, 3899–3913.
[114] Coenye, T. and Vandamme, P. (2004) Use of the genomic signature in bacterial classification and identification. Syst. Appl. Microbiol. 27, 175–185.
[115] Pride, D.T., Meinersmann, R.J., Wassenaar, T.M. and Blaser, M.J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res. 13, 145–158.
[116] Wright, F. (1990) The effective number of codons used in a gene. Gene 87, 23–29.
[117] Andersson, S.G.E. and Sharp, P.M. (1996) Codon usage in the *Mycobacterium tuberculosis* complex. Microbiology-Uk 142, 915–925.
[118] Chen, L.L. and Zhang, C.T. (2003) Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages. Biochem. Biophys. Res. Commun. 306, 310–317.
[119] Podani, J., Oltvai, Z.N., Jeong, H., Tombor, B., Barabasi, A.L. and Szathmary, E. (2001) Comparable system-level organization of archaea and eukaryotes. Nat. Genet. 29, 54–56.
[120] Wa, H.M. and Zeng, A.P. (2004) Phylogenetic comparison of metabolic capacities of organisms at genome level. Mol. Phyl. Evol. 31, 204–213.
[121] Coenye, T. and Vandamme, P. (2004) A genomic perspective on the relationship between the aquificales and the epsilon-proteobacteria. Syst. Appl. Microbiol. 27, 313–322.
[122] Teeling, H., Lombardot, T., Bauer, M., Ludwig, W. and Glockner, F.O. (2004) Evaluation of the phylogenetic position of the planctomycete '*Rhodopirellula baltica*' SH 1 by means of concatenated ribosomal protein sequences, DNA-directed RNA polymerase subunit sequences and whole genome trees. Int. J. Syst. Evol. Microbiol. 54, 791–801.
[123] Gutacker, M.M., Smoot, J.C., Migliaccio, C.A.L., Ricklefs, S.M., Hua, S., Cousins, D.V., Graviss, E.A., Shashkina, E., Kreiswirth, B.N. and Musser, J.M. (2002) Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. Genetics 162, 1533–1543.
[124] Ye, R.W., Wang, T., Bedzyk, L. and Croker, K.M. (2001) Applications of DNA microarrays in microbial systems. J. Microbiol. Methods 47, 257–272.
[125] Fitzgerald, J.R. and Musser, J.M. (2001) Evolutionary genomics of pathogenic bacteria. Trends Microbiol. 9, 547–553.
[126] Cho, J.C. and Tiedje, J.M. (2001) Bacterial species determination from DNA–DNA hybridization by using genome fragments and DNA microarrays. Appl. Environ. Microbiol. 67, 3677–3682.
[127] Hamels, S., Gala, J.L., Dufour, S., Vannuffel, P., Zammatteo, N. and Remacle, J. (2001) Consensus PCR and microarray for diagnosis of the genus *Staphylococcus*, species, and methicillin resistance. Biotechniques 31, 1364-+.
[128] Volokhov, D., Rasooly, A., Chumakov, K. and Chizhikov, V. (2002) Identification of *Listeria* species by microarray-based assay. J. Clin. Microbiol. 40, 4720–4728.
[129] Kakinuma, K., Fukushima, M. and Kawaguchi, R. (2003) Detection and identification of *Escherichia coli*, *Shigella*, and *Salmonella* by microarrays using the gyrB gene. Biotechnol. Bioeng. 83, 721–728.
[130] Fukushima, M., Kakinuma, K., Hayashi, H., Nagai, H., Ito, K. and Kawaguchi, R. (2003) Detection and identification of *Mycobacterium* species isolates by DNA microarray. J. Clin. Microbiol. 41, 2605–2615.
[131] Volokhov, D., Chizhikov, V., Chumakov, K. and Rasooly, A. (2003) Microarray-based identification of thermophilic *Campylobacter jejuni*, *C. coli*, *C. lari*, and *C. upsaliensis*. J. Clin. Microbiol. 41, 4071–4080.
[132] Loy, A., Lehner, A., Lee, N., Adamczyk, J., Meier, H., Ernst, J., Schleifer, K.H. and Wagner, M. (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. Appl. Environ. Microbiol. 68, 5064–5081.
[133] Koizumi, Y., Kelly, J.J., Nakagawa, T., Urakawa, H., El-Fantroussi, S., Al-Muzaini, S., Fukui, M., Urushigawa, Y. and Stahl, D.A. (2002) Parallel characterization of anaerobic toluene- and ethylbenzene-degrading microbial consortia by PCR-denaturing gradient gel electrophoresis, RNA-DNA membrane hybridization, and DNA microarray technology. Appl. Environ. Microbiol. 68, 3215–3225.
[134] Van Leeuwen, W.B., Jay, C., Snijders, S., Durin, N., Lacroix, B., Verbrugh, H.A., Enright, M.C., Troesch, A. and Van Belkum, A. (2003) Multilocus sequence typing of *Staphylococcus aureus* with DNA array technology. J. Clin. Microbiol. 41, 3323–3326.
[135] Akopyants, N.S., Fradkov, A., Diatchenko, L., Hill, J.E., Siebert, P.D., Lukyanov, S.A., Sverdlov, E.D. and Berg, D.E. (1998) PCR-based subtractive hybridization and differences in gene content among strains of *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA 95, 13108–13113.
[136] Winstanley, C. (2002) Spot the difference: applications of subtractive hybridisation to the study of bacterial pathogens. J. Med. Microbiol. 51, 459–467.
[137] Musser, J.M. (1996) Molecular population genetic analysis of emerged bacterial pathogens: selected insights. Emerg. Infect. Dis. 2, 1–17.
[138] Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J.J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M. and Spratt, B.G. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA 95, 3140–3145.
[139] Enright, M.C. and Spratt, B.G. (1999) Multilocus sequence typing. Trends Microbiol. 7, 482–487.
[140] Chan, M.S., Maiden, M.C.J. and Spratt, B.G. (2001) Database-driven multi locus sequence typing (MLST) of bacterial pathogens. Bioinformatics 17, 1077–1083.
[141] Urwin, R. and Maiden, M.C.J. (2003) Multi-locus sequence typing: a tool for global epidemiology. Trends Microbiol. 11, 479–487.
[142] Suerbaum, S., Smith, J.M., Bapumia, K., Morelli, G., Smith, N.H., Kunstmann, E., Dyrek, I. and Achtman, M. (1998) Free recombination within *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA 95, 12619–12624.
[143] Enright, M.C. and Spratt, B.G. (1998) A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology-Uk 144, 3049–3060.
[144] Achtman, M., Zurth, K., Morelli, C., Torrea, G., Guiyoule, A. and Carniel, E. (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. Proc. Natl. Acad. Sci. USA 96, 14043–14048.
[145] Byun, R., Elbourne, L.D.H., Lan, R.T. and Reeves, P.R. (1999) Evolutionary relationships of pathogenic clones of *Vibrio cholerae* by sequence analysis of four housekeeping genes. Infect. Immun. 67, 1116–1124.
[146] Enright, M.C., Day, N.P.J., Davies, C.E., Peacock, S.J. and Spratt, B.G. (2000) Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. J. Clin. Microbiol. 38, 1008–1015.
[147] Enright, M.C., Spratt, B.G., Kalia, A., Cross, J.H. and Bessen, D.E. (2001) Multilocus sequence typing of Streptococcus pyogenes and the relationships between *emm* type and clone. Infect. Immun. 69, 2416–2427.
[148] Dingle, K.E., Colles, F.M., Wareing, D.R.A., Ure, R., Fox, A.J., Bolton, F.E., Bootsma, H.J., Willems, R.J.L., Urwin, R. and

Maiden, M.C.J. (2001) Multilocus sequence typing system for *Campylobacter jejuni*. J. Clin. Microbiol. 39, 14–23.
[149] Nallapareddy, S.R., Duh, R.W., Singh, K.V. and Murray, B.E. (2002) Molecular typing of selected *Enterococcus faecalis* isolates: pilot study using multilocus sequence typing and pulsed-field gel electrophoresis. J. Clin. Microbiol. 40, 868–876.
[150] Meats, E., Feil, E.J., Stringer, S., Cody, A.J., Goldstein, R., Kroll, J.S., Popovic, T.J. and Spratt, B.G. (2003) Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. J. Clin. Microbiol. 41, 1623–1636.
[151] Helgason, E., Tourasse, N.J., Meisal, R., Caugant, D.A. and Kolsto, A.B. (2004) Multilocus sequence typing scheme for bacteria of the *Bacillus cereus* group. Appl. Environ. Microbiol. 70, 191–201.
[152] Godoy, D., Randle, G., Simpson, A.J., Aanensen, D.M., Pitt, T.L., Kinoshita, R. and Spratt, B.G. (2003) Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholdefia mallei*. J. Clin. Microbiol. 41, 2068–2079.
[153] Revazishvili, T., Kotetishvili, M., Stine, O.C., Kreger, A.S., Morris, J.G. and Sulakvelidze, A. (2004) Comparative analysis of multilocus sequence typing and pulsed-field gel electrophoresis for characterizing *Listeria monocytogenes* strains isolated from environmental and clinical sources. J. Clin. Microbiol. 42, 276–285.
[154] Muller-Graf, C.D.M., Whatmore, A.M., King, S.J., Trzcinski, K., Pickerill, A.P., Doherty, N., Paul, J., Griffiths, D., Crook, D. and Dowson, C.G. (1999) Population biology of *Streptococcus pneumoniae* isolated from oropharyngeal carriage and invasive disease. Microbiology-Sgm 145, 3283–3293.
[155] Coenye, T. and Lipuma, J.J. (2002) Multilocus restriction typing: a novel tool for studying global epidemiology of *Burkholderia cepacia* complex infection in cystic fibrosis. J. Infect. Dis. 185, 1454–1462.
[156] bennet, D.E. and Cafferkey, M.T. (2003) Multilocus restriction typing: a tool for *Neisseria meningitidis* strain discrimination. J. Med. Microbiol. 52, 781–787.
[157] Hoffmann, H. and Roggenkamp, A. (2003) Population genetics of the nomenspecies *Enterobacter cloacae*. Appl. Environ. Microbiol. 69, 5306–5318.
[158] Spratt, B.G. and Maiden, M.C.J. (1999) Bacterial population genetics, evolution and epidemiology. Philos. Trans. R. Soc. Lond. B Biol. Sci. 354, 701–710.
[159] Smith, J.M., Feil, E.J. and Smith, N.H. (2000) Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays 22, 1115–1122.
[160] Feil, E.J. and Spratt, B.G. (2001) Recombination and the population structures of bacterial pathogens. Annu. Rev. Microbiol. 55, 561–590.
[161] Cooper, J.E. and Feil, E.J. (2004) Multilocus sequence typing–what is resolved. Trends Microbiol. 12, 373–377.
[162] Posada, D., Crandall, K.A. and Holmes, E.C. (2002) Recombination in evolutionary genomics. Annu. Rev. Genet. 36, 75–97.
[163] Feil, E.J., Enright, M.C. and Spratt, B.G. (2000) Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between *Neisseria meningitidis* and *Streptococcus pneumoniae*. Res. Microbiol. 151, 465–469.
[164] Feil, E.J., Maiden, M.C.J., Achtman, M. and Spratt, B.G. (1999) The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. Mol. Biol. Evol. 16, 1496–1502.
[165] Feil, E.J., Smith, J.M., Enright, M.C. and Spratt, B.G. (2000) Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. Genetics 154, 1439–1450.
[166] Dorrell, N., Mangan, J.A., Laing, K.G., Hinds, J., Linton, D., Al-Ghusein, H., Barrell, B.G., Parkhill, J., Stoker, N.G., Karlyshev, A.V., Butcher, P.D. and Wren, B.W. (2001) Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. Genome Res. 11, 1706–1715.
[167] Salama, N., Guillemin, K., Mcdaniel, T.K., Sherlock, G., Tompkins, L. and Falkow, S. (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. Proc. Natl. Acad. Sci. USA 97, 14668–14673.
[168] Porwollik, S., Wong, R.M.Y. and Mcclelland, M. (2002) Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis. Proc. Natl. Acad. Sci. USA 99, 8956–8961.
[169] Dziejman, M., Balon, E., Boyd, D., Fraser, C.M., Heidelberg, J.F. and Mekalanos, J.J. (2002) Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. Proc. Natl. Acad. Sci. USA 99, 1556–1561.
[170] Hakenbeck, R., Balmelle, N., Weber, B., Gardes, C., Keck, W. and De Saizieu, A. (2001) Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of *Streptococcus pneumoniae*. Infect. Immun. 69, 2477–2486.
[171] Kato-Maeda, M., Rhee, J.T., Gingeras, T.R., Salamon, H., Drenkow, J., Smittipat, N. and Small, P.M. (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res. 11, 547–554.
[172] Dobrindt, U., Agerer, F., Michaelis, K., Janka, A., Buchrieser, C., Samuelson, M., Svanborg, C., Gottschalk, G., Karch, H. and Hacker, J. (2003) Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. J. Bacteriol. 185, 1831–1840.
[173] van Belkum, A., Melchers, W.J., Ijsseldijk, C., Nohlmans, L., Verbrugh, H. and Meis, J.F. (1997) Outbreak of amoxicillin-resistant *Haemophilus influenzae* type b: variable number of tandem repeats as novel molecular markers. J. Clin. Microbiol. 35, 1517–1520.
[174] Frothingham, R. and Meeker-O'connell, W.A. (1998) Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. Microbiology-Uk 144, 1189–1196.
[175] Keim, P., Price, L.B., Klevytska, A.M., Smith, K.L., Schupp, J.M., Okinaka, R., Jackson, P.J. and Hugh-Jones, M.E. (2000) Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. J. Bacteriol. 182, 2928–2936.
[176] Johansson, A., Farlow, J., Larsson, P., Dukerich, M., Chambers, E., Bystrom, M., Fox, J., Chu, M., Forsman, M., Sjostedt, A. and Keim, P. (2004) Worldwide genetic relationships among *Francisella tularensis* isolates determined by multiple-locus variable-number tandem repeat analysis. J. Bacteriol. 186, 5808–5818.
[177] Schouls, L.M., van der Heide, H.G., Vauterin, L., Vauterin, P. and Mooi, F.R. (2004) Multiple-locus variable-number tandem repeat analysis of Dutch *Bordetella pertussis* strains reveals rapid genetic changes with clonal expansion during the late 1990s. J. Bacteriol. 186, 5496–5505.
[178] Onteniente, L., Brisse, S., Tassios, P.T. and Vergnaud, G. (2003) Evaluation of the polymorphisms associated with tandem repeats for *Pseudomonas aeruginosa* strain typing. J. Clin. Microbiol. 41, 4991–4997.
[179] Alland, D., Whittam, T.S., Murray, M.B., Cave, M.D., Hazbon, M.H., Dix, K., Kokoris, M., Duesterhoeft, A., Eisen, J.A., Fraser, C.M. and Fleischmann, R.D. (2003) Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. J. Bacteriol. 185, 3392–3399.

[180] Robertson, G.A., Thiruvenkataswamy, V., Shilling, H., Price, E.P., Huygens, F., Henskens, F.A. and Giffard, P.M. (2004) Identification and interrogation of highly informative single nucleotide polymorphism sets defined by bacterial multilocus sequence typing databases. J. Med. Microbiol. 53, 35–45.
[181] Rombauts, S., Van De Peer, Y. and Rouze, P. (2003) AFLPin-silico, simulating AFLP fingerprints. Bioinformatics 19, 776–777.
[182] Bikandi, J., San Millan, R., Rementeria, A. and Garaizar, J. (2004) In silico analysis of complete bacterial genomes: PCR, AFLP-PCR and endonuclease restriction. Bioinformatics 20, 798–799.
[183] Mayr, E. (1942) Systematics and the Origin of Species. Columbia University Press, New York.
[184] Rossello-Mora, R. (2003) Opinion: the species problem, can we achieve a universal concept. Syst. Appl. Microbiol. 26, 323–326.
[185] Whitman, W.B., Coleman, D.C. and Wiebe, W.J. (1998) Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. USA 95, 6578–6583.
[186] Ward, D.M. (1998) A natural species concept for prokaryotes. Curr. Opin. Microbiol. 1, 271–277.
[187] Lan, R. and Reeves, P.R. (2001) When does a clone deserve a name. A perspective on bacterial species based on population genetics. Trends Microbiol. 9, 419–424.
[188] Amann, R.I., Ludwig, W. and Schleifer, K.H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169.
[189] Hugenholtz, P., Goebel, B.M. and Pace, N.R. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. 180, 4765–4774.
[190] Gil, R., Silva, F.J., Zientz, E., Delmotte, F., Gonzalez-Candelas, F., Latorre, A., Rausell, C., Kamerbeek, J., Gadau, J., Holldobler, B., Van Ham, R.c.h.j., Gross, R. and Moya, A. (2003) The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. Proc. Natl. Acad. Sci. USA 100, 9388–9393.
[191] Rodriguez-Valera, F. (2004) Environmental genomics, the big picture. FEMS Microbiol. Lett. 231, 153–158.
[192] Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R., Loiacono, K.A., Lynch, B.A., Macneil, I.A., Minor, C., Tiong, C.L., Gilman, M., Osburne, M.S., Clardy, J., Handelsman, J. and Goodman, R.M. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. 66, 2541–2547.
[193] Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D.Y., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H. and Smith, H.O. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74.
[194] Lemos, E.G.D., Alves, L.M.C. and Campanharo, J.C. (2003) Genomics-based design of defined growth media for the plant pathogen *Xylella fastidiosa*. FEMS Microbiol. Lett. 219, 39–45.
[195] Renesto, P., Crapoulet, N., Ogata, H., La Scola, B., Vestris, G., Claverie, J.M. and Raoult, D. (2003) Genome-based design of a cell-free culture medium for *Tropheryma whipplei*. Lancet 362, 447–449.
[196] Dopazo, J., Dress, A. and Vonhaeseler, A. (1993) Split decomposition – a technique to analyze viral evolution. Proc. Natl. Acad. Sci. USA 90, 10320–10324.
[197] Feil, E.J., Li, B., Aanensen, D.M., Hanage, W.P. and Spratt, B.G. (2004) eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. 186, 1518–1530.
[198] Smith, N.H., Holmes, E.C., Donovan, G.M., Carpenter, G.A. and Spratt, B.G. (1999) Networks and groups within the genus *Neisseria*: analysis of argF, recA, rho, and 16S rRNA sequences from human *Neisseria* species. Mol. Biol. Evol. 16, 773–783.
[199] Farfan, M., Minana-Galbis, D., Fuste, M.C. and Loren, J.G. (2002) Allelic diversity and population structure in *Vibrio cholerae* O139 Bengal based on nucleotide sequence analysis. J. Bacteriol. 184, 1304–1313.
[200] Suerbaum, S., Lohrengel, M., Sonnevend, A., Ruberg, F. and Kist, M. (2001) Allelic diversity and recombination in *Campylobacter jejuni*. J. Bacteriol. 183, 2553–2559.
[201] Coenye, T. and Lipuma, J.J. (2003) Population structure analysis of *Burkholderia cepacia* genomovar III: varying degrees of genetic recombination characterize major clonal complexes. Microbiology-Sgm 149, 77–88.
[202] Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. and Ikemura, T. (2003) Informatics for unveiling hidden genome signatures. Genome Res. 13, 693–702.
[203] Bryant, D. and Moulton, V. (2004) Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21, 255–265.
[204] Zhaxybayeva, O., Hamel, L., Raymond, J. and Gogarten, J.P. (2004) Visualization of the phylogenetic content of five genomes using dekapentagonal maps. Genome Biol. 5, R20.
[205] Bininda-Emonds, O.R.P., Gittleman, J.L. and Purvis, A. (1999) Building large trees by combining phylogenetic information: a complete phylogeny of the extant carnivora (mammalia). Biol. Rev. 74, 143–175.
[206] Eulenstein, O., Chen, D.H., Burleigh, J.G., Fernandez-Baca, D. and Sanderson, M.J. (2004) Performance of flip supertree construction with a heuristic algorithm. Syst. Biol. 53, 299–308.
[207] Branscomb, E. and Predki, P. (2002) On the high value of low standards. J. Bacteriol. 184, 6406–6409.
[208] Fraser, C.M., Eisen, J.A., Nelson, K.E., Paulsen, I.T. and Salzberg, S.L. (2002) The value of complete microbial genome sequencing (you get what you pay for). J. Bacteriol. 184, 6403–6405.
[209] Ward, N., Eisen, J., Fraser, C. and Stackebrandt, E. (2001) Sequenced strains must be saved from extinction. Nature 414, 148.
[210] Coenye, T. and Vandamme, P. (2004) Bacterial whole-genome sequences: minimal information and strain availability. Microbiology-Sgm 150, 2017–2018.
[211] Vandepoele, K., Saeys, Y., Simillion, C., Raes, J. and Van de Peer, Y. (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. Genome Res. 12, 1792–1801.