International Journal of Systematic and Evolutionary Microbiology (2014), **64**, 316–324 **DOI** 10.1099/ijs.0.054171-0

Review

Integrating genomics into the taxonomy and systematics of the *Bacteria* and *Archaea*

Jongsik Chun[1] and Fred A. Rainey[2]

[1]School of Biological Sciences, Seoul National University, Seoul 151-742, Republic of Korea

[2]Department of Biological Sciences, University of Alaska Anchorage, Anchorage, AK 99508, USA

Correspondence
Jongsik Chun
jchun@snu.ac.kr
Fred A. Rainey
farainey@uaa.alaska.edu

The polyphasic approach used today in the taxonomy and systematics of the *Bacteria* and *Archaea* includes the use of phenotypic, chemotaxonomic and genotypic data. The use of 16S rRNA gene sequence data has revolutionized our understanding of the microbial world and led to a rapid increase in the number of descriptions of novel taxa, especially at the species level. It has allowed in many cases for the demarcation of taxa into distinct species, but its limitations in a number of groups have resulted in the continued use of DNA–DNA hybridization. As technology has improved, next-generation sequencing (NGS) has provided a rapid and cost-effective approach to obtaining whole-genome sequences of microbial strains. Although some 12 000 bacterial or archaeal genome sequences are available for comparison, only 1725 of these are of actual type strains, limiting the use of genomic data in comparative taxonomic studies when there are nearly 11 000 type strains. Efforts to obtain complete genome sequences of all type strains are critical to the future of microbial systematics. The incorporation of genomics into the taxonomy and systematics of the *Bacteria* and *Archaea* coupled with computational advances will boost the credibility of taxonomy in the genomic era. This special issue of *International Journal of Systematic and Evolutionary Microbiology* contains both original research and review articles covering the use of genomic sequence data in microbial taxonomy and systematics. It includes contributions on specific taxa as well as outlines of approaches for incorporating genomics into new strain isolation to new taxon description workflows.

In the current practice of the taxonomy of the *Bacteria* and *Archaea*, a novel species is recognized using the polyphasic approach, in which we consider multidimensional aspects of organisms including phenotypic, genotypic and chemotaxonomic traits (Colwell, 1970; Tindall *et al.*, 2010). In this process, genotypic characterization is an essential component in describing species, as genetic information sheds light on evolutionary relationships between diverse lineages. Phylogenetic analysis based on 16S rRNA gene sequences and determination of similarity between sequences are now routinely carried out as the first step in identifying novel organisms (Stackebrandt & Ebers, 2006; Stackebrandt & Goebel, 1994; Tindall *et al.*, 2010). Over the last 50 years, DNA–DNA hybridization (DDH), which measures indirectly the degree of genetic similarity between two genomes, has been the 'gold standard' for bacterial species demarcation by providing a constant numerical threshold (DDH value >70 %) for the species boundary (Wayne *et al.*, 1987). Since the invention of cost-effective, high-throughput DNA sequencing, known as next-generation sequencing (NGS), sequencing a bacterial or archaeal genome and its direct comparison are readily applicable to microbial taxonomy, even in general laboratories in academic institutions. This has resulted in the use of the genome sequence in microbial taxonomy becoming feasible, and in such approaches being applied in recent taxonomic studies involving formal taxonomic proposals. Here, we review the current status of taxonomy of the *Bacteria* and *Archaea* and the potential use of genomics in classification and identification of these organisms.

Abbreviations: ANI, average nucleotide identity; DDH, DNA–DNA hybridization; GBDP, genome BLAST distance phylogeny; MLSA, multilocus sequence analysis; MLST, multilocus sequencing typing; NGS, next-generation sequencing; OGRI, overall genome relatedness index.

Taxonomy of the *Bacteria* and *Archaea*: where are we now?

At present, the number of validly published names of prokaryotic species is about 12 000. This number is clearly an underestimation of what exists on Earth, given that there are over 1.5 million known animal species. Consider that every animal species likely harbours microbial species whose ecological niche is that particular animal host. Moreover, the way that we define bacterial and archaeal species is based largely on genomic relatedness, which is a much more relaxed definition than that of eukaryote species. For example, if the current criterion used to define bacterial and archaeal species is applied to the animal world, all primates should be classified as a single species

(Whitman *et al.*, 1998). Therefore, direct comparison of species diversity between the domains *Bacteria*, *Archaea* and *Eukarya* has no rational basis due to radical differences in how we define 'species'. Even so, we can still try to measure overall species diversity of the bacterial and archaeal worlds using the current definition of species. There have been a few attempts to estimate the total number of bacterial and archaeal species on Earth, suggesting that this figure can be over a million (Curtis *et al.*, 2002).

Before the dawn of the molecular biology era, we were very slow in discovering novel bacterial and archaeal species due to the lack of efficient and objective methods to recognize and delineate novel species. In 1992, Fox and colleagues first noticed that two bacterial strains belonging to different species can share high levels of genetic similarity, even almost identical 16S rRNA gene sequences (Fox *et al.*, 1992). Later, the utility of 16S rRNA gene sequences in delineating novel species was formulated by Stackebrandt & Goebel (1994), who proposed that, if two strains share less than 97 % 16S rRNA gene sequence similarity, they belong to different species. This simple and objective guideline has greatly increased the rate of discovering novel species when tedious and labour-intensive DDH is avoided (Fig. 1). At present, the most widely used methodology for proving the novelty of a bacterial or archaeal species is the combination of 16S rRNA gene sequencing and DDH; the latter is only applied for cases in which the 16S rRNA gene sequence similarity between two strains is >97 %. It should be noted that novel species are also differentiated using phenotypic characteristics on top of genetic evidence (Wayne *et al.*, 1987; Tindall *et al.*, 2010).

NGS technologies and genome assembly

It has been two decades since the first bacterial genome, of *Haemophilus influenzae* Rd, was sequenced (Fleischmann *et al.*, 1995). Prior to the invention of NGS in 2005, the use of genome sequencing in bacterial and archaeal taxonomy was hampered by the high-cost, labour-intensive and time-consuming process of the conventional Sanger sequencing method. The first NGS platform that was widely used in microbiology was the Roche 454 sequencing system, which adopted the principle of pyrosequencing (Margulies *et al.*, 2005), followed by other NGS platforms. Illumina DNA sequencing technology, which was originally developed by Solexa, is based on bridge-amplification and reversible terminators (Bentley, 2006), and is provided as different instruments including HiSeq and its lower-throughput, bench-top version called MiSeq.

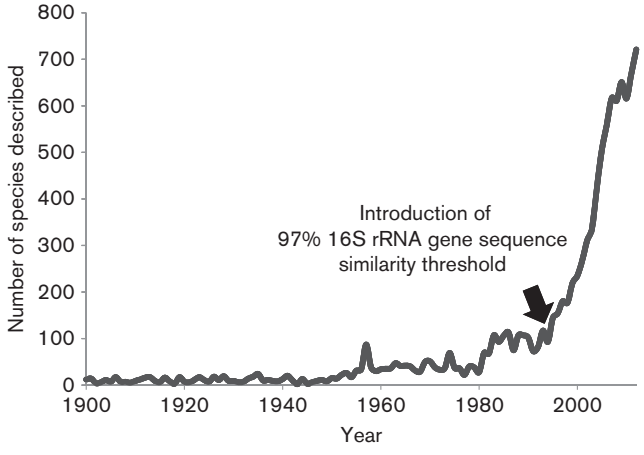

Fig. 1. Number of prokaryotic species described per year since 1900.

The Roche 454 and Illumina platforms are generally called NGS technologies, whereas those subsequently commercialized are often referred to as third-generation sequencing technologies. Among them, Pacific Biosciences provides single-molecule sequencing, meaning that no template amplification is needed. This technology can produce very long, albeit less accurate, sequencing reads, of over 10 000 bp (Eid *et al.*, 2009). Similar to Roche 454, Ion Torrents (Life Technologies) uses emulsion PCR for template DNA amplification, but adopts semiconductor technology to sequence DNA, for which a change in pH is detected instead of light. It is perhaps impractical to compare these DNA sequencing systems objectively at any given time, as all of these technologies and related experimental protocols are constantly being improved, and new ones are frequently introduced into the market. However, it is evident that these massively high-throughput sequencing technologies are becoming readily accessible to general microbiology laboratories, which means that genomics will be within the reach of most microbial taxonomists in the near future.

All of these new DNA sequencing platforms can generate massive quantities of sequence data with relatively shorter read lengths, but are significantly cheaper than the conventional Sanger method. In general, the sequencing accuracy of NGS platforms is over 99 %, which is slightly lower than that of the Sanger method. However, in genome-sequencing projects using NGS, multiple depths of sequencing coverage are always applied; therefore the final assembled sequences (called contigs) are generally of high accuracy. In genome-sequencing projects based on the Sanger method, depths of sequencing coverage usually range from 7- to 10-fold (one nucleotide position of the genome is sequenced seven to ten times on average). In contrast, at least 30× depths are usually applied in NGS-based genome-sequencing projects. Each of the available NGS technologies has pros and cons regarding read length, accuracy, the nature of sequencing errors, its ability to produce paired-end information and cost-effectiveness. It is, therefore, sensible to employ two or more NGS systems simultaneously to gain genome data of better quality. Such an approach is called 'hybrid genome assembly'. For example, preliminary high-quality contigs can be obtained using Illumina MiSeq/Roche 454/Ion Torrents with high depths of sequencing coverage, and the resulting contigs

can be joined to get the complete sequence by applying Pacific Biosciences's long-read sequencing (Koren *et al.*, 2012). Basic features of each sequencing platform used in prokaryotic genome sequencing and their latest updates are summarized in Table 1.

Estimation of genomic relatedness using genome sequencing data

DDH is an experimental method to measure the degree of relatedness between two different genomes by applying nucleic acid hybridization. Since the first introduction of this technique in microbiology in the late 1960s (Johnson & Ordal, 1968), it has been portrayed as the 'gold standard' for species delineation of the *Bacteria* and *Archaea*. Wayne *et al.* (1987) recommended a DDH value of 70 % as the threshold for the bacterial species boundary, although there are some exceptions (Rosselló-Mora, 2006). Due to the labour-intensive and error-prone nature of DDH experiments, there has been a continuous demand for an alternative genotypic standard (Stackebrandt *et al.*, 2002). Comparative studies between 16S rRNA gene sequence similarities and DDH values revealed that 97 % 16S rRNA gene sequence similarity corresponded to the 70 % DDH value (Stackebrandt & Goebel, 1994). It is now generally accepted that DDH is only required for a pair of strains showing 97 % or more 16S rRNA gene sequence similarity when a novel species is proposed (Tindall *et al.*, 2010). Even though the 16S rRNA gene sequence is regarded as the best approach in placing an uncharacterized strain onto the phylogenetic framework of all microbes, in many cases, it is too conserved to distinguish two closely related species. This limitation can be overcome by whole-genome-based comparisons, which have better resolving power in species delineation (Oren & Papke, 2010).

Since genome sequencing is now affordable and accessible to general microbiology laboratories, it has been suggested that comparison of whole-genome sequence data, as a form of digital, *in silico* DDH, would be used to replace DDH for taxonomic purposes (Konstantinidis & Tiedje, 2005). Many efforts have been made to correlate DDH values with digital DDH indices derived from computational comparison of two genome sequences. Average nucleotide identity (ANI) represents a mean of identity values between multiple sets of orthologous regions shared by two genomes. Konstantinidis & Tiedje (2005) first showed that the ANI of shared genes between two genomes is a robust measure of evolutionary distance, as ANI values correlate well with DDH values. Two years later, Goris *et al.* (2007) refined this method by artificially cutting the query genome sequence into fragments of 1020 bp, simulating the DNA fragmentation step of DDH experiments. At present, ANI of Goris *et al.* (2007) is usually calculated as a mean identity of all BLASTN (Altschul *et al.*, 1990) matches between two genome sequences where only matches with at least 30 % overall sequence identity are considered. Richter & Rosselló-Móra (2009) employed MUMmer software (Kurtz *et al.*, 2004) instead of BLASTN to get faster results while not losing accuracy. Generally, ANI based on the BLASTN method (ANIb) is adopted and used more widely than ANI based on the MUMmer algorithm (ANIm). On the basis of comparative studies between ANI and DDH values, ANI values equivalent to the 70 % DDH threshold are 95–96 % (Goris *et al.*, 2007; Richter & Rosselló-Móra, 2009). In addition to DDH, an alignment-free method using tetranucleotide frequency has been shown to represent a good correlation with ANI (Richter & Rosselló-Móra, 2009).

Unlike ANI, which is a similarity-type index, genome BLAST distance phylogeny (GBDP) is a distance-type genome relatedness index (Henz *et al.*, 2005). In the GBDP algorithm, the genome sequence is not artificially cut into small pieces. Instead, two genome sequences are aligned to each other using local alignment tools such as BLAST (Altschul *et al.*, 1990), BLAST+ (Camacho *et al.*, 2009)

Table 1. Basic features of NGS platforms used in prokaryotic genome sequencing

Run time is the time required for the DNA sequencing reaction after the library preparation step is complete. Throughput is the number of prokaryote genomes generated per run when the size of genome is 4 Mb and 12.5× depth of sequencing coverage is needed for each assembly.

Manufacturer/platform	Run time	Read length	Yield per run	Throughput
Roche				
454 GS Junior	10–12 h	400–500 bp	~50 Mb	~1
454 FLX Titanium XL+	~23 h	700–1000 bp	~1 Gb	~20
Illumina				
GAIIx	~14 days	2 × 150 bp	85–95 Gb	170–190
MiSeq	~39 h	2 × 250 bp	7.5–8.5 Gb	15–17
HiSeq 2500 (rapid run mode, single flow cell)	40 h	2 × 150 bp	75–90 Gb	160–180
Pacific Biosciences				
PacBio RS II	~2 h	3–5 kbp (up to >20 kbp)	~250 Mb	~5
Life Technologies				
Ion PGM System (Ion 316 chip v2)	4.9 h	400 bp	600–1000 Mb	12–20
Ion Proton PGM System (Ion PI Chip)	2–4 h	200 bp	~10 Gb	~200

and BLAT (Kent, 2002) to obtain sets of high-scoring segment pairs (HSPs). These are then used to calculate a specific distance formula. Practically, GBDP offers three algorithms, namely, greedy, greedy-with-trimming and coverage, depending on the way overlapped HSPs are processed. In a recent version, the algorithm was further enhanced by using more sophisticated statistical models with confidence-interval estimation (Meier-Kolthoff *et al.*, 2013).

MUMi is another distance-type index that is based on maximal unique exact matches shared by two genomes (Deloger *et al.*, 2009). It was originally developed to provide higher resolution at the intraspecies level. Like the ANIm algorithm, MUMi uses the MUMmer program for rapid pairwise genome comparison using at least 19 bp-long shared segments. MUMi has been shown to correlate well with ANI values (Deloger *et al.*, 2009).

All of the above methods, for which we coin the term overall genome relatedness indices (OGRI), utilize whole-genome sequences, but not individual gene sequences or a set of sequences. Therefore, gene finding/prediction and functional annotation of each predicted gene are not required. Since the gene-finding step, in particular, can be carried out in many different ways, OGRI provides a simple, yet reproducible and objective, way of comparing two genomes.

Unlike OGRI, only conserved parts of genomes can be considered for phylogenetic comparison of two strains. This approach has been widely adopted for molecular epidemiology as a form of multilocus sequencing typing (MLST) (Maiden *et al.*, 1998; Sullivan *et al.*, 2005). In MLST, sequences of 8–12 selected conserved genes are determined, and each sequence is grouped into sequence types (ST) according to their sequence. Every ST is treated or weighted equally regardless of their sequence similarity. MLST is a powerful typing method in differentiating closely related strains within a species, because of its ability to overcome the bias caused by lateral gene transfer. However, if this concept is applied to higher-level classification, especially to differentiate two distinct species, sequence similarity values between the two gene sequences should be considered. In contrast to MLST, this phylogenetic approach is known as multilocus sequence analysis (MLSA). The potential of this method in bacterial species definition was strongly endorsed by the ad hoc committee of the International Committee on Systematics of Prokaryotes (Stackebrandt *et al.*, 2002), given that whole-genome sequencing was not accessible to most microbial taxonomists at that time. MLSA has been used successfully in the classification and identification of many taxa (Guo *et al.*, 2008; Marrero *et al.*, 2013; Martens *et al.*, 2007).

In MLSA, sequences of each selected gene are sequenced using PCR and Sanger sequencing, which restricts analysis based on a large number of genes. In contrast, whole-genome sequencing allows us to choose as many genes as possible for such an analysis. Mende *et al.* (2013) devised a special MLSA method, called specI, in which 40 universal, single-copy, protein-coding genes are selected to calculate a genome similarity based on ANI of the 40 genes. They went on to conduct a large-scale study, suggesting that their method is comparable to ANI and can be used for bacterial species demarcation. Unlike specI, which has been designed to be used in studies of all bacterial phyla, a larger number of genes can be used in specific cases. For example, the phylogenetic structure of *Vibrio cholerae* and related taxa has been elucidated by MLSA based on >1000 genes (Chun *et al.*, 2009; Haley *et al.*, 2010), resulting in the recognition of two novel species of the genus *Vibrio*. The MLSA method differs from OGRI as the former only considers parts of the genome and the input data should be gene sequences, not genomes, which means that the gene finding/prediction step should be carried out in advance for the MLSA. MLSA is more phylogenetically sensible and can be a good alternative to check conclusions derived from OGRI methods.

Among all the digital genomic relatedness indices mentioned above, the ANIb algorithm has been used most widely for classification and identification of bacteria and archaea (Camelo-Castillo *et al.*, 2014; Chan *et al.*, 2012; Haley *et al.*, 2010; Hoffmann *et al.*, 2012; Jiménez *et al.*, 2013; Lee *et al.*, 2013; Löffler *et al.*, 2013; Lucena *et al.*, 2012; Richter & Rosselló-Móra, 2009; Ruvira *et al.*, 2013; Vanlaere *et al.*, 2009; Yi *et al.*, 2012). Bioinformatic tools for calculating *in silico* pairwise genomic relatedness values are available as either standalone computer programs or web-based services, and are summarized in Table 2.

Applying genomics to the taxonomy of the *Bacteria* and *Archaea*

As outlined above, the contribution of 16S rRNA gene sequences to the classification and identification of the *Bacteria* and *Archaea* has been immense (Rosselló-Mora & Amann, 2001). Unlike DDH, once a 16S rRNA gene sequence is obtained from an isolate, it can be compared against speciality databases, such as EzTaxon (Kim *et al.*, 2012) and the Ribosomal Database Project (Cole *et al.*, 2009); these databases hold carefully curated 16S rRNA gene sequences. To make this approach taxonomically meaningful, 16S rRNA gene sequences of type strains of all known species should be available for comparison. Among species whose names have formal standing in nomenclature (n=10 944), 98.8 % have 16S rRNA gene sequences available for their type strains (Table 3). This level of high coverage in a so-called 'DNA barcode' database is unprecedented compared with those available for animals, plants, fungi and protists. Since any publication proposing a novel bacterial or archaeal species is likely to report the 16S rRNA gene sequence of the type strain of the novel taxon, this level should be maintained, at least, if not increased. This is the result of long-term, joint efforts of many microbial taxonomists over decades, including a recent endeavour to fill the remaining gaps in the 16S rRNA gene sequence database (Yarza *et al.*, 2013).

Table 2. Bioinformatic tools and resources for genome-to-genome comparison for taxonomic purposes

Genome relatedness index	Threshold for species demarcation	Gene finding/ annotation required?	Tools	Description	URL
ANI	95–96 %	No	JSpecies	JAVA-based standalone software that calculates ANIb and ANIm (Richter & Rosselló-Móra, 2009)	http://www.imedea.uib.es/jspecies/
			EzGenome	Web-based service for ANIb calculation	http://www.ezbiocloud.net/ezgenome/ani
GBDP	0.258	No	Genome-to-Genome Distance Calculator	Web-based service for calculation of pairwise GBDP distance using NCBI-BLAST, BLAST+, BLAT, BLASTZ and MUMmer (Meier-Kolthoff *et al.*, 2013)	http://ggdc.dsmz.de/
MUMi	Not available	No	MUMi	Web-based service for calculation of MUMi (Deloger *et al.*, 2009)	http://genome.jouy.inra.fr/mumi/
Nucleotide identity	96.5 %	Yes	specI	Web-based service and Linux standalone program for identification using 40 universal, single-copy marker genes (Mende *et al.*, 2013)	http://www.bork.embl.de/software/specI/

If genome sequence comparisons are used in a way similar to that applied to 16S rRNA gene sequences, genome sequences of all type strains should be determined and made available in publicly accessible databases. At present, the number of genome sequences available (n=11 913) is slightly larger than the number of known species. However, sequencing efforts have been focused on a few clinically important bacterial species. For example, 1037 and 440 genome sequences are available for strains belonging to *Escherichia coli* and *Staphylococcus aureus*, respectively. Genome sequence availability over known bacterial diversity is skewed towards cultivability and subjective reasons, not taxonomic importance. Consequently, only 1725 species (15.8 % of the total) have genome sequence information for their type strains (Table 3). In addition, the genomes of strains other than the type strain of 969 species have been sequenced; therefore, these data have inherent limitations for use in classification and identification. It is clear that there is an urgent need to expand our genomic knowledge on hitherto uncharacterized type strains of species with validly published names, which will serve as a crucial framework for future taxonomy as well as other fields of microbiology, including ecology and clinical microbiology. One such attempt is an international consortium spearheaded by the US DOE Joint Genome Institute, called the Genomic Encyclopedia of Bacteria and Archaea project (GEBA; http://www.jgi.doe.gov/programs/GEBA/), which is trying to fill the gaps in genome sequencing throughout the phylogenetic tree of the *Bacteria* and *Archaea* (Wu *et al.*, 2009). Similarly, genome information for unexplored uncultured phyla of which only 16S rRNA gene sequences are available has been unravelled using single-cell genomics technology (Hofer, 2013; Rinke *et al.*, 2013; Swan *et al.*, 2013). This kind of effort will contribute appreciably to bacterial systematics by providing more detailed and comprehensive taxonomic information for circumscribing bacterial and archaeal species. The overall distributions of all bacterial species as well as uncultured phylotypes and the availability of genome information are depicted in Fig. 2.

Moving into the genomics era

It is clear, at this point, that the use of genomics in the classification and identification of bacteria and archaea will be greatly facilitated as the cost of DNA sequencing is

Table 3. Current status of prokaryotic taxonomy and related genome sequencing (as of 31 July 2013)

Domain	Species with:			
	Validly published name	16S rRNA gene sequence (type strain)	Genome sequence (any strain)	Genome sequence (type strain)
Bacteria	10 546	10 420	2518	1567
Archaea	398	397	176	158
Total (prokaryotes)	10 944	10 817	2694	1725
Percentage	100 %	98.8 %	24.6 %	15.8 %

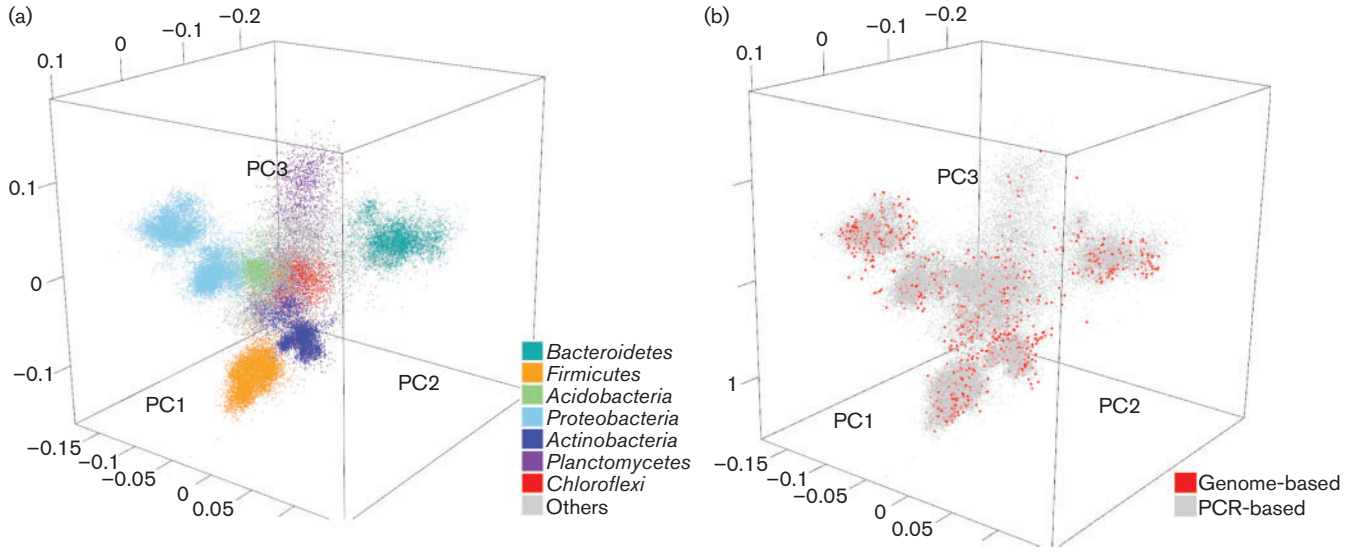

Fig. 2. Known bacterial species diversity based on 16S rRNA sequences (a) and distribution of species for which genome sequences are available (b). 16S rRNA gene sequences of bacteria were aligned manually using the EzEditor software (Jeon *et al.*, 2014) and three-dimensional co-ordinates for each sequence were obtained by calculating pairwise similarity followed by principal co-ordinate analysis (Sokal & Sneath, 1963). Animated and updated versions of figures are available at http://www.ezbiocloud.net/ezgenome/status/.

reduced at an unprecedented speed (http://www.genome.gov/sequencingcosts/). Up to now, a two-step approach has been widely used in recognizing novel species: 16S rRNA gene sequence similarity is first calculated, and DDH is applied to cases with >97 % 16S rRNA gene similarity (Stackebrandt & Goebel, 1994; Tindall *et al.*, 2010). As the number of genome sequences covering species with validly published names grows, DDH will be easily replaced by genome sequence-derived relatedness indices such as ANI. Since genome sequencing encompasses high-quality 16S rRNA gene sequencing, this will be a one-step approach that can be fully automated at the post-DNA-sequencing stage. Additionally, genome sequencing also provides a precise DNA G+C content of the genome, which has been of historical value in bacterial taxonomy.

For genome sequencing to be used on a routine basis in bacterial and archaeal taxonomy, objective and reproducible bioinformatics tools should be available. Fortunately, unlike the highly sophisticated bioinformatics methods required for functional genomics and comparative genomics, routine identification can be carried out by fairly simple bioinformatics processes. In OGRI methods, genome assembly and calculation of genome-based relatedness indices are conducted serially. Both steps have been validated extensively and can be coupled to be fully automated, even through a web-service or standalone software. However, such a service or software is not yet available. Unlike OGRI, MLSA methods require additional gene finding and functional annotation between the aforementioned steps. It should be noted that OGRI, like DDH, shows no resolution at the suprageneric taxonomic level. However, MLSA based on conserved genes has been applied successfully to dissecting phylogenetic structure among phyla and has great potential in suprageneric classification (Lang *et al.*, 2013). The rate at which genome sequencing is carried out at various microbiology laboratories worldwide will be likely to be exponential, given the current growing trends of released genome sequences in public database (Fig. 3). Mining of this huge dataset to understand the true nature of bacterial and archaeal species will be extremely challenging considering the computational cost of both the hardware and software required. For example, saving all available bacterial and archaeal genome sequences onto even a smartphone is possible (4 megabases × 12 000 is only 40 gigabases). However, pairwise comparison of all genomes will take an unimaginably long time (if each comparison takes 1 min, 12 000 × 12 000 comparisons will take 273 years). It is likely that the future direction for species definition of the domains *Bacteria*, *Archaea* and *Eukarya* will be a matter of computational or bioinformatics means.

Since many DNA sequencing platforms with different technical specifications and experimental protocols are being developed, a minimum standard for each case should be set for taxonomic use, preferably by international organizations such as the International Committee on Systematics of Prokaryotes or its taxonomic subcommittees.

Modern bacterial and archaeal taxonomy has been greatly advanced by the introduction of 16S rRNA gene sequencing. We believe that genomics will have equal or even greater impact on how we classify and identify bacteria and

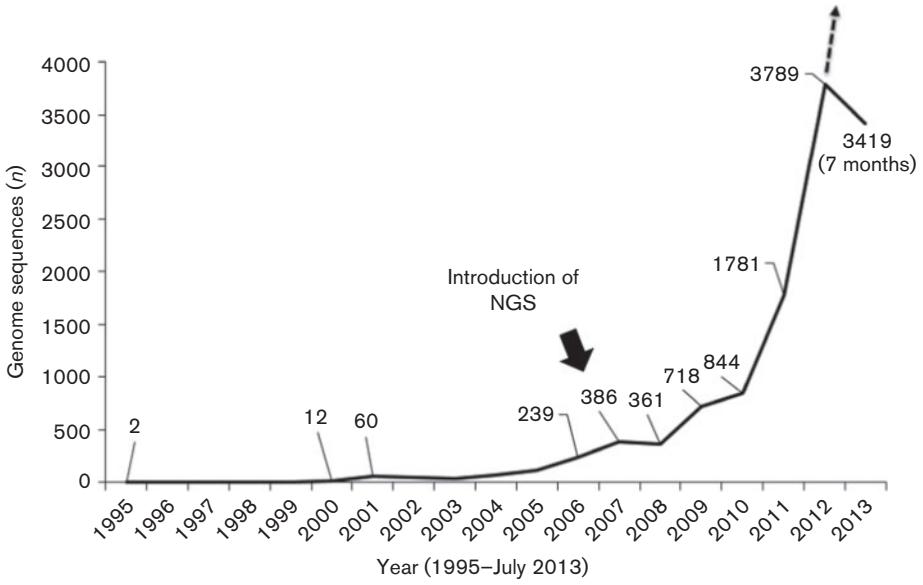

Fig. 3. Number of prokaryotic genome sequences released to public databases per year.

archaea. By coupling with recent advances in computational sciences, such as cloud computing and big data analysis, microbial taxonomists should be able to provide more robust and objective approaches for classification and identification, which have been the foundation of our society ever since the discovery of microbes.

This issue of *International Journal of Systematic and Evolutionary Microbiology* contains a number of contributions, both original research and reviews on topics that cover the use of genome sequence data in microbial taxonomy. As well as covering prokaryotic taxonomy, two papers address the use of genome sequence data in the taxonomy of yeasts (Kutzman, 2014) and algae (Kim *et al.*, 2014b). The contribution of Ramasamy *et al.* (2014) outlines a polyphasic strategy that incorporates genome sequence data for the identification of novel species of bacteria and names the approach 'taxo-genomics'. Amaral *et al.* (2014) provide a taxon-specific strategy for the identification of phenotypes from whole-genome sequences that can be used in the identification of species and strains of the genus *Vibrio*. A paper describing the use of conserved indels and signature proteins in the taxonomy of a group of plant-pathogenic genera not only demonstrates the use of this approach in taxonomy but suggests that these unique signatures have applications and could be used to develop antibacterial agents (Naushad *et al.*, 2014). Meier-Kolthoff *et al.* (2014) address the use of genomic data in the calculation of G+C content of the DNA of organisms, highlighting the age-old problems with the methodologies used for the determination of G+C content and DDH values. Kim *et al.* (2014a) report a large-scale comparison between 16S rRNA gene sequence similarities and ANI, which results in the proposal of a new threshold of 16S rRNA gene sequence similarity in the recognition of novel species. All of these constitutions make the case for using genome sequence-based approaches that are reliable and reproducible.

This collection of papers sets the stage for moving the taxonomy and systematics of the *Bacteria* and *Archaea* to the next level and into the genomic era. It could be considered that bacterial taxonomy is languishing in the phenotypic and chemotaxonomic quagmire of unreliable methods and difficult-to-replicate data and that genome-sequence approaches to determining not only the relationships and identity of micro-organisms but also the expected phenotype could indeed rescue this science, boost its credibility and bring it into the genomic era.

Acknowledgements

We thank Sang-Cheol Park, Jae-Woo Kim, Seok-Hwan Yoon and Mincheol Kim for preparing tables and figures. J. C. was supported by grants from the National Research Foundation of the Republic of Korea (2013-035122 and 2012M3A9D1054622).

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–410.

Amaral, G. R. S., Dias, G. M., Wellington-Oguri, M., Chimetto, L., Campeão, M. E., Thompson, F. L. & Thompson, C. C. (2014). Genotype to phenotype: identification of diagnostic vibrio phenotypes using whole genome sequences. *Int J Syst Evol Microbiol* **64**, 357–365.

Bentley, D. R. (2006). Whole-genome re-sequencing. *Curr Opin Genet Dev* **16**, 545–552.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Camelo-Castillo, A., Benítez-Páez, A., Belda-Ferre, P., Cabrera-Rubio, R. & Mira, A. (2014). *Streptococcus dentisani* sp. nov., a new member of the Mitis group. *Int J Syst Evol Microbiol* **64** (in press).

Chan, J. Z.-M., Halachev, M. R., Loman, N. J., Constantinidou, C. & Pallen, M. J. (2012). Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. *BMC Microbiol* **12**, 302.

Chun, J., Grim, C. J., Hasan, N. A., Lee, J. H., Choi, S. Y., Haley, B. J., Taviani, E., Jeon, Y. S., Kim, D. W. & other authors (2009). Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc Natl Acad Sci U S A* **106**, 15442–15447.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T. & other authors (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**, D141–D145.

Colwell, R. R. (1970). Polyphasic taxonomy of the genus *Vibrio*: numerical taxonomy of *Vibrio cholerae*, *Vibrio parahaemolyticus*, and related *Vibrio* species. *J Bacteriol* **104**, 410–433.

Curtis, T. P., Sloan, W. T. & Scannell, J. W. (2002). Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci U S A* **99**, 10494–10499.

Deloger, M., El Karoui, M. & Petit, M. A. (2009). A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. *J Bacteriol* **191**, 91–99.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P. & other authors (2009). Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A. & other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

Fox, G. E., Wisotzkey, J. D. & Jurtshuk, P., Jr (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol* **42**, 166–170.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P. & Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**, 81–91.

Guo, Y., Zheng, W., Rong, X. & Huang, Y. (2008). A multilocus phylogeny of the *Streptomyces griseus* 16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. *Int J Syst Evol Microbiol* **58**, 149–159.

Haley, B. J., Grim, C. J., Hasan, N. A., Choi, S. Y., Chun, J., Brettin, T. S., Bruce, D. C., Challacombe, J. F., Detter, J. C. & other authors (2010). Comparative genomic analysis reveals evidence of two novel *Vibrio* species closely related to *V. cholerae*. *BMC Microbiol* **10**, 154.

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K. & Schuster, S. C. (2005). Whole-genome prokaryotic phylogeny. *Bioinformatics* **21**, 2329–2335.

Hofer, U. (2013). Environmental microbiology: exploring diversity with single-cell genomics. *Nat Rev Microbiol* **11**, 598–599.

Hoffmann, M., Monday, S. R., Allard, M. W., Strain, E. A., Whittaker, P., Naum, M., McCarthy, P. J., Lopez, J. V., Fischer, M. & Brown, E. W. (2012). *Vibrio caribbeanicus* sp. nov., isolated from the marine sponge *Scleritoderma cyanea*. *Int J Syst Evol Microbiol* **62**, 1736–1743.

Jeon, Y.-S., Lee, K., Park, S.-C., Kim, B.-S., Cho, Y.-J., Ha, S.-M. & Chun, J. (2014). EzEditor: a versatile sequence alignment editor for both ribosomal RNA and protein coding genes. *Int J Syst Evol Microbiol* **64** (in press).

Jiménez, G., Urdiain, M., Cifuentes, A., López-López, A., Blanch, A. R., Tamames, J., Kämpfer, P., Kolstø, A. B., Ramón, D. & other authors (2013). Description of *Bacillus toyonensis* sp. nov., a novel species of the *Bacillus cereus* group, and pairwise genome comparisons of the species of the group by means of ANI calculations. *Syst Appl Microbiol* **36**, 383–391.

Johnson, J. L. & Ordal, E. J. (1968). Deoxyribonucleic acid homology in bacterial taxonomy: effect of incubation temperature on reaction specificity. *J Bacteriol* **95**, 893–900.

Kent, W. J. (2002). BLAT – the BLAST-like alignment tool. *Genome Res* **12**, 656–664.

Kim, O. S., Cho, Y. J., Lee, K., Yoon, S. H., Kim, M., Na, H., Park, S. C., Jeon, Y. S., Lee, J. H. & other authors (2012). Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int J Syst Evol Microbiol* **62**, 716–721.

Kim, M., Oh, H.-S., Park, S.-C. & Chun, J. (2014a). Towards a taxanomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* **64**, 346–351.

Kim, K. M., Park, J.-H., Bhattacharya, D. & Yoon, H. S. (2014b). Applications of next-generation sequencing to unravelling the evolutionary history of algae. *Int J Syst Evol Microbiol* **64**, 333–345.

Konstantinidis, K. T. & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* **102**, 2567–2572.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R. & other authors (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol* **30**, 693–700.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12.

Kurtzman, C. P. (2014). Use of gene sequence analyses and genome comparisons for yeast systematics. *Int J Syst Evol Microbiol* **64**, 325–332.

Lang, J. M., Darling, A. E. & Eisen, J. A. (2013). Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PLoS ONE* **8**, e62510.

Lee, K., Park, S. C., Yi, H. & Chun, J. (2013). *Flavobacterium limnosediminis* sp. nov., isolated from sediment of a freshwater lake. *Int J Syst Evol Microbiol* **63**, 4784–4789.

Löffler, F. E., Yan, J., Ritalahti, K. M., Adrian, L., Edwards, E. A., Konstantinidis, K. T., Müller, J. A., Fullerton, H., Zinder, S. H. & Spormann, A. M. (2013). *Dehalococcoides mccartyi* gen. nov., sp. nov., obligately organohalide-respiring anaerobic bacteria relevant to halogen cycling and bioremediation, belong to a novel bacterial class, *Dehalococcoidia* classis nov., order *Dehalococcoidales* ord. nov. and family *Dehalococcoidaceae* fam. nov., within the phylum *Chloroflexi*. *Int J Syst Evol Microbiol* **63**, 625–635.

Lucena, T., Ruvira, M. A., Arahal, D. R., Macián, M. C. & Pujalte, M. J. (2012). *Vibrio aestivus* sp. nov. and *Vibrio quintilis* sp. nov., related to Marisflavi and Gazogenes clades, respectively. *Syst Appl Microbiol* **35**, 427–431.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K. & other authors (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**, 3140–3145.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J. & other authors (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.

Marrero, G., Schneider, K. L., Jenkins, D. M. & Alvarez, A. M. (2013). Phylogeny and classification of *Dickeya* based on multilocus sequence analysis. *Int J Syst Evol Microbiol* **63**, 3524–3539.

Martens, M., Delaere, M., Coopman, R., De Vos, P., Gillis, M. & Willems, A. (2007). Multilocus sequence analysis of *Ensifer* and related taxa. *Int J Syst Evol Microbiol* **57**, 489–503.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P. & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14**, 60.

Meier-Kolthoff, J. P., Klenk, H.-P. & Göker, M. (2014). Taxanomic use of DNA G+C content and DNA-DNA hybridization in the genomic age. *Int J Syst Evol Microbiol* **64**, 352–356.

Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. (2013). Accurate and universal delineation of prokaryotic species. *Nat Methods* **10**, 881–884.

Naushad, H. S., Lee, B. & Gupta, R. S. (2014). Conserved signature indels and signature proteins as novel tools for understanding microbioal phylogeny and systematics: identification of molecular signatures that are specific for the phytopathogenic genera *Dickeya, Pectobacterium* and *Brenneria*. *Int J Syst Evol Microbiol* **64**, 366–383.

Oren, A. & Papke, R. T. (2010). *Molecular Phylogeny of Microorganisms*. Wymondham, UK: Caister Academic Press.

Ramasamy, D., Mishra, A. K., Lagier, J.-C., Padhmanabhan, R., Rossi, M., Sentausa, E., Raoult, D. & Fournier, P.-E. (2014). A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int J Syst Evol Microbiol* **64**, 384–391.

Richter, M. & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* **106**, 19126–19131.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., Darling, A., Malfatti, S., Swan, B. K. & other authors (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437.

Rosselló-Mora, R. (2006). DNA-DNA reassociation methods applied to microbial taxonomy and their critical evaluation. In *Molecular Identification, Systematics and Population Structure of Prokaryotes*, pp. 23–50. Edited by E. Stackebrandt. Berlin, Heidelberg: Springer.

Rosselló-Mora, R. & Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiol Rev* **25**, 39–67.

Ruvira, M. A., Lucena, T., Pujalte, M. J., Arahal, D. R. & Macián, M. C. (2013). *Marinifilum flexuosum* sp. nov., a new *Bacteroidetes* isolated from coastal Mediterranean Sea water and emended description of the genus *Marinifilum* Na *et al.*, 2009. *Syst Appl Microbiol* **36**, 155–159.

Sokal, R. R. & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*. San Francisco: W. H. Freeman.

Stackebrandt, E. & Ebers, J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiol Today* **33**, 152–155.

Stackebrandt, E. & Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* **44**, 846–849.

Stackebrandt, E., Frederiksen, W., Garrity, G. M., Grimont, P. A. D., Kämpfer, P., Maiden, M. C. J., Nesme, X., Rosselló-Mora, R., Swings, J. & other authors (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.

Sullivan, C. B., Diggle, M. A. & Clarke, S. C. (2005). Multilocus sequence typing: data analysis in clinical microbiology and public health. *Mol Biotechnol* **29**, 245–254.

Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., González, J. M., Luo, H., Wright, J. J., Landry, Z. C. & other authors (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci U S A* **110**, 11463–11468.

Tindall, B. J., Rosselló-Móra, R., Busse, H. J., Ludwig, W. & Kämpfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *Int J Syst Evol Microbiol* **60**, 249–266.

Vanlaere, E., Baldwin, A., Gevers, D., Henry, D., De Brandt, E., LiPuma, J. J., Mahenthiralingam, E., Speert, D. P., Dowson, C. & Vandamme, P. (2009). Taxon K, a complex within the *Burkholderia cepacia* complex, comprises at least two novel species, *Burkholderia contaminans* sp. nov. and *Burkholderia lata* sp. nov. *Int J Syst Evol Microbiol* **59**, 102–111.

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E. & other authors (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Bacteriol* **37**, 463–464.

Whitman, W. B., Coleman, D. C. & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95**, 6578–6583.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., Kunin, V., Goodwin, L., Wu, M. & other authors (2009). A phylogeny-driven genomic encyclopaedia of *Bacteria* and *Archaea*. *Nature* **462**, 1056–1060.

Yarza, P., Spröer, C., Swiderski, J., Mrotzek, N., Spring, S., Tindall, B. J., Gronow, S., Pukall, R., Klenk, H. P. & other authors (2013). Sequencing orphan species initiative (SOS): filling the gaps in the 16S rRNA gene sequence database for all species with validly published names. *Syst Appl Microbiol* **36**, 69–73.

Yi, H., Cho, Y. J., Yoon, S. H., Park, S. C. & Chun, J. (2012). Comparative genomics of *Neisseria weaveri* clarifies the taxonomy of this species and identifies genetic determinants that may be associated with virulence. *FEMS Microbiol Lett* **328**, 100–105.