Cautionary tale of using 16S rRNA gene sequence similarity values in identification of human-associated bacterial species

Morgane Rossi-Tamisier, Samia Benamar, Didier Raoult and Pierre-Edouard Fournier

Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes, UM 63, CNRS 7278, IRD 198, Institut Hospitalo-Universitaire Méditerranée-Infection, Faculté de médecine, Aix-Marseille Université, 27 Boulevard Jean Moulin, 13385 Marseille cedex 05, France

Modern bacterial taxonomy is based on a polyphasic approach that combines phenotypic and genotypic characteristics, including 16S rRNA sequence similarity. However, the 95 % (for genus) and 98.7 % (for species) sequence similarity thresholds that are currently recommended to classify bacterial isolates were defined by comparison of a limited number of bacterial species, and may not apply to many genera that contain human-associated species. For each of 158 bacterial genera containing human-associated species, we computed pairwise sequence similarities between all species that have names with standing in nomenclature and then analysed the results, considering as abnormal any similarity value lower than 95 % or greater than 98.7 %. Many of the current bacterial species with validly published names do not respect the 95 and 98.7 % thresholds, with 57.1 % of species exhibiting 16S rRNA gene sequence similarity rates <98.7 %, and 60.1 % of genera containing species exhibiting a 16S rRNA gene sequence similarity rate <95 %. In only 17 of the 158 genera studied (10.8 %), all species respected the 95 and 98.7 % thresholds. As we need powerful and reliable taxonomical tools, and as potential new tools such as pan-genomics have not yet been fully evaluated for taxonomic purposes, we propose to use as thresholds, genus by genus, the minimum and maximum similarity values observed among species.

INTRODUCTION

Taxonomy enables researchers to classify bacterial isolates into existing or potentially new taxa. Bacterial taxonomy is a dynamic discipline and the criteria used have evolved, following scientific and technical progresses. Currently, the classification of a bacterial isolate is based on a polyphasic approach that combines phenotypic and genotypic characteristics, although the tools used have varied over time. In 1996, Vandamme et al. suggested that polyphasic taxonomy should include DNA G+C content difference, 16S rRNA gene sequence similarity (Fox et al., 1977, 1992; Hugenholtz et al., 1998; Ludwig & Klenk, 2001), DNA–DNA hybridization (DDH) (Brenner et al., 1969; Johnson, 1991), and chemotaxonomic criteria, in particular fatty acid analysis (Vandamme et al., 1996). In 2002, Stackebrandt et al. proposed to add whole-genome profiling (AFLP, RAPD, Rep-PCR, PFGE), multi-locus sequence typing, MALDI-TOF mass spectrometry, Fourier-transformed infrared spectroscopy and/or pyrolysis-mass spectrometry to the previously cited criteria (Stackebrandt et al., 2002). In 2010, Tindall et al. re-evaluated the various available methods and proposed a combination of 16S rRNA gene sequence similarity and phylogeny, genomic DNA G+C content, DDH, cell morphology and Gram-staining properties, as well as chemotaxonomic criteria such as peptidoglycan structure and composition of respiratory lipoquinones and hydrophobic side-chains of lipids or fatty acids (Tindall et al., 2010). However, the species rank is especially difficult to define, as the names of many bacterial species with standing in nomenclature have been validly published on the basis of arbitrary criteria (Rossello-Mora, 2003).

The 16S rRNA gene is a highly conserved gene that is made up of nine hypervariable domains separated by more preserved fragments that can be used to design universal primers. More than three million 16S rRNA gene sequences are currently available in databases (Quast et al., 2013). The use of 16S rRNA gene sequences for taxonomic purposes began in the 1980s, when the development of molecular tools, notably automatic sequencing (Sentausa & Fournier, 2013), enabled the development of large sequence databases.
such as GenBank (Benson et al., 2013), Greengenes (DeSantis et al., 2006), SILVA (Quast et al., 2013) and EzTaxon (Kim et al., 2012), to which a sequence obtained from an isolate may be compared using BLAST (Altschul et al., 1990; Morgulis et al., 2008). Specific rRNA analysis tools have been developed as well, such as ARB (Ludwig et al., 2004).

The massive use of 16S rRNA gene sequences notably had a very significant impact on the number of bacterial species with validly published names, which grew from 1800 in 1980 to nearly 12 500 in 2013 (Parte, 2014).

Many pre-existing taxa were reclassified and a greater number created. In order to facilitate the use of 16S rRNA gene sequences as a genotyping tool, cut-off values have been defined. As early as 1994, two strains were considered as belonging to distinct species if they shared 16S rRNA gene sequence similarities lower than 97% (Stackebrandt & Goebel, 1994), and to discriminate two genera if this value was lower than 95%. In 2006, the cut-off value at the species level was re-evaluated at 98.7% (Stackebrandt & Ebers, 2006). However, several authors have demonstrated that these cut-offs, initially designed to standardize the use of 16S rRNA gene sequences in taxonomy, do not apply to several genera. Extreme examples include species of the genus Edwardsiella (99.3 to 99.8% 16S rRNA gene sequence similarity; Janda & Abbott, 2007) and the genera Streptomyces and Chlorobium in which the lowest inter-species similarities are 78 and 86.1%, respectively (Alexander et al., 2002). Among genera containing human-associated species, several inter-species 16S rRNA gene sequence similarity values <95% or >98.7% were observed in the genus Clostridium. As an example, Clostridium botulinum and Clostridium sporogenes exhibit a 99.7% similarity. However, the species Clostridium sporogenes was conserved as it groups non-toxicogenic strains related genetically to Clostridium botulinum (Olsen et al., 1995). Another example may be found in the genus Rickettsia in which 16S rRNA gene sequence similarity values >99% are found among all 26 species that have names with standing in nomenclature, despite species-specific epidemi-o-clinical characteristics (Fournier & Raoult, 2009). Thus, in some cases, phenotypic and/or chemotypic specificities justify a distinct classification of bacterial strains into species or genera from that suggested by their genetic relatedness. These examples question the use of strict criteria for the interpretation of the results of 16S rRNA-based metagenomic studies in which phenotypic properties are not taken into account (Huse et al., 2008). Herein, as no broad analysis of inter-species 16S rRNA gene sequence similarity among bacterial genera has been performed to date, we studied the variability of this parameter among bacterial genera containing species of medical interest (i.e. species found in humans, pathogenic or not) that have names with standing in nomenclature.

**METHODS**

**Selection of studied bacterial genera and of 16S rRNA gene sequences.** Within the List of Prokaryotic Names with Standing in Nomenclature (LPSN) website (www.bacterio.net; Parte, 2014), we selected bacterial genera that contained at least one species associated with humans, pathogenic or not, and obtained the sequence accession numbers for the 16S rRNA gene from their type strains.

**Collection of 16S rRNA gene sequences.** For every studied genus, we created a FASTA format file containing the 16S rRNA gene sequences of the type strain from each species. Due to the wide heterogeneity in length and quality of the 16S rRNA gene sequences of type strains, we did not use sequences shorter than 1200 nt. We also removed sequences shorter than 1250 nt with more than six degenerate nucleotides, and sequences larger than 1250 nt with more than ten degenerate nucleotides.

**16S rRNA gene sequence analysis: calculation of pairwise 16S rRNA gene sequence similarities.** Sequences were aligned using CLUSTAL Omega with default settings (Chenna et al., 2003; Goujon et al., 2010). In each studied genus, pairwise 16S rRNA gene sequence similarities between all possible species pairs were first estimated using the MEGA phylogeny software version 5 (Tamura et al., 2011). Then, the highest and the lowest values computed by this software were more accurately determined using pairwise BLASTN. We defined as expected values (EV) inter-species 16S rRNA gene sequence similarity percentages that were between 95 and 98.7% (Stackebrandt & Ebers, 2006), and as abnormal values (AV) those that were >98.7 or <95%.

**RESULTS**

Fifty-six species, for which 16S rRNA gene sequence accession numbers were not available on the LPSN website, were not included in the study. Another 74 species, including seven that were shorter than 1200 nt and 67 that were of insufficient quality, were also excluded from the analysis. Overall, 4289 species that have names with standing in nomenclature, classified within 158 genera, 81 families, 39 orders and 12 phyla were included in our study (Table S1, available in the online Supplementary Material).

Among the 4289 studied species, the lowest and highest inter-species pairwise 16S rRNA gene sequence similarities were 68.7 and 100%, respectively. Of the 4289 studied species, 2448 (57.1%) exhibited at least one 16S rRNA gene sequence similarity rate ≥98.7% with another species in the same genus. In 95 of 158 studied genera (60.1%) at least one species exhibited at least one 16S rRNA gene sequence similarity rate <95% with another species in the same genus. In only 17 of the 158 studied genera (10.8%), all species exhibited EVs. These included genera of major medical interest such as Morganella and Chlamydia (Table S1).

Sixty-eight of the 158 studied genera (43.0%) had 50 to 100% EVs, including Kingella (50%), Corynebacterium (53.1%), Bordetella (53.6%), Bartonella (56.9%), Acinetobacter (61.3%), Neisseria (66.7%), Borrelia (70.0%), Yersinia (72.1%), Enterobacter (78.4%), Escherichia (80.0%), Burkholderia (82.1%), Vibrio (82.7%), Streptomyces (83.5%), Mycobacterium (91.1%), Staphylococcus (92.8%) and Nocardia (93.9%) (Fig. 1). It should be noted that bacterial genera from the phylum Actinobacteria [24 of the 33 studied genera (72.7%)] and the classes Alphaproteobacteria (11/20, 55%), Betaproteobacteria (9/15, 60%) and Gammaproteobacteria...
(24/37, 64.8 %) were more likely to have >50 % EVs. In addition, within the family Enterobacteriaceae, 15/19 genera (78.9 %) showed >50 % EVs (Fig. 2, Table S1).

Forty-four of the 158 studied genera (27.8 %) exhibited 20 to 50 % EVs, including Actinomyces (22.0 %), Campylobacter (22.1 %), Legionella (29.9 %), Streptococcus (33.2 %), Rickettsia (35.7 %), Bacillus (39.4 %), Listeria (46.4 %), Bifidobacterium (44.7 %), Pseudomonas (46.5 %) and Helicobacter (49.0 %) (Fig. 1, Table S1).

Twenty-nine of the 158 studied genera (18.4 %) had less than 20 % EVs, including Brucella (0.0 %), Treponema (2.1 %), Clostridium (6.4 %), Mycoplasma (8.4 %), Bacteroides (9.6 %) and Shigella (16.7 %) (Fig. 1). The order Clostridiales and the phylum Tenericutes contained the genera with the lowest rates of EVs, with six of eight genera (75 %) and two of three genera (66.6 %), respectively, having less than 20 % EVs. None of the studied genera of the phylum Tenericutes had more than 50 % EVs (Fig. 2, Table S1). Noteworthy, two out of five genera in the phylum Spirochaetes had very few EVs: Spirochaeta (1.5 %) and Treponema (2.1 %) (Table S1).

### DISCUSSION

Among the genetic tools used to classify bacterial strains, 16S rRNA gene sequence analysis had the most significant impact and enabled the reclassification or creation of a large number of taxa. However, the cut-off values of 95 and 98.7 % used to classify bacterial isolates at the genus and species levels, respectively, were established under the assumption that the level of inter-species 16S rRNA gene sequence variation was homogeneous among genera. However, it was suggested that the speed of evolution of rRNA genes may vary according to the phylum (Clarridge, 2004), and the inadequacy of the current cut-offs for many genera has been reported. It was notably demonstrated that the discriminatory power of 16S rRNA gene sequences could be insufficient at the species level (Bosshard et al., 2006; Mignard & Flandrois, 2006), as observed for Streptococcus pneumoniae and Streptococcus mitis that exhibit as few as 3 bp differences. Conversely, intra-species 16S rRNA gene sequence differences may be important, as is the case for Enterobacter agglomerans for which two strains may exhibit as many as 27 bp differences, which does not validate the 98.7 % cut-off, and thus may justify...
their classification in distinct species. Such a genetic difference was also observed at the genus level, with Clostridium tetani and Clostridium innocuum exhibiting 104 bp differences, which does not validate the 95% threshold and thus may justify the classification of both species in distinct genera. Another drawback is the sequence variability among the multiple copies of 16S rRNA genes present in a bacterial chromosome. In 1997, Wang et al. reported a degree of sequence similarity among 16S rRNA genes copies of Thermobispora bispora of only 92% (Wang et al., 1997). In 2010, Pei et al. identified sequence diversity > 1.3% among 16S rRNA genes in a genome in 11 bacterial species (Pei et al., 2010). Among these, Borrelia afzelii, an agent of Lyme disease in humans, exhibits a similarity of only 79.62% between its two 16S rRNA gene copies (Pei et al., 2010). Thus, a strict application of the 98.7% threshold would classify these bacteria in different species depending on the 16S rRNA gene copy analysed (Acinas et al., 2004; Pei et al., 2010).

However, although variations of inter-species 16S rRNA gene sequence similarity values among genera had been reported, we could not find any article studying this phenomenon at a large scale. In the present study, using a systematic analysis of inter-species 16S rRNA gene sequence similarity of genera containing bacteria of medical interest, we demonstrated that the current thresholds are applicable to only 42.9% of studied species and 39.9% of studied genera. In addition, when both cut-off values were taken into account, only 10.8% of studied genera had all their species adequately classified. We also observed important discrepancies among bacterial genera, some of which contain species of major medical importance. The proportion of species exhibiting EVs ranged from 0 (Brucella, Tissierella) to 100% (Kytooccus, Chlorobium), and there were discrepancies regarding these rates within most of the studied phyla (Fig. 1, Table S1).

The phylum Chlorobi was the only phylum within which the 95 and 98.7% thresholds were respected for all studied taxa, but only one genus, Chlorobium, was studied. Other taxa, including the Chlamydia, Actinobacteria, Bacilli, Alphaproteobacteria and Gammaproteobacteria (and more specifically the order Enterobacteriales) contained many genera for which both thresholds were respected. Additionally, the cut-offs were globally well-adapted for some genera such as Bordetella (53.6%), Bartonella (56.9%), Acinetobacter (61.3%), Enterococcus (63.2%), Neisseria (66.7%), Borrelia (70%), Yersinia (72.1%), Enterobacter (78.4%), Escherichia (80.0%), Burkholderia (82.1%), Vibrio (82.7%), Streptomyces (83.5%), Mycobacterium (91.2%), Staphylococcus (92.3%), Klebsiella (93.3%), Nocardia (93.9%) and Chlamydia (100%). In contrast, the current cut-offs were poorly applicable to the order Clostridiales and phyla Bacteroidetes and Spirochaetes. In the order Clostridiales and phyla Bacteroidetes, Tenericutes and Spirochaetes no genus exhibited 100% EVs (Fig. 2). We acknowledge the fact that the limited number of studied genera of the phyla Tenericutes and Spirochaetes may in part explain these results. Additionally, at the genus level, some genera had few EVs, including Brucella (0%), Treponema (2.1%), Clostridium (6.4%), Mycoplasma (8.4%), Pasteurella (15.4%), Shigella (16.7%), Actinomyces (22%), Campylobacter (22.1%), Haemophilus (26.9%), Legionella (29.9%), Streptococcus (33.2%), Rickettsia (35.7%), Bacillus (39.4%), Leptospira (43.3%), Listeria (46.4%) and Pseudomonas (46.5%) (Fig. 1, Table S1). For each studied genus, the ranges of 16S rRNA gene sequence similarities have been made available in a free-

**Fig. 2.** Distribution of genera that respect current thresholds (95% and 98.7%) according to orders or classes. For example, in the phylum Actinobacteria, 70% of genera have > 50% inter-species sequence similarity rates between the two thresholds and 12% of genera have < 20% inter-species sequence similarity rates between the two thresholds. The phylum Bacteroidetes shows a different pattern, as 25% of genera have > 50% inter-species sequence similarity rates between the two thresholds and 63% of genera have < 20% inter-species sequence similarity rates between the two thresholds.

Therefore, on the basis of these results and observed variations, we confirm that the 95 and 98.7 % inter-species 16S rRNA gene sequence similarity thresholds may only be used as indicators, and not as a definite tool for the classification of bacterial strains. This was already the case for several bacterial species exhibiting specific phenotypic characteristics but inter-species 16S rRNA gene sequence similarity values greater than 98.7 %, such as Clostridium botulinum and Clostridium sporogenes (Olsen et al., 1995), Rickettsia prowazekii and Rickettsia rickettsii (Fournier & Raoul, 2009), or Nocardia paucivorans and Nocardia brevicatena (Roth et al., 2003). We recently proposed to include genomic data in the taxonomic classification of new bacterial isolates (Ramasamy et al., 2014). In this strategy, we systematically compare the 16S rRNA gene sequence similarity values obtained between the new isolate and its closest phylogenetic neighbour to those observed among members of the same genus and/or closely related genera. Our findings in the present study support our strategy of computing extreme 16S rRNA gene sequence similarity values, to better assign a strain within a species or a genus, and suggest that applying strict thresholds developed by studying a limited number of taxa may lead to misclassification and contribute to confusing situations.

In conclusion, the degree of conservation of 16S rRNA gene sequence similarity may vary greatly among bacterial genera. Therefore, rather than just referring to two cut-off values, we propose that when classifying a new strain, in addition to its phenotypic characteristics, its 16S rRNA gene sequence similarity to its phylogenetically closest species should be compared, in the studied genus, to the range of similarity values observed among species that have names with standing in nomenclature.

REFERENCES


