INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

RESEARCH ARTICLE
Chun *et al*., *Int J Syst Evol Microbiol* 2018;68:461–466
DOI 10.1099/ijsem.0.002516

MICROBIOLOGY SOCIETY
OPEN MICROBIOLOGY

Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes

Jongsik Chun,[1,*] Aharon Oren,[2] Antonio Ventosa,[3] Henrik Christensen,[4] David Ruiz Arahal,[5] Milton S. da Costa,[6] Alejandro P. Rooney,[7] Hana Yi,[8] Xue-Wei Xu,[9] Sofie De Meyer[10] and Martha E. Trujillo[11,*]

Abstract

Advancement of DNA sequencing technology allows the routine use of genome sequences in the various fields of microbiology. The information held in genome sequences proved to provide objective and reliable means in the taxonomy of prokaryotes. Here, we describe the minimal standards for the quality of genome sequences and how they can be applied for taxonomic purposes.

INTRODUCTION

One of the ultimate goals of microbial taxonomy is to devise a process of classification and identification that is stable, objective and readily usable by those who do not have special skills. Given the vast diversity of prokaryotes in nature, an ability to build a database that is searchable and comparable is also a fundamental feature for the next generation taxonomy [1, 2].

In this context, genomics has become a promising methodology as it provides a reproducible, reliable, highly informative means to infer phylogenetic relationships among prokaryotes which allows the continuation of our tradition to natural classification [3].

The replacement of DNA–DNA hybridization (DDH) as ‘the gold standard’ in prokaryote taxonomy with pairwise genome-sequence derived similarity has been proposed by several authors [4–10]. There is sufficient experimental evidence to adopt this proposal, which is supported by DNA sequencing platforms that generate high throughput data with low cost and high quality as well as adequate bioinformatics tools for classification and identification of prokaryotes. The aim of this article is to provide a general guideline to apply genome sequence data to taxonomic purposes and propose the minimal standards of quality for genome sequence data.

Use of whole genome sequence data in delineating new species

There have been a series of efforts to develop a bioinformatic method to replace DDH for differentiating species. Because the DDH value basically reflects relatedness or similarity between two genomes, these efforts focused on devising values analogous to DDH values. These values, as forms of similarity or distance, were coined as the overall genome related index (OGRI) [3].

Like DDH, OGRIs can be used to check if a strain belongs to a known species by calculating the relatedness between genome sequences of the strains and type strain of a species. Average nucleotide identity (ANI) and digital DDH (dDDH) have been most widely used, and relevant software tools are readily available as web-services and as standalone tools (Table 1). The proposed and generally accepted species boundary for ANI and dDDH values are 95~96 and 70 %, respectively [4, 5, 7]. Even though there has been a considerable effort in obtaining genome data for type strains, less than 50 % of species with validly published names are represented by genome sequences of their type strains as of the time of writing. In

Author affiliations: [1]Department of Biological Sciences and Institute of Molecular Biology and Genetics, Seoul National University, Seoul 08826, Republic of Korea; [2]The Institute of Life Sciences, The Hebrew University of Jerusalem, The Edmond J. Safra Campus, 91904 Jerusalem, Israel; [3]Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Sevilla, 41012 Sevilla, Spain; [4]Department of Veterinary Disease Biology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; [5]Departamento de Microbiología y Ecología, and Colección Española de Cultivos Tipo (CECT), Universitat de València, València, Spain; [6]Center for Neurosciences and Cell Biology, University of Coimbra, 3004-504 Coimbra, Portugal; [7]U.S. Department of Agriculture, Agricultural Research Service, Crop Bioprotection Research Unit, Peoria, IL 61604, USA; [8]Department of Public Health Sciences, Graduate School, Korea University, Seoul, Republic of Korea; [9]Laboratory of Marine Ecosystem and Biogeochemistry, Second Institute of Oceanography, State Oceanic Administration, Hangzhou 310012, PR China; [10]Centre for Rhizobium Studies, Murdoch University, Murdoch, WA, Australia; [11]Departamento de Microbiología y Genética, Campus Miguel de Unamuno, Universidad de Salamanca, 37007, Salamanca, Spain.
*Correspondence: Jongsik Chun, jchun@snu.ac.kr; Martha E. Trujillo, mett@usal.es

Keywords: genome sequence; minimal standards; prokaryotic taxonomy; average nucleotide identity; phylogenomics.

Abbreviations: ANI, average nucleotide identity; dDDH, digital DNA–DNA hybridization; DDH, DNA–DNA hybridization; GGDC, Genome-to-Genome Distance Calculator; MLSA, multilocus sequence analysis; NGS, next generation sequencing; OGRI, overall genome related index.

Table 1. Web-services and standalone software tools for taxonomic purposes

Algorithm	Function	Type	URL/Reference
OrthoANI with usearch	Calculation of ANI	Standalone	https://www.ezbiocloud.net/tools/orthoaniu [9]
OrthoANI with usearch	Calculation of ANI	Web service	https://www.ezbiocloud.net/tools/ani [9]
Genome-to-Genome Distance Calculator	Calculation of dDDH	Web service	http://ggdc.dsmz.de/ggdc.php/ [7]
ANI calculator	Calculation of ANI	Web service	http://enve-omics.ce.gatech.edu/ani/
JSpecies	Calculation of ANI	Standalone	http://imedea.uib-csic.es/jspecies/ [5]
JSpeciesWS	Calculation of ANI	Web service	http://jspecies.ribohost.com/ [30]
CheckM	Checking contamination	Standalone	http://ecogenomics.github.io/CheckM/ [29]
ContEst16S	Checking contamination	Web service	https://www.ezbiocloud.net/tools/contest16s [28]
BBMap	Calculation of sequencing depth of coverage	Standalone	https://sourceforge.net/projects/bbmap/
Amphora2	Phylogenomic treeing	Standalone	http://wolbachia.biology.virginia.edu/WuLab/Software.html [21]
BIGSdb	Phylogenomic treeing	Standalone	https://pubmlst.org/software/database/bigsdb/ [31]
bcgTree	Phylogenomic treeing	Standalone	https://github.com/iimog/bcgTree [32]
Phylophlan	Phylogenomic treeing	Standalone	https://huttenhower.sph.harvard.edu/phylophlan[22]
UBCG	Phylogenomic treeing	Standalone	https://www.ezbiocloud.net/tools/ubcg

contrast, an almost complete database of 16S rRNA gene (16S) sequences is available for the type strains of prokaryotic species [11, 12]. Therefore, a combination of 16S similarity and OGRI can be used in a systematic process to identify and recognize a new species (Fig. 1). In this two-step approach, the list of species that is required to compare to the strain in question using genome sequences is obtained using a 16S-based search [11]; only species showing 98.7 % or higher 16S similarity are selected for calculating OGRIs [7, 13, 14]. It is noteworthy that the use of a 98.7 % cutoff, higher than the previously accepted 97 % [15], requires assurance in the quality of 16S sequences under consideration [14]. If genome sequence data of the type strains of the hit species (showing ≥98.7 % 16S similarity) are not available, it is recommended to obtain their genome sequences, not only to measure OGRIs but also to extend and improve the public genome database for taxonomic purposes.

Use of genome data in recognizing subspecies

At this stage, we do not have sufficient data to provide a general guideline for defining subspecies using genome data. However, a good practice should involve the following criteria: (i) OGRIs between subspecies and other species should be lower than the species-level cutoff value, (ii) OGRIs between subspecies should be higher than the species-level cutoff, (iii) strains belonging to different subspecies should be genomically coherent and form distinguishable clades by OGRIs and phylogenomic treeing, (iv) subspecies should be differentiated by a sufficient number of phenotypes, and (v) there should be a sound rationale why subspecies should be created and separately recognized, such as showing different host specificity in the case of pathogens.

Use of genome data in the classification of genera and higher taxa

A genome sequence of a prokaryote contains as little as a few hundreds to over ten thousand genes (www.ezbiocloud.net/dashboard). Every gene in the genome has its own evolutionary history, and these histories may contradict each other. Phylogenetic methods assume that evolution follows a tree-like structure and although it can easily depict the evolutionary history of a single gene these methods have difficulty in reconciling the conflicting signals of reticulate evolution in nucleotide sequence data. Many genes have undergone horizontal transfer events and it is a daunting task to elucidate the phylogenetic relationships amongst genomes.

Most of the widely used phylogenetic methods were developed to infer the phylogeny of a gene, but not of whole genome sequence. Because many genes have undergone horizontal transfer events, it is a daunting task to elucidate precise phylogenetic relationships among genomes.

A generally accepted process to infer whole genome phylogeny is to use multiple genes that are thought to be orthologous (or unlikely underwent lateral transfer events). Functionally important genes, such as those encoding ribosomal proteins, are usually selected for such phylogenomic treeing. The number of chosen genes varies depending on the taxonomic scope of the study and on the algorithm used to select orthologous, usually single-copy, genes [16–18].

Because OGRI does not have a taxonomic resolution above the species level, a multigene-based phylogenomic treeing approach should be the choice for defining genera or higher taxa, complementing 16S-derived phylogeny. It can be differentiated from a similar method called multilocus sequence analysis (MLSA) [19, 20] by applying a substantially higher number of orthologous genes that were rationally selected using large scale comparative genomics [21–23]. The combination of phylogenomic treeing and highly conserved phenotypes, including chemotaxonomic markers, should play a significant role in the classification of genera and higher taxa. We expect this approach will, in the future, help resolve the poorly classified taxa and enable us to move towards a more standardized, balanced classification scheme across different phyla.

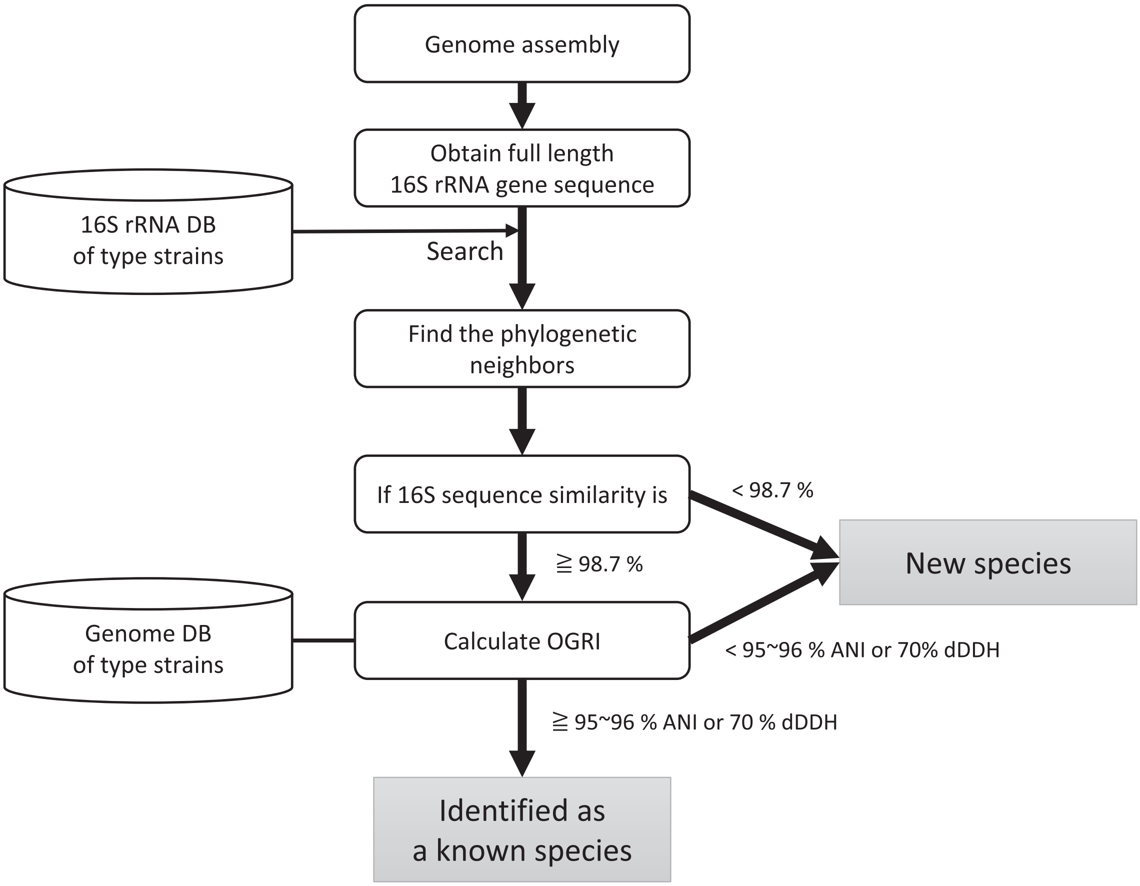
Genome assembly
Obtain full length 16S rRNA gene sequence
16S rRNA DB of type strains
Search
Find the phylogenetic neighbors
If 16S sequence similarity is
< 98.7 %
≧ 98.7 %
New species
Genome DB of type strains
Calculate OGRI
< 95~96 % ANI or 70% dDDH
≧ 95~96 % ANI or 70 % dDDH
Identified as a known species

Fig. 1. Workflow of genome based classification at the species level. To recognize new genera, phylogenomic treeing should be used.

PROPOSED MINIMAL STANDARDS FOR THE USE OF GENOME SEQUENCE DATA FOR TAXONOMIC PURPOSES

DNA sequencing platforms

In the last decade, several next generation sequencing platforms were commercially introduced and proved to provide adequate genome data for taxonomic purposes in microbiology [3, 24, 25]. Even though different microbiological disciplines require different standards in accuracy and quality of final assembled contigs, genome sequences that are generated for taxonomic purposes should also serve for the other fields such as environmental and clinical microbiology. It is clear that new next generation sequencing (NGS) platforms will be invented and become available in the near future, some, if not all, of which will meet the quality requirement for such a purpose. At present, DNA sequencing platforms provided by Illumina (USA), Ion Torrent (Thermo Fisher Scientific, USA) and Pacific Biosciences (USA) have been shown to generate DNA sequence data that meet the general standards, if used with adequate experimental protocols. Any other NGS platform that will be available in the future should be subject to rigorous evaluation before it can be used in prokaryotic taxonomic studies.

Quality of raw NGS data and assembled genome sequences

In general, NGS platforms provide their own means of accessing sequencing quality of raw data. Because these were usually designed to match the sequencing quality statistics originally developed for the Sanger method [26], they can be used to compare the quality of raw sequence data generated by different NGS platforms to some extent. However, the important statistic is the quality of the final assembly, not that of the raw data.

Individual NGS reads of low quality are usually filtered out before the genome assembly process. Various software tools can be used to assemble the filtered raw reads into contigs [27]. It is evident that completion of genome sequences without any undetermined bases provides the best quality over a draft status of the genome assembly which results in the form of contigs. However, fragmented assemblies also provide sufficient information for taxonomic purposes if sequencing is carried out with a sufficient redundancy.

Several statistical parameters have been developed and used to describe the quality of the final genome assembly. The indices below are recommended to make sure that the quality of a genome sequence is suitable for taxonomic purposes:

- *Genome size*. The genome size is defined as the length sum of all contigs. If the genome sequence is not completely determined, this value represents only an approximation.
- *The number of contigs and N50*. A genome assembly process results in contigs of various lengths. It is general practice to exclude very short contigs from the final assembly. Because there is no clear standard in how to select the contigs, the number of contigs cannot be a good indicator of the quality of the genome sequence. Instead, one can use N50 which is known to give a better assessment of the final assembly. If the lengths of the contigs are summed from the largest to shortest, the N50 is defined as the length of the shortest contig that accumulatively show 50 % or more of the genome size.
- *Sequencing depth of coverage*. This value is usually expressed as the folds (e.g. 40.5X means that each base in the final assembly was read in 40.5 times on average). This statistic can be measured for all DNA sequencing platforms with adequate genome assembler software. It may be difficult to recommend a single value for all NGS platforms which have different accuracy and read-lengths. Theoretically, the more sequencing reads are generated, the better the quality of the assembly is. Given the ever-decreasing cost of NGS, we recommend ≥50X for the currently available NGS platforms (Illumina, Ion Torrent, and Pacific Biosciences). Not all assembler software tools provide the value of sequencing depths for the resultant genome assembly. In such cases one can use a short-read mapping software, such as BBMap (Table 1), that map all quality-filtered short reads to the final assembly to precisely estimate the sequencing depth of coverage.

Authenticity of the genome assembly

Strains and their genome data are often mistakenly mislabelled in the process of genome sequencing, leading to wrong taxonomic interpretation. In part, this is because genome sequencing is carried out in central sequencing facilities where the chance of mislabelling and contamination is relatively high. It is therefore important to check if a genome sequence indeed belongs to the strain under investigation.

Because 16S sequences of almost all type strains of known species are available, this gene can be used to confirm the authenticity of the final genome assembly. In cases of species that have a very similar 16S sequence to the phylogenetic neighboring species, protein-coding genes, such as *gyr*B, *rpo*B and *rec*A, can be used to further support the authenticity of genome data. To achieve this, full-length 16S or protein-coding genes should be extracted from the genome assembly. If there are substantial differences among 16S sequences in multiple rRNA operons, this matter should be resolved by either a cloning experiment followed by the Sanger sequencing or complete genome sequencing. For describing new species, a full-length 16S sequence of type strain should be obtained by the conventional Sanger and compared with 16S sequence extracted from whole genome assembly to ensure the authenticity of genome data. In case that full-length 16S sequence cannot be extracted from the genome assembly, protein-coding genes can be used instead of 16S. In any case, the full-length 16S sequence must be determined and provided for the type strains.

Contamination in the genome assembly

The cost-effectiveness of NGS allows us to generate high-throughput data with a higher sequencing depth of coverage than the conventional Sanger sequencing. One of the drawbacks of this phenomenon is that contaminating DNA sequences, even in a minor amount, can be incorporated into the genome assembly. A recent survey using the 16S sequences showed that contamination events could occur in both culturing and DNA sequencing steps [28]. At present, only a few bioinformatic tools for detecting potential contaminations are available using 16S [28] and protein-coding [29] genes. Due to the high frequency of lateral gene transfer events in the realm of prokaryotes, interpreting results of checking contaminations using protein-coding genes should be carefully performed.

Bioinformatics for taxonomic purposes

Two major categories of bioinformatics analysis can be carried out for taxonomic purposes.

- *OGRI*. OGRI represents any measurements indicating how similar two genome sequences are [3]. It is a direct descendant of DDH which has been used to define the species boundary of prokaryotes and still serves as a gold standard in prokaryotic taxonomy. The taxonomic resolution of OGRI is limited to differentiate only closely related species. It is also worth noting that OGRI is not suitable for phylogenetic inference, especially at the suprageneric rank level. Among the OGRIs, average nucleotide identity (ANI) has been most widely used. Several software tools for calculating the original ANI and its improved versions were developed for taxonomic purposes [5, 8, 9, 30]. An alternative to ANI is digital DDH (Genome-to-Genome Distance Calculator; GGDC) which has also been widely used for taxonomic purposes [7]. It is recommended that authors who propose new species should provide OGRI values between the type strain of proposed species and type strains of related species that show ≥98.7 % 16S sequence similarity. The general procedure for genome-based classification is depicted in Fig. 1.
- *Phylogenomic treeing*. Given the plethora of information residing in genome sequences, we are now able to explore the more precise phylogenetic relationship at various taxonomic levels. Application of genome data to

phylogenetic analysis, called phylogenomic treeing, can be achieved by inferring phylogenetic trees on the basis of multiple genes, instead of a single gene such as 16S. Phylogenomics should provide a better taxonomic framework, especially at the genus and higher levels. We expect that poorly classified taxa can be reorganized using this phylogenomic approach in the future. This is an active area of research with different scientific views. Here we recommend the minimum number of genes to be 31, which is higher than that used in the traditional multilocus sequence analysis (MLSA). Software tools that can be used for phylogenomic treeing for which sets of 31 to 400 house-keeping core genes are available (Table 1).

Choice of reference genome data from the public domain

There are now many cases in which multiple genome sequences are available for the same type strains (>1200 species with validly published names at the time of writing). It is therefore important that authentic genome sequences of the best quality are chosen for OGRI and phylogenomic treeing. A recommended criterion for selecting genome sequences with the best quality among multiple choices is N50 static rather than the number of contigs. The sequencing depth of coverage can also be a useful measure, but this value is usually not available for the genome assemblies available in public databases. The reference genomes in the taxonomic study should be those of the type strains of the species under investigation.

Deposition of sequencing data

The final assembly should be deposited to GenBank/EMBL/DDBJ database. Since raw NGS data can be used for improving assembly by a third party, depositing raw NGS data can be useful for the scientific community. However, depositing raw NGS data is not mandatory. Depositing only raw NGS data (without the final assembly) should be avoided, as the genome assembly process can be very difficult to reproduce.

CONCLUSIONS

We recommend the following when a genome sequence is used for taxonomic purposes.

- The sequencing instrument, library reagents and method for genome assembly should be described in detail.
- At least the following statistics should be given for the final genome assembly: (i) the obtained genome size, (ii) DNA G+C ratio, (iii) the number of contigs, (iv) N50 and (v) the sequencing depth of coverage. We recommend at least ≥50X depth for Illumina, Ion Torrent and Pacific Biosciences DNA sequencing platforms. Potential contamination should be checked. Full-length 16S sequence of the proposed type strains should be determined by the Sanger sequencing method. To check the authenticity of the final genome assembly, full-length 16S sequence that was extracted from the genome assembly should be compared with that of the Sanger method. Alternatively, protein-coding gene sequences can be used to check the authenticity. In any case, the full-length 16S sequence must be provided for the type strain.
- For the proposal of new species, OGRI values should be calculated with all phylogenetically related species (Fig. 1). For the classification of genera or higher taxa, at least one method for phylogenomic treeing should be used in which at least 30 genes are included.
- The final genome assembly should be deposited in a publicly accessible database that requires no login process.

Funding information
The authors received no specific grant from any funding agency.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Whitman WB. The need for change: embracing the genome. In: Goodfellow M, Sutcliffe I and Chun J (editors). *New Approaches to Prokaryotic Systematics. Methods in Microbiology*, vol. 41. London: Academic Press; 2014. pp. 1–12.
2. Rosselló-Móra R, Amann R. Past and future species definitions for *Bacteria* and *Archaea*. *Syst Appl Microbiol* 2015;38:209–216.
3. Chun J, Rainey FA. Integrating genomics into the taxonomy and systematics of the *Bacteria* and *Archaea*. *Int J Syst Evol Microbiol* 2014;64:316–324.
4. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P *et al.* DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007;57:81–91.
5. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci USA* 2009;106:19126–19131.
6. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* 2005; 102:2567–2572.
7. Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 2013;14:60.
8. Lee I, Ouk Kim Y, Park SC, Chun J. OrthoANI: animproved algorithm and software for calculating average nucleotide identity. *Int J Syst Evol Microbiol* 2016;66:1100–1103.
9. Yoon SH, Ha SM, Lim J, Kwon S, Chun J. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek* 2017;110:1281–1286.
10. Colston SM, Fullmer MS, Beka L, Lamy B, Gogarten JP *et al.* Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using *Aeromonas* as a test case. *MBio* 2014;5: e02136.
11. Yoon SH, Ha SM, Kwon S, Lim J, Kim Y *et al.* Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int J Syst Evol Microbiol* 2017; 67:1613–1617.
12. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E *et al.* The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res* 2014;42:D643–D648.
13. Stackebrandt E, Ebers J. Taxonomic parameters revisited: tarnished gold standards. *Microbiol Today* 2006;33:152–155.

14. **Kim M, Oh HS, Park SC, Chun J.** Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* 2014;64:346–351.
15. **Stackebrandt E, Goebel BM.** Taxonomic Note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Evol Microbiol* 1994;44:846–849.
16. **Chun J, Grim CJ, Hasan NA, Lee JH, Choi SY *et al.*** Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc Natl Acad Sci USA* 2009;106:15442–15447.
17. **Sangal V, Goodfellow M, Jones AL, Schwalbe EC, Blom J *et al.*** Next-generation systematics: an innovative approach to resolve the structure of complex prokaryotic taxa. *Sci Rep* 2016;6:38392.
18. **Thompson CC, Vicente AC, Souza RC, Vasconcelos AT, Vesth T *et al.*** Genomic taxonomy of vibrios. *BMC Evol Biol* 2009;9:258.
19. **Glaeser SP, Kämpfer P.** Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Syst Appl Microbiol* 2015;38:237–245.
20. **Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ.** Universal trees based on large combined protein sequence data sets. *Nat Genet* 2001;28:281–285.
21. **Wu M, Scott AJ.** Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* 2012;28:1033–1034.
22. **Segata N, Börnigen D, Morgan XC, Huttenhower C.** PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* 2013;4:2304.
23. **Hahnke RL, Meier-Kolthoff JP, García-López M, Mukherjee S, Huntemann M *et al.*** Genome-based taxonomic classification of bacteroidetes. *Front Microbiol* 2016;7:2003.
24. **Goodwin S, Mcpherson JD, Mccombie WR.** Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;17:333–351.
25. **Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M *et al.*** Application of next generation sequencing in clinical microbiology and infection prevention. *J Biotechnol* 2017;243:16–24.
26. **Ewing B, Hillier L, Wendl MC, Green P.** Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 1998;8:175–185.
27. **Liao YC, Lin SH, Lin HH.** Completing bacterial genome assemblies: strategy and performance comparisons. *Sci Rep* 2015;5:8747.
28. **Lee I, Chalita M, Ha SM, Na SI, Yoon SH *et al.*** ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences. *Int J Syst Evol Microbiol* 2017;67:2053–2057.
29. **Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW.** CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–1055.
30. **Richter M, Rosselló-Móra R, Oliver Glöckner F, Peplies J.** JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* 2016;32:929–931.
31. **Jolley KA, Maiden MC.** BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010;11:595.
32. **Ankenbrand MJ, Keller A.** bcgTree: automatized phylogenetic tree building from bacterial core genomes. *Genome* 2016;59:783–791.