The Taxonomy of Bacteria in the Genomic Era

Lorena Carro, Álvaro Peix, and Encarna Velázquez

1 A Brief Introduction to the History of Bacterial Taxonomy

If we open a student book of microbiology, we will find the definition of the taxonomy as "the science by which organisms are characterized, named, and placed into groups according to several defined criteria" and together with the phylogeny conforms the systematics, "the study of the diversity of organisms and their relationships" (Madigan et al. 2012). However, what is behind these definitions is something much more complex, with hundreds of laboratory techniques applied for a proper characterization of microorganisms that have been evolving over years. Taxonomy has been applied from the very beginning of human conscious of its environment, classifying the organisms of their surrounding; however, we will focus on bacterial systematics, whose origin goes back to the "animalcules" descriptions generated by Antonie Van Leeuwenhoek in the middle of the seventeenth century. Few improvements were obtained until the nineteenth century, where the first genera of bacteria were described, with the term *Bacterium* given for the first time to classify rod-shaped cells (Murray and Holt 2005).

The species concept, so easily defined for higher organisms, has been a source of discussion from the very beginning of prokaryotic taxonomy and still generates controversy among researchers. First bacterial species definitions include terms such

L. Carro (✉)
Department of Conservation of Natural Resources, Soil Microbial Ecology Group, NEIKER, Derio, Spain
e-mail: lcg@usal.es

Á. Peix
Natural Resources and Agrobiology Institute, IRNASA-CSIC, Salamanca, Spain

E. Velázquez
Microbiology and Genetics Department, University of Salamanca, Salamanca, Spain

T. G. Villa, T. de Miguel Bouzas (eds.), *Developmental Biology in Prokaryotes and Lower Eukaryotes*, https://doi.org/10.1007/978-3-030-77595-7_12

as close resemblance and essential and distinguishing features, which induced to differentiate species according to their morphology, source of isolation, and pathogenicity. These features, although useful at the beginning, were lately shown to be highly imprecise and subjective (Brenner et al. 2001).

First principles for bacterial characterization, classification, and identification were proposed in the second half of the twentieth century, when a sole character despite its importance was proposed to be not enough for species definition. Instead, they proposed the use of a large list of biochemical tests and strain samples, to better characterize members of new defined species. This classification ended in the proposal of what was called numerical taxonomy, proposed by Sokal and Sneath (1964). Within this method, a whole range of tests, more than 100, were analyzed, and coefficients were established to calculate similarity between strains and species.

However, it was not until the taxonomists were able to extract nucleic acids from cells that a more "natural classification" based on nucleic acid analysis was possible to better define bacterial species. In the 1960s, the development of methodologies for nucleic acid extraction and renaturation (Marmur 1961; Marmur and Doty 1961) led to the first studies based on DNA homologies and genetic comparisons through DNA-DNA and DNA-RNA hybridization to be used in bacterial classification. A few years later, in the 1980s, the profiles of stable low molecular weight RNA (LMW RNA) were proposed for bacterial species differentiation (Höfle 1988), which separated by a modified electrophoresis technique named Staircase Electrophoresis (SE) increased the resolution of LMW RNA profiles allowing the differentiation of genera and species of both prokaryotic and eukaryotic microorganisms (Velázquez et al. 2001).

Also, the discovery in the 1980s of the polymerase chain reaction (PCR) by Mullis in 1983 opens the development of several fingerprinting techniques based on DNA amplification of given fragments or genes. Restriction Fragment Length Polymorphisms (RFLP) technology was described in 1980 (Botstein et al. 1980) and then applied to generate DNA profiling, which allowed us to find differences among strains. Other techniques for genotyping characterization of microorganisms were developed, such as ARDRA (Amplified rDNA Restriction Analysis), RAPD (Random Amplification of Polymorphic DNA), BOX-PCR (Repetitive extragenic palindromic sequences), and ERIC-PCR (Enterobacterial Repetitive Intergenic Consensus), allowing a fast classification into genetic groups of the microorganisms isolated in a sample (Carro and Nouioui 2017).

Apart from these applications, the development of the polymerase chain reaction allowed the use of a gold marker in prokaryotic taxonomy, the 16S rRNA gene. This gene was selected by a group of characteristics that made this molecule a unique taxonomic marker: its size, its slow evolution rate, and its ubiquity in bacteria. These characteristics allow us to propose a phylogenetic classification of Prokaryotes based on this gene (Winker and Woese 1991). In 1994, its use was already widespread among taxonomists to generate phylogenetic reconstructions of new taxa. In spite of its limitations to define by itself further than genus classification, a similarity cutoff value was proposed for species delineation, the 97% similarity in 16S rRNA gene sequence (Stackebrandt and Goebel 1994). This value was later updated to

98.7–99% similarity, depending on the genus (Stackebrandt and Ebers 2006), and validated with empirical datasets and statistical probabilities of failure (Meier-Kolthoff et al. 2013b). Strains presenting values over this range should be analyzed by DNA-DNA hybridization methods to define novel species.

The 16S rRNA gene belongs to the ribosomal operon, which also encompasses in bacteria the 23S and 5S rRNA genes as well as several intergenic regions. In addition to the 16S rRNA gene, the 23S rRNA gene and the intergenic spacer (ITS) between the 16S rRNA and 23S rRNA genes have been used with taxonomic purposes (Ludwig and Schleifer 1994; D'Auria et al. 2006; Yarza et al. 2010). The ITS region contains hypervariable sequence regions allowing the differentiation of bacterial genera, species, and strains (Peix et al. 2005). Also, this region has different sizes in different bacterial groups, which facilitates their use for metagenomic analysis of bacterial populations through a technique named RISA or ARISA (Ranjard et al. 2000). Currently, the complete ribosomal operons are used for the identification of bacteria in complex samples by metagenomics (Kerkhof et al. 2017; Cuscó et al. 2019; Martijn et al. 2019).

The 16S rRNA gene contains highly conserved regions in bacteria, which allowed its amplification and sequencing with universal primers annealing in these regions (Edwards et al. 1989). Two of these primers were used to obtain TP-RAPD (Two Primers Random Amplified Polymorphic DNA) patterns, which are not strain-dependent being able to differentiate among different bacterial species (Rivas et al. 2001). Universal primers can also be used to amplify a region of the 16S rRNA gene in Prokaryotes and 16S rRNA genes in Eukaryotes called UARR (Universal Amplified Ribosomal Region), which contains the V6, V7, and V8 domains (Rivas et al. 2004). These regions, particularly V6, are useful for metagenomic analysis of bacterial populations through NGS (Kumar et al. 2011; Temperton and Giovannoni 2012; Tremblay et al. 2015; Yang et al. 2016; Winand et al. 2019).

DNA-DNA hybridization allows the comparison of two complete genomes and therefore the calculation of a similarity percentage between them on the basis of the dissociation of the DNA strands. The fast development of this technique allowed the determination of a numerical value as a threshold to define whether two microorganisms belong to the same species, set in 70% by Wayne in 1987 (Wayne et al. 1987). Several methods were developed to improve this technique avoiding radioactive labeling as used to be performed since its proposal at the early sixties, including filter competition, optical renaturation rates, hydroxyapatite, fluorimetric, or microplates (Mehlen et al. 2004). However, the availability limited to highly specialized laboratories and low reproducibility of the results obtained between laboratories is always claimed for more stable methods (Carro et al. 2012).

Among those methods, a first proposal was the use of multilocus sequence analysis (MLSA), the analysis of a small set of protein-encoding genes, also known as housekeeping genes (Chimetto Tonon and Moreira 2016). As sequencing technologies became affordable for most laboratories, the use of this technique increased for classification of novel taxa. However, the first works on this technique showed that the selection of genes is relevant for the results obtained and what is worst, variable among genera, which led to the search for valid genes to be used for

each genus to be analyzed (Carro et al. 2012). Nevertheless, once those genes were found for a given genus, the phylogeny of the concatenated sequence of them was robust and allows a better definition of the genus diversity and evolution (Adékambi and Drancourt 2004; Guo et al. 2008). More recently, the availability of whole-genome sequences and the generation of several overall genome relatedness index (OGRI), which will be discussed in the next section, have generated the searched stability for DNA-DNA hybridization, since the comparison of genome sequences offers the same values regardless of the performing laboratory and tools have been developed to allow the general use of this approach (Chun et al. 2018).

The compilation of all these analyses was called polyphasic taxonomy, which is the combination of phenotypic, environmental, and genotypic characteristics, together with the phylogeny of the strains, to generate a whole view of the microorganisms to be described for properly proceeding to their classification and identification. This polyphasic taxonomy is always in evolution including new developed tools for bacterial characterization.

2 Whole-Genome Sequences: How to Use Them in Bacterial Taxonomy

During the last decade, the number of genomes available exponentially increased (Fig. 1). This explosion in numbers is due to several reasons, including the improvement of next-generation sequencing (NGS) technologies, which allowed us to generate a whole draft genome in less than 48 h with a drastic reduction in the

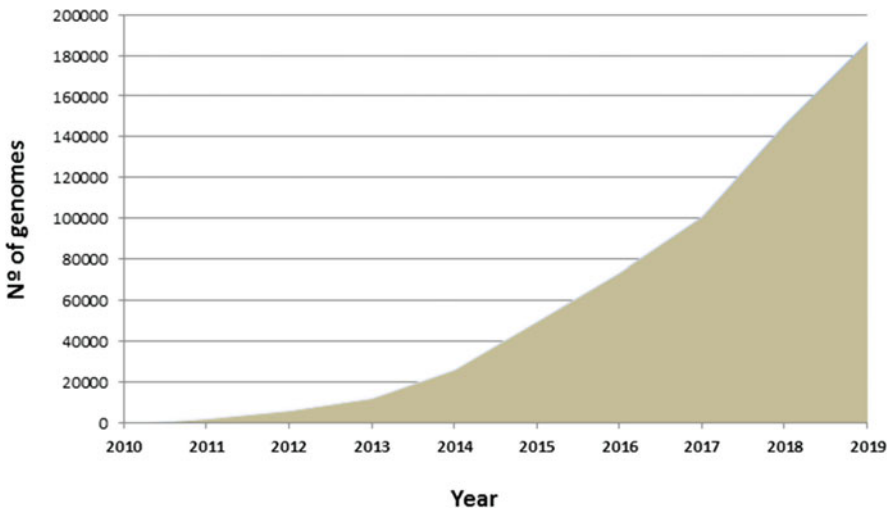

Fig. 1 Number of prokaryotic genomes publicly available generated during the last decade. Source of information: EZBioCloud Statistics

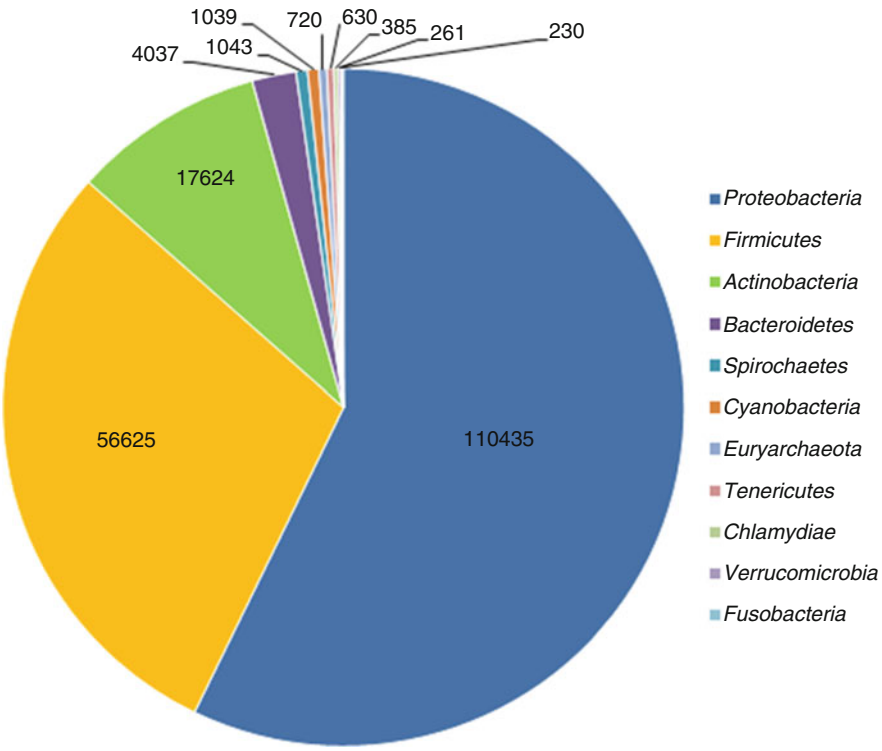

Fig. 2 Number of prokaryotic genomes by phylum. Source of information: EZBioCloud Statistics

cost (Kremer et al. 2017), and multigenome sequencing projects such as GEBA (The genomic encyclopedia of *Bacteria* and *Archaea*), which aimed to filling the gaps of type-strain genome in the tree of life (Mukherjee et al. 2017). Half of the genomes generated until now correspond to strains of the phylum *Proteobacteria*, with the most abundant phylum being *Firmicutes* and *Actinobacteria* in second and third position, respectively, and only 5% of the genomes belonging to other phyla (Fig. 2). Since early 2019, the genome sequence is considered a key feature to present for the type strain of every species proposed, according to the instructions for authors of the International Journal of Systematic and Evolutionary Microbiology Journal. This prerequisite, included for other microbial description journals, has also helped to increase the number of available genomes. However, in spite of the potentialities hidden in the genomic information of those strains, most papers only meet the requirement of genome draft production, without trying to also unravel the useful information related to the proposed taxa. But how can we generate good sequencing data? Which of them should be used in taxonomic descriptions? How should they be presented? Answers to those questions are in the following subsections.

2.1 Technologies to Generate Whole-Genome Sequences

Major breakthrough in DNA sequencing arrived with Sanger's chain-termination technique in 1977, which was rapidly extended and widely used for the following three decades (Heather and Chain 2016). The automation of this technology gave which are now called the first-generation DNA sequencing machines, routinely applied in many laboratories to date. Using this technology, first genomes of prokaryotic organisms were produced, with *Haemophilus influenzae* completed in 1995 (Fleischmann et al. 1995) the first. The development of this technology allowed the simultaneous sequencing of hundreds of samples and was even applied for the generation of the first decoded human genome (Lander et al. 2001).

Improvements of the methodologies used for DNA sequencing have not stopped, and the following group of technologies has been called next-generation sequencing (NGS) or second generation of DNA sequencing. These technologies are based on DNA fixed to a solid phase and the measuring of pyrophosphate, which is determined when is converted into ATP by ATP sulfurylase, and the ATP is used as substrate for a measurable luciferase. These pyrosequencing technologies were firstly developed by Roche, allowing the mass parallelization of sequencing reactions (Margulies et al. 2005). Other companies developed their own systems, including Illumina, IonTorrent, or Life technologies, with Illumina being, with their HiSeq and MiSeq technologies, one of the most frequently used nowadays. Each of these sequencing platforms present advantages and disadvantages, as shown in Table 1. Some of these platforms, such as Illumina and SOLiD, are also known as "short-read" technologies, as the data generated by them are short reads between 30 and 500 bp, largely smaller than that obtained by Sanger sequencing (around 1000 bp) (Kremer et al. 2017). Nevertheless, NGS allowed a drastic reduction in the cost per base, mainly due to the higher sequencing coverage obtained by the higher throughput, generating millions or billions of DNA strands sequenced in parallel.

The third generation of DNA sequencing arrives with the technologies that avoid the direct action of the DNA polymerase, allowing long reads from very limited DNA samples. Among the technologies grouped under this name are the single molecule sequencing and real-time sequencing, with no DNA amplification. The single molecule real-time (SMRT) platform from Pacific Biosciences is the most used nowadays from this generation of DNA sequencers. PacBio platforms are able to produce really long reads, over 10 kb in length, which are especially useful for the generation of de novo genome sequences (Heather and Chain 2016). However, main flaw of this technology is the high error rate (Table 1), which has hampered the generalization of its use, as well as a higher price than other available systems. The second technology pushing up in this third generation is developed by Oxford Nanopore Technologies. What is expected from this technology is the generation of very long reads at really low cost; in addition, they have presented compact machines as MinION, having the size of a smartphone, allowing direct sequencing of material at sampling sites (Loman and Quinlan 2014). Poor-quality profiles obtained and high error rate are still the challenges for these technologies.

Table 1 A comparison of main characteristics of most used next-generation sequencing instruments

	Roche 454 FLX titanium	Life technologies solid	Ion torrent PGM	Ion proton	Illumina MiSeq	Illumina HiSeq2500	PacBio RS	Oxford nanopore
Detection	Light	Ligation	pH	Proton	Fluorescence	Fluorescence	Fluorescence	Nanopore
Time for run	9 h	8 day	2 h	2 h	27 h	11 day/27 h	2 h	6 h
Read length (bp)	400	85	100	200	2 × 250	2 × 100	10,000	2,000,000
Read per run	1×10^6	1000×10^6	1×10^6	660×10^6	1×10^6	3000×10^6	1×10^6	1×10^6
Gb per run	0.7	150	0.1	100	8	300	9	42
DNA requirements	3000–20,000 ng	0.2 ng	100–1000 ng	10–100 ng	50–1000 ng	50–1000 ng	250–15,000 ng	0.01–1000 ng
Observed raw error rate	0.6%	0.06%	1.7%	1%	0.8%	0.3%	12.9%	6%
Insert size	800	120	250	200	700	700	15,000	50 Gb
Advantages	Long reads	Low error rate	Short run time	Short run time	High throughput/cost	High throughput/cost	No PCR/Longest read	Long reads/portable
Disadvantages	High error rate and no platform support	Short reads and long run time	High hands-on time and low reads per cost	High hands-on time and low reads per cost	Short reads and few reads per cost than HiSeq	High instrument cost and short reads	High error rate and high capital cost	High error rate and high cost per read

The platforms provided by Illumina, Ion Torrent, and PacBio are considered to meet the general standards for the description of new species (Chun et al. 2018). The combination of the most used platform of each generation is also highly encouraged, and some of the most important sequencing centers, as the Joint Genome Institute, apply this methodology. PacBio platforms allow obtaining long sequences, key to be able to close de novo genomes, while Illumina platforms allow the validation of the sequence, due to its low error rate. This combination offer best results in bacterial genome sequencing, but its application is often limited due to the cost involved in this double sequencing.

2.2 Minimal Standards for Genomic Data in Taxonomy

Once the reads of the genomic sequencing have been obtained, it is necessary to determine the quality of the results obtained as well as the assembling of them into contigs and scaffolds, for which a lot of specific software has been developed (Velvet, SPAdes, QUAST, etc.). NGS platforms provide their own statistics for sequencing raw data; however, the most important statistics to take into account for taxonomic purposes are the ones obtained from the final assembly. A good review of the available software tools was presented by Kremer and his colleagues (Kremer et al. 2017). Some of the key parameters according to Chun et al. (2018) that should be checked in a genome assembly and included in the genome description of the strain include

- *Number of contigs*: the perfect number of contigs is one, and accordingly, this number should be as low as possible with the obtained data. However, for taxonomic purposes, this value could be higher if the redundancy or coverage of the data is enough, with values sometimes as high as 600 accepted.
- *N50*: this is a good parameter to measure the quality of an assembly. N50 represents the minimum contig length needed to cover 50% or more of the genome when the contigs are summed from the largest to shortest.
- *Coverage*: another important value is the sequencing depth of coverage and indicates how many times each base of the final assembly has been read on average, indicated as the folds. A recommended value proposed as minimum for a good coverage is $\geq$50X.
- *Genome size*: most of the genomes generated until now are not closed, and this implicates that the genome size, taken as the sum of all the contigs length, is just an approximation, but this value gives an idea of completeness compared with other members of the genus and by in vitro calculation.
- *G+C content*: this value is also an indicator of the quality, and it should be coherent with the expected data for the strain of study.
- *16S rRNA sequence*: in addition to obtaining this marker gene as an indicator of completeness of the genome sequence, it should be used to verify the authenticity of the genome, in order to verify that it matches with the Sanger sequence

obtained from the strain for which the genome has been generated. This will avoid possible mistakes regarding strain contamination or labeling. Other housekeeping genes could also be used in the case of doubt.

- *Contamination of samples*: it is possible that contamination of the DNA samples to be sequenced occurs, and even if these are in a minor amount, they could be incorporated into the final genome sequence. One tool that has been created for this purpose is CheckM (Parks et al. 2015), which is also used to study other quality parameters, it indicates the percentage of possible contamination in a genome. However, lateral gene transfer events should be carefully interpreted, as it is a common event in prokaryotes. Another tool, ContEst16S, focused on the presence of different 16S rRNA genes in the assembly has also been developed to find possible contaminations (Lee et al. 2017).

On the other hand, the information on how the genomes have been generated and make this information available is also essential when genomic data are used in taxonomic descriptions of species. Some of the main points to be taken into account include

- *Deposit genome information in public databases*: two main databases should be used for the deposit of information at GenBank/EMBL/DDBJ database:
  - *WGS database*: assembled and quality checked genome should be deposited to allow comparison between your genome and others.
  - *SRA database*: raw sequencing data should be deposited too, as it could be used for improving the assembly once more information or better methods will be developed in the future.
- *Properly describe the sequencing, assembling, and annotation methods*: including the sequencing instrument, the reagents used for library preparation, and all the software used in the process to obtain the final genome.

2.3 Overall Genome Relatedness Index (OGRI)

The OGRIs is a generic name to group all the bioinformatics methods defined to replace the wet-lab DNA-DNA hybridization (DDH) for the differentiation of species (Chun and Rainey 2014) in a reproducible and objective way. These methods utilize whole-genome sequences, and gene annotation is not previously required. Calculation of OGRI or DDH values with all the closely related species presenting a similarity value of 16S rRNA gene over 98.7% is compulsory when a new species is proposed, at least with one of the known methods.

Even before the genome sequencing was affordable for most microbiology laboratories, the utilization of in silico methods to replace DDH was proposed (Henz et al. 2004; Konstantinidis and Tiedje 2005). One of the methods used to correlate DDH values with digital DDH by computational comparison of genome sequences was the *average nucleotide identity* (ANI), representing a mean of identity

values between multiple sets of orthologous regions (Konstantinidis and Tiedje 2005). An initial cutoff value of 94% was proposed to correspond to the traditional 70% for DDH, but this boundary was later adjusted to 95–96% after refining the method and simulating an artificially cutting of the genome similar to what was occurring in the DDH method (Goris et al. 2007). The implementation of MUMmer software (*ANIm*) instead of BLASTN (*ANIb*) helped to obtain faster results for ANI method (Richter and Rosselló-Móra 2009). Another implementation to solve the differences in reciprocal ANI values was proposed by generating a new algorithm, the *OrthoANI*, which uses only orthologous fragment pairs to calculate nucleotide identities (Lee et al. 2016).

Other OGRI proposed in parallel and widely distributed is a distance genome relatedness index, the *genome BLAST distance phylogeny* (GBDP) (Henz et al. 2004), for which calculation two genome sequences are aligned to each other and generate high-scoring segment pairs to apply a specific distance formula. The algorithm was lately improved with confidence-interval estimation thanks to a new statistical model proposed by Meier-Kolthoff and colleagues (Meier-Kolthoff et al. 2013a). These implementations generate the *digital DDH* (*dDDH*), which mimics the results of the classical DDH with confidence-internal estimation, enabling the user to statistically evaluate the outcomes. Therefore, the species boundary for dDDH values is 70%, the same as the one proposed for classical DDH (Wayne et al. 1987). A web-based tool was implemented to carry on those analyses known as *genome to genome distance calculator* (GGDC), which is available online.

Another distance-type index is maximal unique matches index (*MUMi*) (Deloger et al. 2009); however, its use has been much more limited, probably because it was proposed to provide higher resolution at the intraspecies level and analyzes the exact matches shared by the two sequences of study.

2.4 Genome Phylogeny

As the genomes of more type strains become available, the generation of phylogenomic trees using whole-genome sequences directly, or groups of genes obtained from them, should be compulsory to better determine the phylogenetic relationship of the strains of study with other species of the genus. This is even more important in genera for which the 16S rRNA gene has been shown not enough for species differentiation. In addition, phylogenomic approach is able to provide a better taxonomic framework for genus and higher taxa. This has been shown in many works published recently; those have allowed the analysis of whole phyla and proposed several reclassifications within its orders, families, genera, and species. Within these analyses, we can find several works on whole phyla, as *Bacteroidetes* (García-López et al. 2019; Hahnke et al. 2016) and *Actinobacteria* (Nouioui et al. 2018; Salam et al. 2020), as well as families, genera, and species, like *Rhodobacteraceae* (Simon et al. 2017), *Micromonospora* (Carro et al. 2018), and *Pseudomonas fluorescens* complex (Garrido-Sanz et al. 2016).

Within the available methods for generate whole-genome phylogenies, two main approaches have been proposed: a) the use of core or conserved genes between the genomes, which can vary from below one hundred genes to several thousand depending on the method applied, and b) the used of whole-genome sequences, based on amino acid or nucleotides. For the use of conserved genes, several approaches have been proposed, between them, the up-to-date bacterial core gene (UBCG) method proposed by Na et al. (2018) and freely available, which is increasingly used. This method has defined a set of 92 core genes to be concatenated that are conserved for all taxonomic ranks of *Bacteria*, allowing standard comparison regardless of the number of strains included in the analysis. Other methods currently applied include CSI phylogeny (Kaas et al. 2014), a webserver that identifies the single nucleotide polymorphisms (SNPs) and infers a phylogeny based on the concatenated alignment of them, or M1CR0B1AL1Z3R (Avram et al. 2019), a web server that finds orthologous groups, aligns them, and generates the corresponding phylogeny. For the use of whole-genome sequences, several approaches have also been proposed, REALPHY was developed by Bertels et al. (2014), and the pipeline is freely available online. In this method, sequences are mapped against reference genomes by bowtie 2 and the phylogenies inferred with PhyML. More recently, another website has been developed to generate whole-genome phylogenies, the Type (Strain) Genome Server (TYGS) (Meier-Kolthoff and Göker 2019), which is increasingly used probably due to its user-friendly interface, although it is limited to comparison of up to 20 genomes at present. The TYGS methodology is based on the Genome BLAST Distance Phylogeny method (GBDP) (Meier-Kolthoff et al. 2013a). The number and methodologies are continuously increasing and improving according to the technologies available, but online servers help to share these analyses, which are usually highly demanding on computer resources, with the research community, making them available to anyone with a computer and internet connection.

2.5 *Genome Characterization: Where Should One Begin?*

A huge amount of taxonomic papers that have included the genome of the corresponding type strain has limited its use to the dDDH or ANI calculations, lacking even the generation of the genome phylogeny. This fact is a pity, taking into account all the information that could be extracted from a genome sequence. So, we will list here some of the analyses that can be performd to improve the use of genomes in taxonomic manuscripts:

- *OGRI calculations*: as shown before, several tools are available to calculate the relatedness of the studied strains with close relatives in the genus or family.
- *Construct the genome phylogeny*: although not all the genomes of the species of interest are available, it should be equally constructed, as the 16S rRNA gene is not enough to define closely related strains for many genera, and a better idea of

the position of the strain will be given. The ideal would also be to obtain the genomes of the closely related type strains and, in that way, completing the availability of the genomic information of the genus. In the genera with a group of well-defined housekeeping genes for taxonomy, these could be directly obtained from the genome too to generate an MLSA phylogenetic tree, including all the type strains for which these data are available.

- *Core and pangenome*: the core genome refers to the genes that are shared by all known members of a taxonomic group without exception, while the pangenome refers to all the genes contained in all the strains belonging to the same taxonomic group. The pangenome includes the core genome and the accessory genome, not necessary for the survival of the species, and could be really high, with hundreds of strains probably necessary to complete one (Medini et al. 2005). Both concepts, core and pangenome, are important from an evolutionary point of view and should be analyzed when a relevant number of strains are known for the same species.
- *Bioclusters* determination: some tools have been developed to determine the potential capacity of a strain to produce antibiotics or other secondary metabolites from genome information. Between them, antiSMASH has been gaining attention and is increasingly used to give a general idea of the potential activity that a new isolate can have. First version of this website available tool was proposed in 2011 (Medema et al. 2011), which is nowadays in version 5.0 (Blin et al. 2019). Another tool to look for biosynthetic gene clusters is ABC (Atlas of Biosynthetic gene Clusters), developed by the Joint Genome Institute and available at the IMG (Integrated Microbial Genomes) database (Hadjithomas et al. 2015). This tool has been developed based on predictions for all available genomes in IMG, and a last updated version has been recently released (Palaniappan et al. 2019).
- *Ecologic and phenotypic analysis*: the genome sequences are full of information to better characterize the new taxa. However, it is sometimes difficult to decide what to look for or how to find it. Several approaches could be used, for example, we can decide first a series of characteristics that we are interested in and look for the genes already described that have those functions and then search for homologous genes within the genome. Another approach is to check generally the annotated genes of our genomes and decide which of them should be further characterized or analyzed. Within all the information that could be analyzed in a genome, two main characteristics should be proposed in the description of new species, the genes related to the ecological role of the studied microorganism (adaptation to the environment conditions, interaction with closely related organisms, etc.) and the genes related to the phenotypic abilities. Frequently, when the ability of a microorganism to use a carbon source or produce a specific compound is tested, differential results are observed among laboratories and even within the same laboratory (Riesco et al. 2018). On the other hand, the analysis of the genome allows determining the presence or absence of specific genes encoding for the production of specific compounds by a strain, although sometimes the laboratory conditions do not allow observing a positive result due to specific conditions needed or the nonfunctionality of the genes. Several tools and

approaches could be used to determine or study those characteristics, such as SEED viewer, an intuitive and friendly user platform (Overbeek et al. 2014), after using the RAST server for genome annotation (Aziz et al. 2008).

3 Metagenomic Analysis: Do They Fit in Classical Taxonomy?

Last sequencing technologies developed have led to a huge step forward in microbial ecology studies, allowing the in situ characterization and identification of millions of bacteria that were never detected by classical isolation methods. Nevertheless, metagenomic results depend enormously on the taxonomy, as we will be able to identify the organisms that are properly described in the literature, while the other ones will generate an increase "microbial dark matter." On the other hand, the classification of those organisms is in conflict with classical taxonomy, which needs culturable microorganisms to apply polyphasic taxonomy and deposit of strains in two independent culture collections to fulfill the postulates of the International Code of Nomenclature of Prokaryotes (ICNP) for name validation.

Whitman proposed in 2015 the use of genome sequences as the type material for taxonomic descriptions of prokaryotes (Whitman 2015), an article that have generated a debate among the taxonomists of the twenty-first century. As previously exposed, whole-genome sequences have induced a huge evolution in microorganisms classification and characterization in very few years, a fact that none has discussed. However, the absence of a whole organism that could be maintained and all its information reproduced generate uncertainties for the definition of new species, which concern a good number of researchers. At that time, Whitman proposed the possibility to deposit DNA in public collections that should be based on either a clonal population or a single cell. The change of the code to allow gene sequences as type material for the description of prokaryotic species was proposed by Whitman the following year in a taxonomic note (Whitman 2016). In this proposal, he remarks the importance of naming the prokaryotic diversity to allow the communication among researchers from different fields without misidentifications and justified the change in the code to allow the validation of the names from *Candidatus* taxa, the way in which are known uncultivated microorganisms that could not be validated since their existence is only known based on genome or genes amplification. Many taxonomic groups have been described as *Candidatus* in the last few years thanks to metagenomic analyses; however, the names proposed have not priority according to the Bacteriological Code of Nomenclature, and therefore, if a strain of these taxa is isolated and described could be given a completely different name. This situation generate two problems, the absence of a list of *Candidatus* names with the corresponding sequences where it can be checked if they were already described and the generation of different names to classify the same organisms. Nevertheless, the solution of this problem could be easily solved

without introducing the validation of genomes or genes as type material (with the risk that this has as we will discuss later), only by giving priority to *Candidatus* names, a proposal that has already been made to the International Committee on Systematics of Prokaryotes (ICSP). In this way, the uncultivable strains will be named as *Candidatus* before the proposed name until a strain of these taxa would be cultivable, moment at which the proposed species will be updated with the same name to the standard taxonomy.

A different proposal was made by Konstatinidis and colleagues in 2017 (Konstantinidis et al. 2017), who suggested to generate an independent nomenclature for not-yet-cultivated taxa, proposing a series of standards and guidelines for the description of these taxa, based on genome sequences obtained from single-cell amplification or population binning that will be used as type material. They also proposed to have their own list of validly published names. Although they proposed some minimal standards of quality, they recognized that "metagenomes typically constitute a mosaic of different genotypes of a single population coexisting in the same environment," which generate some uncertainties about the species that would be described.

Several concerns to the inclusion of genes or genome sequences as type materials have been claimed by other taxonomists, including the quality control, the limitation of the original source to reproduce the results (indeed, some of the already described species with valid names are no longer considered as types because the culture was lost at some point), the complexity of the maintenance of DNA, and the difficulties for the culture collections to distribute the DNA to other researchers in order to repeat the experiments. Moreover, a genome sequence could be artificially generated, and even if we would believe on the honorability of researchers, chimeras out of their control could be described as real diversity. Some of the arguments against this proposal are summarized in the article of Bisgaard et al. (2019), including

- DNA material could be damaged or lost, and as it is not a proliferating material, data will not be reproducible.
- Species descriptions will need to be revised as its DNA sequence is replaced by new version within the development of sequencing and assembling technologies.
- Functional assessments for genomes are limited.
- Genomic data do not always agree with gene expression, which could generate errors to establish taxonomic relationships.
- Minimal standards for new taxa descriptions will be difficult to define and could induce to produce a high amount of new taxa descriptions based on single DNA sequences, generating a taxonomic and nomenclatural chaos.
- Motivation for the isolation and phenotypic characterization of strains will decrease, and therefore, the study of intra and interspecies diversity will be reduced.

Similar arguments were also presented by Zamora et al. (2018) to evidence the problems of allowing DNA sequence data as type material in fungal taxa, trying to put on evidence the consequences of accepting the proposals to amend the International Code of Nomenclature for algae, fungi, and plants (ICN). They argue that

using DNA as a type, it will be used just as information from a character of an organism, instead of the organism itself, severely limiting the characterization, and names should be given to organisms, not to characters of them. In addition, a major concern raised is the reliability of the DNA sequence data and a proper method to be checked, generating irreproducible science.

The proposal of a different nomenclature for not-yet-cultivated taxa was also discussed by some authors: Oren and Garrity expressed their concerns showing several examples of how the unregulated naming of taxa has previously led to chaos (Oren and Garrity 2018); while Overmann et al. argued that, in addition to the arguments previously exposed for genomes as type material (technological and conceptual limitations), confusion will be unavoidable if two different nomenclatures are created without links that avoid the generation of synonyms (Overmann et al. 2019). The limitation of phenotypic information that will be given for not-yet-cultured microorganisms is another point of concern. These authors are positive to give priority to *Candidatus* names when those microorganisms are described with sufficient morphological-cytological, metabolic, and ecological traits to clearly distinguish them from other taxa, to be informative, and induce new more successful cultivation attempts based on the generated information (the so-called "reverse metagenomics").

According to Konstantinidis et al. (2020), the advantages outweigh concerns in the use of genome sequences as type material, as they try to demonstrate in their last correspondence letter to Environmental Microbiology, where they argue that the use of genomes will not generate weakness in the standards of prokaryotic taxonomy, trying to answer the authors who have expressed concerns like "who will take the time to grow and deposit their strains if a genome sequence is valid." In addition, they discuss that only high-quality genomes should be used as type material, avoiding the future revision of the sequences discussed by Bisgaard et al. (2019).

However, it seems that for the moment, many taxonomists are not convinced and the use of type material, which is expected to continuously be available for researchers, will still be a culturable type strain. Nonetheless, this fact does not prevent the proposal of change of priority in the Code of Nomenclature for Prokaryotes, and noncultivable bacteria identified by metagenomics could be proposed as new species and named *Candidatus*, which name should be maintained once a strain of the *Candidatus* taxa could be finally isolated and cultivated. This simple solution will allow generating a whole new taxonomy based on "uncultivable" strains, which names could be valid and just waiting to develop the capacity to grow them under laboratory conditions.

References

Adékambi T, Drancourt M (2004) Dissection of phylogenetic relationships among 19 rapidly growing *Mycobacterium* species by 16S rRNA, hsp65, sodA, recA and rpoB gene sequencing. Int J Syst Evol Microbiol 54:2095–2105

Avram O, Rapoport D, Portugez S, Pupko T (2019) M1CR0B1AL1Z3R—a user-friendly web server for the analysis of large-scale microbial genomics data. Nucleic Acids Res 47:W88–W92

Aziz RK, Bartels D, Best A, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:1–15

Bertels F, Silander OK, Pachkov M, Rainey PB, van Nimwegen E (2014) Automated reconstruction of whole-genome phylogenies from short-sequence reads. Mol Biol Evol 31:1077–1088

Bisgaard M, Christensen H, Clermont D, Dijkshoorn L, Janda JM, Moore ERB, Nemec A, Nørskov-Lauritsen N, Overmann J, Reubsaet FAG (2019) The use of genomic DNA sequences as type material for valid publication of bacterial species names will have severe implications for clinical microbiology and related disciplines. Diagn Microbiol Infect Dis 95:102–103

Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 47:W81–W87

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Brenner DJ, Staley JT, Krieg NR (2001) Classification of procaryotic organisms and the concept of bacterial speciation. In: Garrity GM, Boone D, Castenholz R (eds) Bergey's manual of systematic bacteriology. Springer, Cham, pp 27–31

Carro L, Nouioui I (2017) Taxonomy and systematics of plant probiotic bacteria in the genomic era. AIMS Microbiol 3:383–412

Carro L, Spröer C, Alonso P, Trujillo ME (2012) Diversity of Micromonospora strains isolated from nitrogen fixing nodules and rhizosphere of *Pisum sativum* analyzed by multilocus sequence analysis. Syst Appl Microbiol 35:73–80

Carro L, Nouioui I, Sangal V, Meier-Kolthoff JP, Trujillo ME, Montero-Calasanz MDC, Sahin N, Smith DL, Kim KE, Peluso P, Deshpande S, Woyke T, Shapiro N, Kyrpides NC, Klenk H-P, Göker M, Goodfellow M (2018) Genome-based classification of micromonosporae with a focus on their biotechnological and ecological potential. Sci Rep 8:525

Chimetto Tonon L, Moreira APB (2016) Molecular taxonomy of environmental prokaryotes. In: Rodrigues TB, Silva AT (eds) Molecular diversity of environmental prokaryotes. CRC Press, Boca Raton, pp 3–36

Chun J, Rainey FA (2014) Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int J Syst Evol Microbiol 64:316–324

Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, Rooney AP, Yi H, Xu XW, De Meyer S, Trujillo ME (2018) Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int J Syst Evol Microbiol 68:461–466

Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O (2019) Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the rrn operon. F1000Res 7:1755

D'Auria G, Pushker R, Rodriguez-Valera F (2006) IWoCS: analyzing ribosomal intergenic transcribed spacers configuration and taxonomic relationships. Bioinformatics 22:527–531

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

Edwards U, Rogall T, Blöcker H, Emde M, Böttger EC (1989) Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. Nucleic Acids Res 17:7843–7853

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 80(269):496–512

García-López M, Meier-Kolthoff JP, Tindall BJ, Gronow S, Woyke T, Kyrpides NC, Hahnke RL, Göker M (2019) Analysis of 1,000 type-strain genomes improves taxonomic classification of bacteroidetes. Front Microbiol 10:2083

Garrido-Sanz D, Meier-Kolthoff JP, Göker M, Martín M, Rivilla R, Redondo-Nieto M (2016) Genomic and genetic diversity within the *Pseudomonas fluorescens* complex. PLoS One 11: e0150183–e0150183

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91

Guo YP, Zheng W, Rong XY, 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

Hadjithomas M, Chen I-MA, Chu K, Ratner A, Palaniappan K, Szeto E, Huang J, Reddy TBK, Cimermančič P, Fischbach MA, Ivanova NN, Markowitz VM, Kyrpides NC, Pati A (2015) IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. MBio 6:e00932–e00915

Hahnke RL, Meier-Kolthoff JP, García-López M, Mukherjee S, Huntemann M, Ivanova NN, Woyke T, Kyrpides NC, Klenk H-P, Göker M (2016) Genome-based taxonomic classification of bacteroidetes. Front Microbiol 7:2003

Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107:1–8

Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC (2004) Whole-genome prokaryotic phylogeny. Bioinformatics 21:2329–2335

Höfle MG (1988) Identification of bacteria by low molecular weight RNA profiles: a new chemotaxonomic approach. J Microbiol Methods 8:235–248

Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O (2014) Solving the problem of comparing whole bacterial genomes across different sequencing platforms. PLoS One 9:e104984–e104984

Kerkhof LJ, Dillon KP, Häggblom MM, McGuinness LR (2017) Profiling bacterial communities by MinION sequencing of ribosomal operons. Microbiome 5:116

Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 102:2567–2572

Konstantinidis KT, Rosselló-Móra R, Amann R (2017) Uncultivated microbes in need of their own taxonomy. ISME J 11:2399–2406

Konstantinidis KT, Rosselló-Móra R, Amann R (2020) Advantages outweigh concerns about using genome sequence as type material for prokaryotic taxonomy. Environ Microbiol 22:819–822

Kremer FS, McBride AJA, Pinto, L. da S. (2017) Approaches for *in silico* finishing of microbial genome sequences. Genet Mol Biol 40:553–576

Kumar PS, Brooker MR, Dowd SE, Camerlengo T (2011) Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. PLoS One 6:e20956

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML,

Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowki J, International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lee I, Kim YO, Park SC, Chun J (2016) OrthoANI: an improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol 66:1100–1103

Lee I, Chalita M, Ha SM, Na SI, Yoon SH, Chun J (2017) ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences. Int J Syst Evol Microbiol 67:2053–2057

Loman NJ, Quinlan AR (2014) Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics 30:3399–3401

Ludwig W, Schleifer KH (1994) Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS Microbiol Rev 15:155–173

Madigan MT, Martinko JM, Sthal DA, Clark DP (2012) Brock, biology of microorganisms, 13th edn. Pearson Education, San Francisco, CA

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Marmur J (1961) A procedure for the isolation of deoxyribonucleic acid from microorganisms. J Mol Biol 3:208–218

Marmur J, Doty P (1961) Thermal renaturation of deoxyribunucleic acids. J Mol Biol 3:585–594

Martijn J, Lind AE, Schön ME, Spiertz I, Juzokaite L, Bunikis I, Pettersson OV, Ettema TJG (2019) Confident phylogenetic identification of uncultured prokaryotes through long read amplicon sequencing of the 16S-ITS-23S rRNA operon. Environ Microbiol 21:2485–2498

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39:W339–W346

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15(6):589–594. https://doi.org/10.1016/j.gde.2005.09.006

Mehlen A, Goeldner M, Ried S, Stindl S, Ludwig W, Schleifer KH (2004) Development of a fast DNA-DNA hybridization method based on melting profiles in microplates. Syst Appl Microbiol 27:689–695

Meier-Kolthoff JP, Göker M (2019) TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat Commun 10:1–10

Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M (2013a) Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinf 14:60

Meier-Kolthoff JP, Göker M, Spröer C, Klenk HP (2013b) When should a DDH experiment be mandatory in microbial taxonomy? Arch Microbiol 195:413–418

Mukherjee S, Seshadri R, Varghese NJ, Eloe-Fadrosh EA, Meier-Kolthoff JP, Göker M, Coates RC, Hadjithomas M, Pavlopoulos GA, Paez-Espino D, Yoshikuni Y, Visel A, Whitman WB, Garrity GM, Eisen JA, Hugenholtz P, Pati A, Ivanova NN, Woyke T, Klenk H-P, Kyrpides NC (2017) 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nat Biotechnol 35:676–683

Murray R, Holt J (2005) The history of Bergey's manual. In: Garrity G, Boone D, Castenholz R (eds) Bergey's manual of systematic bacteriology. Springer, New York, pp 1–14

Na SI, Kim YO, Yoon SH, Ha SM, Baek I, Chun J (2018) UBCG: up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. J Microbiol 56:281–285

Nouioui I, Carro L, García-López M, Meier-Kolthoff JP, Woyke T, Kyrpides NC, Pukall R, Klenk HP, Goodfellow M, Göker M (2018) Genome-based taxonomic classification of the phylum actinobacteria. Front Microbiol 9:1–119

Oren A, Garrity GM (2018) Uncultivated microbes—in need of their own nomenclature? ISME J 12:309–311

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R (2014) The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res 42:206–214

Overmann J, Huang S, Nübel U, Hahnke RL, Tindall BJ (2019) Relevance of phenotypic information for the taxonomy of not-yet-cultured microorganisms. Syst Appl Microbiol 42:22–29

Palaniappan K, Chen I-MA, Chu K, Ratner A, Seshadri R, Kyrpides NC, Ivanova NN, Mouncey NJ (2019) IMG-ABC v.5.0: an update to the IMG/atlas of biosynthetic gene clusters knowledgebase. Nucleic Acids Res 48:D422–D430

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055

Peix A, Rivas R, Velázquez E, Mateos PF, Martínez-Molina E, Munoz-Herrera A, Gomez-Alonso-A, Velázquez E (2005) Application of horizontal staircase electrophoresis in agarose minigels to the random intergenic spacer analysis of clinical samples. Electrophoresis 26:4402–4410

Ranjard L, Poly F, Combrisson J, Richaume A, Gourbière F, Thioulouse J, Nazaret S (2000) Heterogeneous cell density and genetic structure of bacterial pools associated with various soil microenvironments as determined by enumeration and DNA fingerprinting approach (RISA). Microb Ecol 39:263–272

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

Riesco R, Carro L, Román-Ponce B, Prieto C, Blom J, Klenk HP, Normand P, Trujillo ME (2018) Defining the species *Micromonospora saelicesensis* and *Micromonospora noduli* under the framework of genomics. Front Microbiol 9:1–16

Rivas R, Velázquez E, Valverde A, Mateos PF, Martínez-Molina E (2001) A two primers random amplified polymorphic DNA procedure to obtain polymerase chain reaction fingerprints of bacterial species. Electrophoresis 22:1086–1089

Rivas R, Velázquez E, Zurdo-Piñeiro JL, Mateos PF, Martínez-Molina E (2004) Identification of microorganisms by PCR amplification and sequencing of a universal amplified ribosomal region present in both prokaryotes and eukaryotes. J Microbiol Methods 56(3):413–426

Salam N, Jiao J-Y, Zhang X-T, Li W-J (2020) Update on the classification of higher ranks in the phylum *Actinobacteria*. Int J Syst Evol Microbiol 70:1331–1355

Simon M, Scheuner C, Meier-Kolthoff JP, Brinkhoff T, Wagner-Döbler I, Ulbrich M, Klenk HP, Schomburg D, Petersen J, Göker M (2017) Phylogenomics of *Rhodobacteraceae* reveals evolutionary adaptation to marine and non-marine habitats. ISME J 11:1483–1499

Sokal RR, Sneath PHA (1964) Principles of numerical taxonomy. Freeman and Company, San Francisco, CA

Stackebrandt E, Ebers J (2006) Taxonomic parameters revisited: tarnished gold standards. Microbiol Today 06:152–155

Stackebrandt E, Goebel BM (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

Temperton B, Giovannoni SJ (2012) Metagenomics: microbial diversity through a scratched lens. Curr Opin Microbiol 15:605–612

Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG (2015) Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol 6:771

Velázquez E, Trujillo ME, Peix A, Palomo JL, García-Benavides P, Mateos PF, Ventosa A, Martínez-Molina E (2001) Stable low molecular weight RNA analyzed by staircase electrophoresis, a molecular signature for both prokaryotic and eukaryotic microorganisms. Syst Appl Microbiol 24:490–499

Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, Moore LH, Moore WEC, Murray RGE, Stackebrandt E, Starr MP, Truper HG (1987) Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol 37:463–464

Whitman WB (2015) Genome sequences as the type material for taxonomic descriptions of prokaryotes. Syst Appl Microbiol 38:217–222

Whitman WB (2016) Modest proposals to expand the type material for naming of prokaryotes. Int J Syst Evol Microbiol 66:2108–2112

Winand R, Bogaerts B, Hoffman S, Lefevre L, Delvoye M, Van Braekel J, Fu Q, Roosens NHC, De Keersmaecker SDJ, Vanneste K (2019) Targeting the 16S rRNA gene for bacterial identification in complex mixed samples: comparative evaluation of second (Illumina) and third (Oxford Nanopore Technologies) generation sequencing technologies. Int J Mol Sci 21:298

Winker S, Woese CR (1991) Definition of the Domains Archaea, Bacteria and Eucarya in terms of small subunit ribosomal RNA characteristics. Syst Appl Microbiol 14:305–310

Yang B, Wang Y, Qian PY (2016) Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinf 17:135

Yarza P, Ludwig W, Euzéby J, Amannd R, Schleifer K-H, Glöckner FO, Rosselló-Móra R (2010) Update of the all-species living tree project based on 16S and 23S rRNA sequence analyses. Syst Appl Microbiol 33:291–299

Zamora JC, Svensson M, Kirschner R, Olariaga I, Ryman S, Parra LA, Geml J, Rosling A, Adamčík S, Ahti T, Aime MC, Ainsworth AM, Albert L, Albertó E, García AA, Ageev D, Agerer R, Aguirre-Hudson B, Ammirati J, Andersson H, Angelini C, Antonín V, Aoki T, Aptroot A, Argaud D, Sosa BIA, Aronsen A, Arup U, Asgari B, Assyov B, Atienza V, Bandini D, Baptista-Ferreira JL, Baral HO, Baroni T, Barreto RW, Beker H, Bell A, Bellanger JM, Bellù F, Bemmann M, Bendiksby M, Bendiksen E, Bendiksen K, Benedek L, Bérešová-Guttová A, Berger F, Berndt R, Bernicchia A, Biketova AY, Bizio E, Bjork C, Boekhout T, Boertmann D, Böhning T, Boittin F, Boluda CG, Boomsluiter MW, Borovička J, Brandrud TE, Braun U, Brodo I, Bulyonkova T, Burdsall HH, Buyck B, Burgaz AR, Calatayud V, Callac P, Campo E, Candusso M, Capoen B, Carbó J, Carbone M, Castañeda-Ruiz RF, Castellano MA, Chen J, Clerc P, Consiglio G, Corriol G, Courtecuisse R, Crespo A, Cripps C, Crous PW, da Silva GA, da Silva M, Dam M, Dam N, Dämmrich F, Das K, Davies L, De Crop E, De Kesel A, De Lange R, Bonzi BDM, dela Cruz TEE, Delgat L, Demoulin V, Desjardin DE, Diederich P, Dima B, Dios MM, Divakar PK, Douanla-Meli C, Douglas B, Drechsler-Santos ER, Dyer PS,

Eberhardt U, Ertz D, Esteve-Raventós F, Etayo Salazar JA, Evenson V, Eyssartier G, Farkas E, Favre A, Fedosova AG, Filippa M, Finy P, Flakus A, Fos S, Fournier J, Fraiture A, Franchi P, Molano AEF, Friebes G, Frisch A, Fryday A, Furci G, Márquez RG, Garbelotto M, García-Martín JM, Otálora MAG, Sánchez DG, Gardiennet A, Garnica S, Benavent IG, Gates G, da Gerlach ACL, Ghobad-Nejhad M, Gibertoni TB, Grebenc T, Greilhuber I, Grishkan B, Groenewald JZ, Grube M, Gruhn G, Gueidan C, Gulden G, Gusmão LFP, Hafellner J, Hairaud M, Halama M, Hallenberg N, Halling RE, Hansen K, Harder CB, Heilmann-Clausen J, Helleman S, Henriot A, Hernandez-Restrepo M, Herve R, Hobart C, Hoffmeister M, Høiland K, Holec J, Holien H, Hughes K, Hubka V, Huhtinen S, Ivancevic B, Jagers M, Jaklitsch W, Jansen AE, Jayawardena RS, Jeppesen TS, Jeppson M, Johnston P, Jørgensen PM, Kärnefelt I, Kalinina LB, Kantvilas G, Karadelev M, Kasuya T, Kautmanová I, Kerrigan RW, Kirchmair M, Kiyashko A, Knapp DG, Knudsen H, Knudsen K, Knutsson T, Kolarík M, Kõljalg U, Košuthová A, Koszka A, Kotiranta H, Kotkova V, Koukol O, Kout J, Kovács GM, Kríž M, Kruys A, Kucera V, Kudzma L, Kuhar F, Kukwa M, Arun Kumar TK, Kunca V, Kušan I, Kuyper TW, Lado C, Læssøe T, Lainé P, Langer E, Larsson E, Larsson KH, Laursen G, Lechat C, Lee S, Lendemer JC, Levin L, Lindemann U, Lindström H, Liu X, Hernandez RCL, Llop E, Locsmándi C, Lodge DJ, Loizides M, Lokös L, Luangsa-Ard J, Lüderitz M, Lumbsch T, Lutz M, Mahoney D, Malysheva E, Malysheva V, Manimohan P, Marin-Felix Y, Marques G, Martínez-Gil R, Marson G, Mata G, Matheny PB, Mathiassen GH, Matocec N, Mayrhofer H, Mehrabi M, Melo I, Mešic A, Methven AS, Miettinen O, Romero AMM, Miller AN, Mitchell JK, Moberg R, Moreau PA, Moreno G, Morozova O, Morte A, Muggia L, González GM, Myllys L, Nagy I, Nagy LG, Neves MA, Niemelä T, Nimis PL, Niveiro N, Noordeloos ME, Nordin A, Noumeur SR, Novozhilov Y, Nuytinck J, Ohenoja E, Fiuza PO, Orange A, Ordynets A, Ortiz-Santana B, Pacheco L, Pál-Fám F, Palacio M, Palice Z, Papp V, Pärtel K, Pawlowska J, Paz A, Peintner U, Pennycook S, Pereira OL, Daniëls PP, Capella MÀPDG, del Amo CMP, Gorjón SP, Pérez-Ortega S, Pérez-Vargas I, Perry BA, Petersen JH, Petersen RH, Pfister DH, Phukhamsakda C, Piatek M, Piepenbring M, Pino-Bodas R, Esquivel JPP, Pirot P, Popov ES, Popoff O, Álvaro MP, Printzen C, Psurtseva N, Purahong W, Quijada L, Rambold G, Ramírez NA, Raja H, Raspé O, Raymundo T, Réblová M, Rebriev YA, de Reyes García JD, Ripoll MÁR, Richard F, Richardson MJ, Rico VJ, Robledo GL, Barbosa FR, Rodriguez-Caycedo C, Rodriguez-Flakus P, Ronikier A, Casas LR, Rusevska K, Saar G, Saar I, Salcedo I, Martínez SMS, Montoya CAS, Sánchez-Ramírez S, Sandoval-Sierra JV, Santamaria S, Monteiro JS, Schroers HJ, Schulz B, Schmidt-Stohn G, Schumacher T, Senn-Irlet B, Ševcíková H, Shchepin O, Shirouzu T, Shiryaev A, Siepe K, Sir EB, Sohrabi M, Soop K, Spirin V, Spribille T, Stadler M, Stalpers J, Stenroos S, Suija A, Sunhede S, Svantesson S, Svensson S, Svetasheva TY, Swierkosz K, Tamm H, Taskin H, Taudière A, Tedebrand JO, Lahoz RT, Temina M, Thell A, Thines M, Thor G, Thüs H, Tibell L, Tibell S, Timdal E, Tkalcec Z, Tønsberg T, Trichies G, Triebel D, Tsurykau A, Tulloss RE, Tuovinen V, Sosa MU, Urcelay C, Valade F, Valenzuela Garza R, van den Boom P, Van Vooren N, Vasco-Palacios AM, Vauras J, Santos JMV, Vellinga E, Verbeken A, Vetlesen P, Vizzini A, Voglmayr H, Volobuev S, von Brackel W, Voronina E, Walther G, Watling R, Weber E, Wedin M, Weholt Ø, Westberg M, Yurchenko E, Zehnálek P, Zhang H, Zhurbenko MP, Ekman S (2018) Considerations and consequences of allowing DNA sequence data as types of fungal taxa. IMA Fungus 9:167–175