AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology

MINIREVIEW

A New Genomics-Driven Taxonomy of *Bacteria* and *Archaea*: Are We There Yet?

George M. Garrity

Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan, USA; NamesforLife, LLC, East Lansing, Michigan, USA

Taxonomy is often criticized for being too conservative and too slow and having limited relevance because it has not taken into consideration the latest methods and findings. Yet the cumulative work product of its practitioners underpins contemporary microbiology and serves as a principal means of shaping and referencing knowledge. Using methods drawn from the field of exploratory data analysis, this minireview examines the current state of the field as it transitions from a taxonomy based on 16S rRNA gene sequences to one based on whole-genome sequences and tests the validity of some commonly held beliefs.

More than a quarter century has passed since Woese et al. (1) proposed a natural system for classifying *Bacteria*, *Archaea*, and *Eukarya* based on comparisons of gene sequences of the small ribosomal subunit (SSU). In the ensuing years, this method has become a standard for identifying *Bacteria* and *Archaea* (prokaryotes) (2). Ideally, when supplemented with additional phenotypic and genotypic data (polyphasic data), it provides a rich description of each strain that places it into a biologically meaningful context (3). Coverage of the 16,343 species, subspecies, and higher taxa with validly published names is nearly complete, with each being anchored by one or more high-quality 16S rRNA reference sequences derived from type strains (4). This represents a unique and unparalleled resource in the life sciences and is analogous to a base map onto which related strains and metadata can be mapped (5, 6). It also serves as the ultimate reference source against which new identification methods must be continuously evaluated and validated, and the source of new knowledge and new names that will ultimately be incorporated into systems intended for use in clinical and quality control microbiology laboratories.

Although 16S rRNA sequence analysis will remain one of the first steps in contemporary identification schemes for both cultured and uncultured prokaryotes in the foreseeable future, there is already a movement under way to employ whole-genome sequences (WGS) as the next logical step in classification schemes (7–12). Comparisons of WGS may overcome the principal shortcoming of 16S rRNA sequence analysis (limited taxonomic resolution) (2, 3); can provide direct evidence about the metabolic, structural, and functional potential of an organism; and can provide indirect evidence by inference about the same properties of closely related species. However, WGS currently has shortcomings when it is used to place species into higher taxa.

Although the transition to a genome-based taxonomy will introduce new steps into the taxonomist's workflow, it will likely proceed in a fashion similar to that for the adoption of new technologies in the past (10). Algorithms and heuristics for analyzing and interpreting genome-based classifications have already been adopted by the broader community (7, 11, 13, 14), and coordinated efforts are under way to assemble a collection of complete or high-quality draft genomes for each type strain (12, 15–19). What differs about this transition compared to earlier ones is the speed at which it is occurring because of the ease and low cost of producing sequence data and the widespread availability of the technology. This is leading to an increase in the number of novel taxa being asserted by a much larger community of microbiologists, much of it outside the taxonomic literature. However, data interpretation, taxonomic inference, and hypothesis generation will likely remain a human activity, at least for now. So too, will be the challenge of linking the scientific, technical, medical, legal, and general literature where inferences are made based on names that may or may not be current, be correct, or have any meaning (20).

Considerable progress has been made in developing methods to delimit species-, subspecies-, and strain-level relationships based on pairwise comparisons of genome sequences. The theories of techniques such as average nucleotide identity (ANI), average amino acid identity (AAI), and digital DNA-DNA hybridization (DDH) are presented elsewhere (7, 11, 13, 14, 21, 22), but it is noteworthy that all are rooted in earlier methods. These will be discussed in the context of the challenges that they present in developing an all-encompassing taxonomy. The focus of this review will be the technical and social challenges we will likely encounter and will draw on insights gained from the 2-decade transition to a 16S rRNA-based taxonomy. Data used in the analyses presented here are from the NamesforLife database (www.namesforlife.com; NamesforLife, LLC, East Lansing, MI), which contains a complete record of the taxonomic and nomenclatural record events of *Bacteria* and *Archaea*, modeled according to the systems and methods of Garrity and Lyons (23, 24).

TAXONOMY IN BRIEF

In the strictest sense, taxonomy deals with the theory and practice of classification, including the principles, rules, and methods (25, 26). In the life sciences, taxonomists typically engage in the development and maintenance of systems and methods for classifying and identifying different groups of organisms (taxa), followed by the formation and application of a name in conformance with the specific code of nomenclature (27). When novel species or higher

Accepted manuscript posted online 18 May 2016

Citation Garrity GM. 2016. A new genomics-driven taxonomy of *Bacteria* and *Archaea*: are we there yet? J Clin Microbiol 54:1956–1963. doi:10.1128/JCM.00200-16.

Editor: C. S. Kraft, Emory University

Address correspondence to garrity@msu.edu.

For a commentary on this article, see doi:10.1128/JCM.01082-16

TABLE 1 Validly published named prokaryotes used in the analyses[a]

Rank	No. validly published	No. covered (16S)[b]	No. of genomes	No. of outliers	No. of singletons	No. with taxon size[f] of:		
						2 or 3	4 to 9	≥10
Species	12,981	11,800	4,092[c]	274	11,800	0	0	0
Genus	2,716	2,422	1,333	131	1,166	616	437	203
Family	389	451[d]	343	86	121	71	79	180
Order	174	202[d]	161	47	52	28	34	90
Class	83	85[d]	69	19	23	13	14	35
Phylum	0	34[e]	33	9	4	4	6	20
Domain	0	2	2	2	0	0	0	2

[a] As of 5 December 2015, there were 12,981 validly published named species and subspecies of *Bacteria* and *Archaea*.
[b] A review of the taxonomic literature yielded 12,481 sequences (16S rRNA) that could be confirmed as derived from type strains, of which 11,800 were considered of sufficient quality to be used in analyses (>1,300 nt, <3% ambiguity); A total of 407 sequences were excluded from the analyses because of quality considerations, and 274 were flagged as outliers during the first pass through the self-organizing self-correcting classifier.
[c] At present, there are 4,733 genome sequences of varying quality and completeness that can be identified with some certainty as derived from type strains. Of these, 641 are replicates that have been sequenced 2 to 10 times and appear in GenBank with a sequence name can be positively associated with a known strain identifier or culture collection accession number for the type strain. Three of the genomes have been sequenced multiple times as synonyms.
[d] There are currently 92 species that have not been placed into families, orders, and/or classes. In these instances, the lower taxa are placed into unnamed, numbered ranked categories designated as *Incertae sedis*.
[e] Currently, the names of the phyla and domains are not covered by the International Code of Nomenclature of Prokaryotes and cannot be considered validly published.
[f] Taxon size is the number of members in the next lower rank. For example, 4 phyla have 2 or 3 classes, 6 phyla have 4 to 9 classes, and 20 phyla have ≥10 classes.

taxa are discovered or when existing taxa are revised by either combining or splitting them, taxonomists prepare and publish formal descriptions of each to establish the names and circumscribe the corresponding taxonomic concepts or to emend existing ones to reflect their new findings. Taxonomists also will prepare special reference material of each new species or subspecies (type specimens) and deposit that material in the appropriate public repositories for future use by others. In botany and zoology, the type material is fixed (preserved and nonviable). For prokaryotes, the type material must be preserved and viable. The published descriptions follow specific formats and establish each name, its rank, the etymology of the name, a description of the diagnostic characteristics of the taxon by which it can be recognized, and information about where the type material is available (27). More recently, INSDC (International Nucleotide Sequence Database Consortium) identifiers are being included for deposited sequences. The description also establishes the boundaries of the taxon (e.g., other members and the parent and child taxa within the taxonomic hierarchy). The description is attributed to the taxonomist(s) who authored it and the page on which it is found in a published article or a monograph that is available to the scientific community. In the case of a taxonomic revision, emendation, or revival of a name, the names of other authors may also be also cited, according to the Code (27).

A biological name gains standing in nomenclature only if it conforms to all of the rules of the governing code of nomenclature. The code establishes how names must be formed and applied to taxonomic concepts, as well as which name is correct when referring to a given taxon When done properly, a biological name stands for all that was known about a taxon at the time it was first described or subsequently emended. But, once names and taxonomic concepts come into use (e.g., in identification schemes used in clinical microbiology), other members and close relatives are recognized and other genotypic and phenotypic properties that were not in the original description are revealed. This leads to gradual deviation between the taxonomic concept and the name that is used in discourse and the literature and the original or emended published description. This is a source of confusion for many microbiologists. A solution to this problem has been proposed by Garrity and Lyons (23, 24).

A common misconception is that there is an official taxonomy. That is true only for viruses (28). The remaining codes all protect freedom of taxonomic thought, which is essential to guarantee that taxonomies reflect current knowledge and do not become rooted in obsolete concepts. A second common misconception is that taxonomic descriptions and the corresponding names are facts. They are not. Rather they are hypotheses and falsifiable when synonymized, revised, or emended (29).

THE GLOBAL TAXONOMY OF *BACTERIA* AND *ARCHAEA*

The universal applicability of 16S rRNA sequence analysis to *Bacteria* and *Archaea* has led to the emergence of a taxonomy that encompasses all of the species within these two domains (Table 1), whether or not a given strain is cultivable. It also changed the manner in which most taxonomists work, significantly democratizing the field. Where most taxonomists traditionally focused on specific regions of "the tree" or on groups that had comparable physiological properties or ecological roles, sequence-based methods have allowed contemporary taxonomists to work much more broadly and to make inferences about similarities in phenotype and ecotype based on what is now commonly referred to as the "phylotype" (30).

The current global taxonomy of *Bacteria* and *Archaea* represents a consensus view that has been undergoing constant refinement, revision and expansion for 35 years. It is the work product of over 17,500 authors of more than 20,600 taxonomic descriptions appearing in 12,195 effective publications. The taxonomy includes the most recently discovered cultivable species, as well as a small number of historically relevant taxa that predate the approved lists (31), some of which were described more than 150 years ago. At present, the average age of a taxon with a validly published name is 16.8 years (excluding those names published in 2015). The rate of synonymizing/reclassifying taxa is 13.8%; the rate of explicit emendation of existing taxa is 7.8%. These rates will likely change once WGS methods of classification are applied more frequently to the current type strains (11), much the same as

happened with 16S rRNA sequence analysis. The possible relaxation of Rules 27 and 30 of the Code (27), to allow genome sequences to serve as type material, might also have a dramatic effect on the number of new taxa proposed, if the trend observed with discovery of "novel species" by 16S rRNA sequence analysis is any indication of future trends (20). The wisdom of such an action would likely lead to a much needed debate on the value of naming such taxa (see Principal 1 in the Code [27] and Sneath [32]) or if some alternative approach to tracking putatively identified novel taxa might be more useful (e.g., the *Candidatus* concept [27] or the semantic model of Garrity and Lyons [23, 24]).

Further examination of the bibliometric data reveals that approximately 90% of the authors of taxonomic proposals can be described as "occasional taxonomists," having described 10 or fewer taxa during their careers. However, the top 100 taxonomists published 93 to 637 proposals during careers that averaged 22.1 years (see Table 1S posted at https://www.researchgate.net/publication/302545989_JCM00200-16-supplement). Within this group, productivity appears to be a function of when they worked and the size of their collaborative network. Some of the more prolific authors began their careers more recently and are affiliated with large laboratories, national culture collections, or comparable organizations that are equipped with the latest technology. Other trends that can be gleaned from the published record include the number of authors on proposals. Many of the recent papers have 20 or more authors. It would appear that taxonomy has become a part of contemporary "big science."

EXPLORING PROKARYOTIC TAXONOMY; RECONCILIATION OF THE NOMENCLATURE AND SEQUENCE DATA

Names of organisms fall into two broad categories: informal (vernacular) and formal (scientific) (32). Vernacular names are used as a matter of convenience and may include lab strain designations that may be used in combination with a taxonomic name and/or other labels such as serovar, pathovar, or biovar designation or culture collection accession number. Vernacular names are unregulated. Most lack a clear definition and may be a synonym, a polyseme, or both. On the other hand, the formation and use of scientific names are governed, and each name stipulates a precise rank and location of a named organism or group of named organisms within a taxonomic hierarchy. Use of a scientific name implies agreement with the published description and taxonomy in which it was established. Since names and taxa can be synonymized and emended, any given organism may bear one or more names that can be applied at a given time. The correct name within a given circumscription, position, and rank is the earliest one that conforms to the Code (27, 32, 33). Reconciliation of an existing nomenclature with a new or emerging theory of classification (e.g., 16S rRNA phylogeny or WGS taxonomy) is a time-consuming task and requires testing the hierarchy implied by the names against alternative models to identify areas where there is agreement or disagreement and to adjust the prevailing theory to fit reality. The same approach can be used when different taxonomies are compared. It is important to note that evaluation of taxonomies is an ongoing process. Refinements, explanations, and insights accumulate continuously. Adjustments must also be made continuously, while maintaining backwards compatibility and may be particularly relevant for identification schemes used in regulated environments.

Historically, our approach (6, 34–36) to reconciling nomenclature and 16S rRNA phylogeny was to employ methods from the field exploratory data analysis (37, 38). Rather than relying on phylogenetic trees, which either lacked resolution when collapsed or were too unwieldy when viewed in full (>10,000 nodes), we applied principal-component analysis (PCA) to the 16S rRNA sequence data. Preliminary experiments with Peter Sneath provided early evidence that this approach would be workable (39). Using a novel approach to both fix the derived coordinates and overlay subsets of sequences based on assigned names, we were able to quickly identify areas of disagreement between the existing nomenclature and the phylogenetic model. Scree plots revealed that the first three dimensions in the derived coordinate system accounted for >90% of the total variance, indicating the reliability of the projections (34, 35). PCA also provided a unique view of the topology of bacterial and archaeal domains, which has remained remarkably stable, robust, and resistant to perturbation by different alignments, different implementations of PCA, or an almost 3-fold increase in the number of validly named species. The PCA plots suggest that there are a finite number of major groups into which the validly named species fall based on 16S rRNA sequence analysis, which correspond to the major phyla; the clear separation of the *Bacteria* and *Archaea* and the presence of "white space" in the plots appear to be inviolate (Fig. 1; see also Fig. 1S posted at https://www.researchgate.net/publication/302545989_JCM00200-16-supplement).

The white space suggests evolutionary constraints in the primary and secondary structures of the 16S rRNA genes of *Bacteria* and *Archaea* and to a lesser extent, those of the dominant bacterial phyla, especially the *Actinobacteria*. Subsequent experiments focused on potential distortion arising from the reduction in dimensionality by PCA, as the method is known to introduce distortions (arching and horseshoe effects) when input data are tailed or are composed mainly of closely spaced values. While views of the data can be corrected by rotation, distortion remains. Garrity and Lilburn (6) found that such distortion could be eliminated by using estimates of the distance to external reference points rather than pairwise measurements. This principle was successfully demonstrated with a test case having a known solution (geographic data). However, a similar solution for phylogenetic data has not yet been found.

A follow-up investigation employed heatmaps for visualizing large matrices of sequence similarity data. Heatmaps are colorized, shaded matrices and provide a distortion-free method for visualizing discrepancies in a classification. The technique is simple and scalable and has an added advantage in that it can reveal the presence of nested hierarchies in data sets. The method is applicable to both symmetric (reflected) and asymmetric matrices and is therefore useful in developing identification schemes as well as taxonomies. Reordering of the underlying matrix is remarkably simple as it is based on the order or appearance of (26) a set of taxa in an input tree (or other classification).

A self-organizing self-correcting classifier (SOSCC) was subsequently developed to automate detection and correction of classification errors (40). The algorithm applies an simple criterion (e.g., a 2-standard deviation within-taxon spread in 16S rRNA similarity) at all levels of an input taxonomy while preserving nomenclatural integrity. The algorithm also smooths by reclustering taxa at each level of the hierarchy. The effects of the SOSCC on the global taxonomy can be seen in a series of heatmaps in Fig. 2S posted at https://www.researchgate.net/publication/30

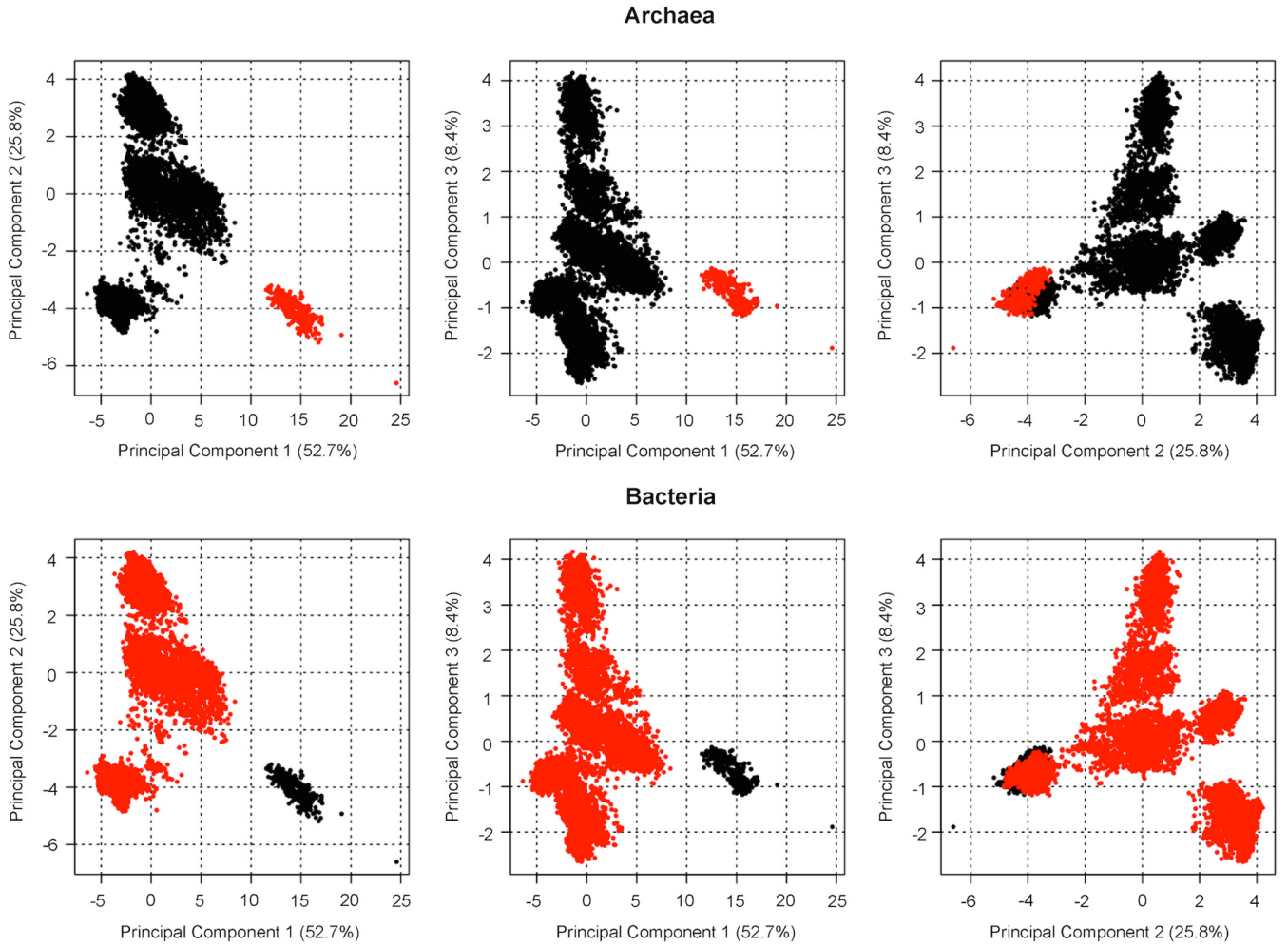


FIG 1 Principal-component analysis of 16S rRNA sequences from type strains of validly named species and subspecies of *Bacteria* and *Archaea*. PCA was used to visualize the distribution pattern of 16S rRNA sequence in a manner similar to that previously described by Garrity and Lilburn (25) with the following modifications. High-quality, curated sequences, matched to the published taxonomic descriptions of type strains, were aligned using the most recent Greengenes alignment, and uncorrected pairwise evolutionary distances were determined in Mothur (version 1.34.1) using the default parameters (Needleman-Wunsch method, kmer searching with 8 mer). The resulting matrix was then subjected to a PCA in R (version 3.3.2, 64 bit, Mac OS). The first three principal components accounted for 86.9% of the total variance. Areas highlighted in red indicate locations of members of the two prokaryotic domains. The two outliers located in the lower right quadrant are *Methanocorpusculum sinense* (GenBank accession no. AF095268.1) and *Halostagnicola bangensis* (HF544345.1). Additional PCA plots showing the locations of the validly named orders and classes are available in Fig. 1S posted at https://www.researchgate.net/publication/302545989_JCM00200-16-supplement. A complete set of figures is available for viewing at http://dx.doi.org/10.1601/tx.1.

2545989_JCM00200-16-supplement. Refinements and reconciliation of the nomenclature and taxonomy occurring between 1980 and 2015 are shown in Fig. 3S posted at https://www.researchgate.net/publication/302545989_JCM00200-16-supplement.

The SOSCC also supports testing of hypotheses about taxon membership based on 16S sequence similarity. For nearly 2 decades, the community has applied a heuristic that strains that share ≤97% sequence similarity between 16S RNA gene sequences (≥1,300 nucleotides [nt] in length) are members of different species within the same genus, and that threshold has been gradually increasing compared to limited subsets of species for which DNA hybridization data were available (41, 42). Additional proposals have emerged for defining taxonomic relationships from the family to the phylum level based on 16S sequence similarity (42). However, an examination of the data for 11,824 type strains comprising 13,991 species and subspecies with validly published names reveals that the distribution of pairwise similarities varies significantly. Many higher taxa are nonnormally distributed, especially at the genus level (see Fig. 4S posted at https://www.researchgate.net/publication/302545989_JCM00200-16-supplement). There is also significant overlap in the distribution of 16S rRNA sequence scores within the higher taxa, calling into question whether the ranks of family, class, and order are justified (Fig. 2). While the grand mean of genus-level 16S rRNA sequence similarity among all of the type strains is 95.4% (n = 10,788), the range of within-genus means is significantly greater, suggesting that the consensus heuristic poorly reflects the genus- or higher-level relationships in the consensus taxonomy (see Table 2S posted at https://www.researchgate.net/publication/302545989_JCM00200-16-supplement). This is a particularly interesting problem, given that recently described taxa were formed by the application of this heuristic. This also suggests that if one were to adopt a taxonomy with rigid

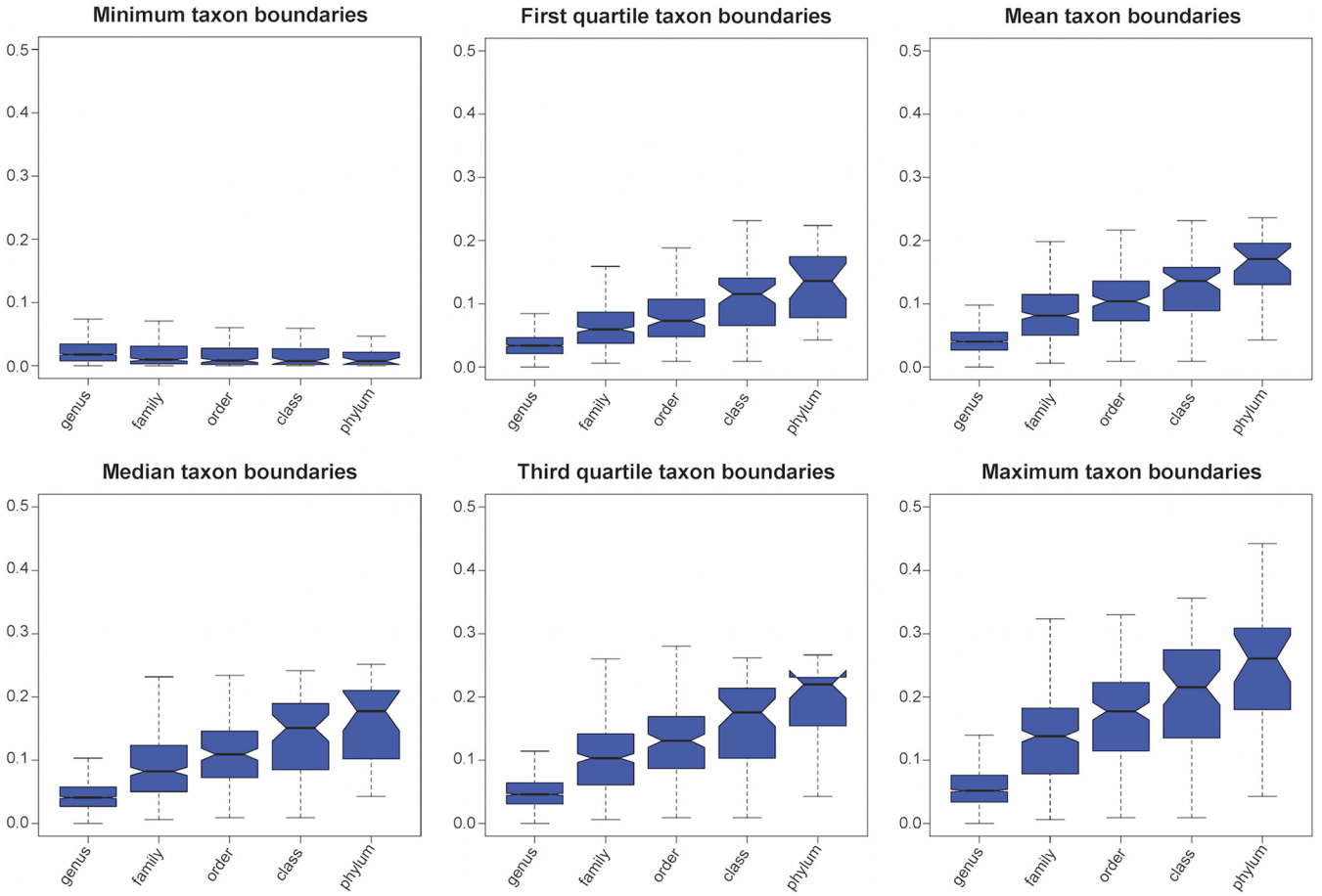


FIG 2 Distribution of sequence similarity among taxon members. The distribution of 16S sequence similarity was determined from the output of the SOSCC analysis of the complete matrix (n = 11,814) for each taxon for the rank of genus to domain, excluding singletons. Pairwise similarities were extracted from the lower triangle of the matrix (excluding self-identity scores on the diagonal). The five number summaries (mean, median, maximum, and first and third quartiles) (11) were computed for each taxon, and boxplots summarizing the distribution of scores at each rank were drawn. Output of the SOSCC is provided in Table S2 posted at https://www.researchgate.net/publication/302545989_JCM00200-16-supplement.

cutoffs as recently suggested by Yarza et al. (42), the net effect would be a significant inflation in the number of species and higher taxa. Whether or not such an action is warranted in the absence of supporting data is an issue worthy of public debate as such a move would have a significant impact on the literature and public databases.

INCORPORATION OF WHOLE-GENOME SEQUENCE DATA INTO TAXONOMIC MODELS

As noted above, the recasting of the taxonomy of the prokaryotes based on WGS is already under way. New bioinformatics methods that overcome the technical limitations of classical DNA hybridization measurements used to delineate prokaryotic species are now used routinely, the most common of which is ANI (7, 13, 14, 21). ANI provides an estimate of sequence similarity between pairs of genomes (designated the query and reference genome) and computes this value for regions in the genome that exhibit at least 70% sequence similarity across 35% or more of the alignable region, using either BLAST or MUMmer. Improvements in the method in which either the query or the query and the reference genomes are "shredded" into "fragments" of 1,020 bases to more closely mimic DDH have recently appeared.

The emerging heuristic is that species-level relationships occur in the range of 94 to 96% ANI. However, it is important to note that ANI scores between a query and reference genome are often asymmetric because of differences in gene complements and genome sizes. This asymmetry is not entirely surprising as it was often observed in reciprocal hybridization studies using labeled DNA probes in the past. However, there was no plausible explanation for the phenomenon at the time. Like DDH, the current implementation of ANI also has limited utility: defining species-, subspecies-, and strain-level relationships. The level of nucleotide identity occurring over more distantly related strains falls off sharply; thus, ANI in its current form does little for redefining higher taxa. This was recently discussed by Kim et al. (43) in a comparative study of 6,787 genomes from 22 phyla. These authors proposed narrowing the species-level ANI cutoff to 96% with a corresponding cutoff of 16S rRNA sequence similarity of 98.65%. It does not appear, however, that this study encompassed a significant number of species from the major phyla (e.g., the *Actinobacteria*), and their data set was heavily biased toward medically relevant species, which included as many as 512 genomes of a single species. The ramification of their recommendation is a significant inflation in the numbers of species and genera. One would also need to consider if it would be possible to differentiate those spe-

cies using alternative methods, thereby substantiating such proposals. The impact on the higher taxonomy is unclear.

In a similar study, Varghese et al. (11) examined 13,151 microbial genomes representing 3,082 named species by ANI paired with an estimate of the alignment fraction of genomic DNA for each pair, creating a paired-value microbial species identifier. These authors reported a high correlation between the two measures and used the combined score to group the genomes into cliques, clique groups, and singletons by application of the Bron-Kerbosch algorithm. They also reported a number of misplaced/misclassified species (ca. 18%); however, since the study was neither restricted to type strains of species with validly published names nor did it fully cover all of the validly published species of *Bacteria* and *Archaea*, these authors wisely chose to not reclassify or rename any of the taxa based on their method.

A NEW CONSENSUS TAXONOMY

Given that taxonomy represents a consensus opinion and the prevailing view is that the current taxonomy of prokaryotes may not take into consideration all of our new knowledge gained through genome sequencing, it is time to recast it again. However, rather than disregarding the current and past taxonomies, it may be prudent to assess where the new methods add to existing knowledge, what limitations may exist, and how to integrate the new methods so as to refine our knowledge rather than lose our knowledge. Currently, <40% of the type strains of validly published species and subspecies have been sequenced. Although a working strategy was recently proposed to fill in gaps with "proxytypes" (15), these cannot replace the types. The time needed to fully complete the collection and keep pace with newly described species will be longer than most would expect. So too will documenting all of the sequences to establish provenance. Linkage of ecological and phenotypic data to genome sequences is also essential to establish the role that each organism may play in a microbiome or what effect it may have on its environment. While genomes may support inference of a property, confirmation requires experimental proof, much of which may already exist. Linking such observational data to the correct genomes requires accurate taxonomic information that is usually inferred through a name. Persistent linking of the correct data and metadata to particular species will become increasingly important as taxonomic work accelerates to keep pace with applications of new sequencing methodologies. So too, will the demand for taxonomists who assert novel taxa to verify their findings with experimental evidence, viable type material, and detailed comparisons with previously described taxa. Simply stating that a particular genome sequence falls above or below an arbitrary threshold is not adequate as a hypothesis. Hopefully the tendency toward "assembly line" taxonomy will fade away and be replaced by more rigorous approaches. Taxonomic hypotheses require a willingness to interpret the data and literature to determine not only what is new but also what is known. WGS has made data generation incredibly easy, but data analysis and interpretation and synthesis of knowledge still require significant effort, insight, and domain expertise. New tools to mine the literature and reason across complex phenotypic and environmental data are in development and will aid taxonomists in the future as they deal with an ever-increasing body of prior knowledge. However, the ultimate decision will still fall to the individual(s) who will make the judgment call in determining the novelty of a taxon.

ACKNOWLEDGMENTS

I am grateful for constructive comments from Charles T. Parker, Brian J. Tindall, Nikos Kyrpides, David Ussery, Miriam Land, Terrence Marsh, and the three anonymous peer reviewers. I am indebted to the late P.H.A. Sneath and John W. Tukey for their numerous and helpful discussions about numerical taxonomy and exploratory data analysis early in my career.

Portions of this work were funded and supported by the Office of Science (BER), U.S. Department of Energy, under grant numbers DE-FG02-04ER63933, DE-FG02-07ER86321, and DE-SC0006191.

NamesforLife semantic resolution technology and the SOSCC systems and methods are covered under U.S. patents 7,925,444 and 8,036,997 under license from the Michigan State University Board of Trustees to NamesforLife, LLC.

NamesforLife, LLC, is an Michigan State University start-up company formed to commercialize systems and methods for classification, searching, and indexing text and data. The author is a cofounder and principal in the company and principal investigator on five STTR awards from the Department of Energy to further develop the technology and grants and loans from the Michigan Economic Development Corporation and the Michigan Universities Commercialization Initiative to develop prototype products for commercial applications.

FUNDING INFORMATION

This work, including the efforts of George M. Garrity, was funded by United States Department of Energy Office of Science (DE-SC0006191, DE-FG02-04ER63933, and DE-FG02-07ER86321).

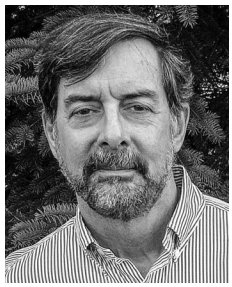
Portions of the funding are through the SBIR/STTR program.

REFERENCES

1. **Woese CR, Kandler O, Wheelis ML.** 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A **87:**4576–4579. http://dx.doi.org/10.1073/pnas.87.12.4576.
2. **Stackebrandt E, Frederiksen W, Garrity GM, Grimont PA, Kampfer P, Maiden MC, Nesme X, Rossello-Mora R, Swings J, Truper HG, Vauterin L, Ward AC, Whitman WB.** 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol **52:**1043–1047. http://dx.doi.org/10.1099/00207713-52-3-1043.
3. **Tindall BJ, Rossello-Mora R, Busse HJ, Ludwig W, Kämpfer P.** 2010. Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst Evol Microbiol **60:**249–266. http://dx.doi.org/10.1099/ijs.0.016949-0.
4. **Yarza P, Sproer C, Swiderski J, Mrotzek N, Spring S, Tindall BJ, Gronow S, Pukall R, Klenk HP, Lang E, Verbarg S, Crouch A, Lilburn T, Beck B, Unosson C, Cardew S, Moore ER, Gomila M, Nakagawa Y, Janssens D, De Vos P, Peiren J, Suttels T, Clermont D, Bizet C, Sakamoto M, Iida T, Kudo T, Kosako Y, Oshida Y, Ohkuma M Arahal DR, Spieck E, Pommerening Roeser A, Figge M, Park D, Buchanan P, Cifuentes A, Munoz R, Euzeby JP, Schleifer KH, Ludwig W, Amann R, Glockner FO, Rossello-Mora R.** 2013. Sequencing orphan species initiative (SOS): filling the gaps in the 16S rRNA gene sequence database for all species with validly published names. Syst Appl Microbiol **36:**69–73. http://dx.doi.org/10.1016/j.syapm.2012.12.006.
5. **Garrity GM.** 2009. Ground truth. Stand Genomic Sci **1:**91–92, 2009. http://dx.doi.org/10.4056/sigs.50595.
6. **Garrity GM, Lilburn TG.** 2002. Mapping taxonomic space: an overview of the road map to the 2nd ed of *Bergey's Manual of Systematic Bacteriology*. WFCC Newsl **35:**5–15.
7. **Auch AF, Von Jan M, Klenk H-P, Göker M.** 2010. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci **2:**117–134. http://dx.doi.org/10.4056/sigs.531120.
8. **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. http://dx.doi.org/10.1099/ijs.0.054171-0.
9. **Garrity GM, Banfield J, Eisen J, Van Der Lelie N, McMahon T, Rusch**

D, Delong E, Moran MA, Currie C, Furhman J, Hallam S, Hugenholtz P, Moran N, Nelson K, Roberts R, Stepanauskas R. 2013. Prokaryotic Super Program Advisory Committee DOE Joint Genome Institute, Walnut Creek, CA, March 27, 2013. Stand Genomic Sci **8:**561–570. http://dx.doi.org/10.4056/sigs.4638348.

10. **Oren A, Garrity GM.** 2014. Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. Antonie Van Leeuwenhoek **106:**43–56. http://dx.doi.org/10.1007/s10482-013-0084-1.
11. **Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A.** 2015. Microbial species delineation using whole genome sequences. Nucleic Acids Res **43:**6761–6771. http://dx.doi.org/10.1093/nar/gkv657.
12. **Whitman WB, Woyke T, Klenk H-P, Zhou Y, Lilburn TG, Beck BJ, De Vos P, Vandamme P, Eisen JA, Garrity G, Hugenholtz P, Kyrpides NC.** 2015. Genomic encyclopedia of bacterial and archaeal type strains, phase III: the genomes of soil and plant-associated and newly described type strains. Stand Genomic Sci **10:**26. http://dx.doi.org/10.1186/s40793-015-0017-x.
13. **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. http://dx.doi.org/10.1099/ijs.0.64483-0.
14. **Konstantinidis KT, Ramette A, Tiedje JM.** 2006. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci. **361:**1929–1940. http://dx.doi.org/10.1098/rstb.2006.1920.
15. **Federhen S, Rosella-Mora R, Klenk H-P, Tindall BJ, Konstantinidis KT, Whitman WB, Brown D, Labeda D, Ussery D, Garrity GM, Colwell RR, Nur H, Graf J, Parte A, Yarza P, Goldber B, Sichtig H, Karsch-Mizrachi I, Cark K, Mcveigh R, Pruitt K, Tatusov T, Falk R, Turner S, Madden T, Kits P, Klimke W, Kimchi A, Agarwala R, Dicuccio M, Ostell J.** 2016. Meeting report: GenBank microbial genomic taxonomy workshop (12−13 May 2015). Stand Genomic Sci **11:**15. http://dx.doi.org/10.1186/s40793-016-0134-1.
16. **Göker M, Klenk HP.** 2013. Phylogeny-driven target selection for large-scale genome-sequencing (and other) projects. Stand Genomic Sci **8:**360–374. http://dx.doi.org/10.4056/sigs.3446951.
17. **Kyrpides NC, Hugenholtz P, Eisen JA, Woyke T, Goker M, Parker CT, Amann R, Beck BJ, Chain PS, Chun J, Colwell RR, Danchin A, Dawyndt P, Dedeurwaerdere T, Delong EF, Detter JC, De Vos P, Donohue TJ, Dong XZ, Ehrlich DS, Fraser C, Gibbs R, Gilbert J, Gilna P, Glockner FO, Jansson JK, Keasling JD, Knight R, Labeda D, Lapidus A, Lee JS, Li WJ, Ma J, Markowitz V, Moore ER, Morrison M, Meyer F, Nelson KE, Ohkuma M, Ouzounis CA, Pace N, Parkhill J, Qin N, Rossello-Mora R, Sikorski J, Smith D, Sogin M, Stevens R, Stingl U, Suzuki K, Taylor D, Tiedje JM, Tindall B, Wagner M, Weinstock G, Weissenbach J, White O, Wang J, Zhang L, Zhou YG, Field D, Whitman WB, Garrity GM, Klenk HP.** 2014. Genomic encyclopedia of Bacteria and Archaea: sequencing a myriad of type strains. PLoS Biol **12:**e1001920. http://dx.doi.org/10.1371/journal.pbio.1001920.
18. **Kyrpides NC, Woyke T, Eisen JA, Garrity G, Lilburn TG, Beck BJ, Whitman WB, Hugenholtz P, Klenk HP.** 2014. Genomic encyclopedia of type strains, phase I: the one thousand microbial genomes (KMG-I) project. Stand Genomic Sci **9:**1278–1284. http://dx.doi.org/10.4056/sigs.5068949.
19. **Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng JF, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk HP, Eisen JA.** 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature **462:**1056–1060. http://dx.doi.org/10.1038/nature08656.
20. **Garrity GM, Oren A.** 2013. Response to Sutcliffe et al.: regarding the International Committee on Systematics of Prokaryotes. Trends Microbiol **21:**53–55. http://dx.doi.org/10.1016/j.tim.2012.12.003.
21. **Lee I, Kim YO, Park SC, Chun J.** OrthoANI: an improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol, in press. http://dx.doi.org/10.1099/ijsem.0.000760.
22. **Richter M, Rossello-Mora R.** 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A **106:**19126–19131. http://dx.doi.org/10.1073/pnas.0906412106.
23. **Garrity G, Lyons C.** 12 April 2011. Systems and methods for resolving ambiguity between names and entities.US patent 7,925,444.
24. **Garrity GM, Lyons C.** 2003. Future-proofing biological nomenclature. OMICS **7:**31–33. http://dx.doi.org/10.1089/153623103322006562.
25. **Dunn G, Everitt B.** 1982. An introduction to mathematical taxonomy. Cambridge University Press, Cambridge, United Kingdom.
26. **Sokal RR, Sneath PHA.** 1963. Principles of numerical taxonomy. WH Freeman, San Francisco, CA.
27. **Parker CT, Tindall BJ, Garrity GM.** International Code of Nomenclature of Prokaryotes. Int J Syst Evol Microbiol, in press. http://dx.doi.org/10.1099/ijsem.0.000778.
28. **David J, Garrity GM, Greuter W, Hawksworth DL, Jahn R, Kirk PM, Mcneill J, Michel E, Knapp S, Patterson DJ, Tindall BJ, Todd JA, Van Tol J, Turland NJ.** 2012. Biological nomenclature terms for facilitating communication in the naming of organisms. ZooKeys **2012:**67–72. http://dx.doi.org/10.3897/zookeys.192.3347.
29. **Gaston KJ, Mound LA.** 1993. Taxonomy, hypothesis testing and the biodiversity crisis. Proc R Soc B Biol Sci **251:**139–142. http://dx.doi.org/10.1098/rspb.1993.0020.
30. **Denniston C.** 1974. An extension of the probability approach to genetic relationships: one locus. Theor Popul Biol **6:**58–75. http://dx.doi.org/10.1016/0040-5809(74)90031-8.
31. **Skerman VBD, McGowan V, Sneath PHA.** 1980. Approved lists of bacterial names. Int J Syst Evol Microbiol **30:**225–420. http://dx.doi.org/10.1099/00207713-30-1-225.
32. **Sneath PA.** 2005. Bacterial nomenclature, p 83−88. *In* Brenner D, Krieg N, Staley J, Garrity G (ed), Bergey's manual of systematic bacteriology, 2nd ed. Springer, New York, NY.
33. **Tindall BJ.** 1999. Misunderstanding the Bacteriological Code. Int J Syst Bacteriology **49**(Pt 3)**:**1313–1316. http://dx.doi.org/10.1099/00207713-49-3-1313.
34. **Garrity GM, Bell JA, Lilburn T.** 2015. The revised road map to the *Manual*. John Wiley & Sons, Inc., Hoboken, NJ.
35. **Garrity GM, Holt J.** 2001. The road map to the *Manual*, p 119–166. *In* Boone DR, Castenholz RW, Garrity GM (ed), Bergey's manual of systematic bacteriology, 2nd ed. Springer, New York, NY.
36. **Lilburn TG, Garrity GM.** 2004. Exploring prokaryotic taxonomy. Int J Syst Evol Microbiol **54:**7–13. http://dx.doi.org/10.1099/ijs.0.02749-0.
37. **Tukey JW.** 1977. Exploratory data analysis. Addison-Wesley Publishing Co., Reading, MA.
38. **Venables WN, Ripley BD, Venables WN.** 2002. Modern applied statistics with S, 4th ed. Springer, New York, NY.
39. **Krieg N, Garrity G.** 2005. On using the *Manual*, p 15−20. *In* Brenner D, Krieg N, J Staley, Garrity G (ed), Bergey's manual of systematic bacteriology, 2nd ed. Springer, New York, NY.
40. **Garrity GM, Lilburn TG.** 2005. Self-organizing and self-correcting classifications of biological data. Bioinformatics **21:**2309–2314. http://dx.doi.org/10.1093/bioinformatics/bti346.
41. **Stackebrandt E, Ebbers J.** 2006. Taxonomic parameters revisited tarnished gold standards. Microbiol Today **33:**152−155.
42. **Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R.** 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol **12:**635–645. http://dx.doi.org/10.1038/nrmicro3330.
43. **Kim M, Oh HS, Park SC, Chun J.** 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol **64:**346–351. http://dx.doi.org/10.1099/ijs.0.059774-0.



George Garrity is Professor of Microbiology and Molecular Genetics at Michigan State University (MSU). His doctorate is from the University of Pittsburgh Graduate School of Public Health (1980). He was Editor-in-Chief of *Bergey's Manual of Systematic Bacteriology* (1996 to 2006), vice-chair of the Judicial Commission of the International Committee on Systematics of Prokaryotes (ICSP) (2005 to 2008), and chair of the ICSP (2008 to 2014). He is a core member of the Genomic Standards Consortium, founding editor of *Standards in Genomic Sciences*, and a nomenclature/list editor for the *International Journal of Systematic and Evolutionary Microbiology*. He is a Fellow of the American Association for the Advancement of Science and of the Society for Industrial Microbiology (SIMB), a recipient of the van Niel International Prize for Studies in Bacterial Systematics (2011), and President-Elect of the SIMB. He has coauthored >170 publications and 121 taxonomic proposals, and is a coinventor on 75 patent applications/grants. Before joining MSU, he was a scientist in the natural products screening program at Merck & Co. (1981 to 1996). His research interests are in semantics, classification, and visualization of big data and knowledge engineering.