Microbial taxonomy in the post-genomic era: Rebuilding from scratch?

Citation: Thompson, Gilda R. Amaral, Mariana Campeão, Robert A. Edwards, Martin F. Polz, Bas E. Dutilh, David W. Ussery, Tomoo Sawabe, Jean Swings, and Fabiano L. Thompson. "Microbial taxonomy in the post-genomic era: Rebuilding from scratch?." Archives of Microbiology 197:3 (April 2015), pp. 359-370.

As Published: http://dx.doi.org/10.1007/s00203-014-1071-2

Publisher: Springer Berlin Heidelberg

Persistent URL: http://hdl.handle.net/1721.1/104799

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Arch Microbiol (2015) 197:359–370
DOI 10.1007/s00203-014-1071-2

OPINION PAPER

Microbial taxonomy in the post-genomic era: Rebuilding from scratch?

Cristiane C. Thompson · Gilda R. Amaral · Mariana Campeão · Robert A. Edwards · Martin F. Polz · Bas E. Dutilh · David W. Ussery · Tomoo Sawabe · Jean Swings · Fabiano L. Thompson

Received: 25 October 2014 / Revised: 4 December 2014 / Accepted: 5 December 2014 / Published online: 23 December 2014

Abstract Microbial taxonomy should provide adequate descriptions of bacterial, archaeal, and eukaryotic microbial diversity in ecological, clinical, and industrial environments. Its cornerstone, the prokaryote species has been re-evaluated twice. It is time to revisit polyphasic taxonomy, its principles, and its practice, including its underlying pragmatic species concept. Ultimately, we will be able to realize an old dream of our predecessor taxonomists and build a genomic-based microbial taxonomy, using standardized and automated curation of high-quality complete genome sequences as the new gold standard.

Keywords Bacteria · Archaea · Microbes · Taxonomy · Genomics · Evolution · Open access

Communicated by Erko Stackebrandt.

C. C. Thompson (✉) · G. R. Amaral · M. Campeão · R. A. Edwards · B. E. Dutilh · J. Swings · F. L. Thompson
Laboratory of Microbiology, Institute of Biology, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil
e-mail: thompsoncristiane@gmail.com

R. A. Edwards
San Diego State University, 5500 Campanile Dr, San Diego, CA 92182, USA

R. A. Edwards
Argonne National Laboratory, 977 S. Cass Ave, Argonne, IL 60439, USA

M. F. Polz
Massachusetts Institute of Technology, Cambridge, MA, USA

B. E. Dutilh
Radboud University, Nijmegen, The Netherlands

D. W. Ussery
BioSciences Division, Oak Ridge National Labs, Oak Ridge, TN, USA

T. Sawabe
Laboratory of Microbiology, Faculty of Fisheries Sciences, Hokkaido University, Hakodate, Japan

J. Swings
Laboratorium voor Microbiologie, Ghent University, K. L. Ledeganckstraat 35, 9000 Ghent, Belgium

F. L. Thompson
SAGE-COPPE-UFRJ, CT2 Rua Moniz de Aragão, no. 360 - Bloco 2, Rio de Janeiro, Brazil

Short history of microbial taxonomy

The history of microbial taxonomy during the last 100 years is one of a scientific field in which progress and conservatism meet. It is progressive as it incorporates the most advanced technologies, yet conservative because it adheres to standards and rules. A number of technological driving forces were operative during its development as a scientific discipline: the introduction of metabolic and phenotypic characterization of bacteria, numerical analysis of phenotypic data (Sneath and Sokal 1973), DNA–DNA hybridizations (DDH) and %G+C determinations (De Ley 1970), 16S rRNA gene sequencing (Woese and Fox 1977), and multilocus sequence analysis (MLSA) (Gevers et al. 2005) before the introduction of complete genome sequencing (Coenye and Vandamme 2003; Coenye et al. 2005; Thompson et al. 2009). The seminal work of Carl Woese in 1977 on the

discovery of the three domains of life, triggered by the use of ribosomal rRNAs as evolutionary chronometers, was a new paradigm in microbial taxonomy (Woese and Fox 1977). Today, within the so-called polyphasic taxonomic approach, all schemes include measurements of evolutionary relationships using gene sequences (most notably the 16S rRNA gene) (Yarza et al. 2008) to determine the phylogenetic position of an isolate, combined with chemotaxonomic, physiological, and cultural properties (Colwell 1970; Stackebrandt et al. 2002). Polyphasic taxonomy is based on the phylogenetic framework. A comprehensive practical guide to polyphasic taxonomy has been published by Tindall et al. (2010) who state that novel taxa should be characterized as comprehensively as possible.

Polyphasic microbial taxonomy is recognized as an orthodox field meaning that the following fixed rules are applied for species delineation: (a) DDH values of at least 70 %; (b) at least 97 % rRNA gene sequence similarity (recently 98.7 % was proposed by Stackebrandt and Ebers (2006); (c) maximum 2 % of G+C span; and (d) differentiating chemotaxonomic and phenotypic features where great weight is placed on the phenotypic (chemotaxonomic) characterization using specialized technologies such as fatty acid methyl ester (FAME), polyamines, peptidoglycan types sphingolipids, and matrix-assisted laser desorption/ionization—time-of-flight mass spectrometer (MALDI–TOF MS); however, in most cases, these are not very useful for discriminating all species, i.e., all species in each major lineage, nor do they shed light on the biology of the microorganisms. Rosselló-Móra (2012) has given a comprehensive state of the art on microbial taxonomy, its principles, practice, and most recent developments (Rosselló-Móra 2012). This author favors the application of genome sequences information in microbial taxonomy.

We contend that current rules are impeding progress both in the description of new species and in the development of taxonomy as a scientific discipline. First, DDH is still considered a gold standard for species delineation in spite of demonstration that other techniques such as multilocus sequence analysis (MLSA) average aminoacid identity (AAI), and genome-to-genome distance (GGD) are portable and have greater discriminatory power (see, e.g., Gevers et al. 2005; Konstantinidis and Tiedje 2005; Auch et al. 2010). In fact, many journals specializing in taxonomy have not yet accepted alternative techniques to DDH. Moreover, because of technological and methodological hurdles, DDH is only performed by few laboratories that are highly specialized in taxonomy (Wayne et al. 1987; Stackebrandt and Ebers 2006) and performing DDH experiments might take years, slow description of new species considerably. Finally, journals such as systematic and applied microbiology (SAM) and International Journal of Systematic and Evolutionary Microbiology (IJSEM) require the concomitant extensive phenotypic characterization of closely related type strains every time a new species is being described (journals.elsevier.com/systematic-and-applied-microbiology/, ijs.sgmjournals.org/) even if data are available for the same type strains using the same reagents and machines in the same laboratories. Because this is time-consuming and unnecessary, it likely keeps many scientists from formally describing new microbial species. That taxonomy is a conservative science is not a new observation, and it is interesting to note that it took two decades for the acceptance of DNA–DNA hybridization as a reliable standard. However, we do not mean to imply that all rules should be overturned. In fact, deposition of strains in public collections and sequences in public databases must continue no matter what new taxonomic schemes are agreed on.

The failure of polyphasic taxonomy

Let us first state that, in the past, application of polyphasic taxonomy has enabled considerable progress and stability in microbial taxonomy and its nomenclatural legacy will be safeguarded. However, the "gold standards" of polyphasic taxonomy are increasingly outdated since orthodox microbial polyphasic taxonomy is neither able to keep up with the progress in environmental and evolutionary microbiology nor with the needs of clinical microbiologists and epidemiologists. Additionally, there is a mounting uneasiness with the definition of the microbial species itself (including bacteria and archaea).

Polyphasic taxonomy cannot keep up with the explosion in genome sequences, even at the broadest levels of taxonomic classification; at the time of writing, there are about 200 bacterial genomes in GenBank where the phyla are listed as 'unclassified.' Further, as tens of thousands of genomes are becoming available, the diversity within a species—much of which arises due to recombination between lineages—has led to the proposal of 'fuzzy species' (Fraser et al. 2007; Hanage 2013). Much of the recent progress in microbiology is due to the dramatic plunge in sequencing cost and speed; currently, sequencing a hundred small bacterial genomes at 10× coverage is <$10—that is, literally a few cents per genome, and third-generation sequencing methodology allows completion of a bacterial genome in a few hours. Additional costs related to, e.g., DNA extraction and library preparation, and bioinformatics (computer time), need to be taken into consideration, but will not undermine the use of genomes in species descriptions. In spite of the potential of genome sequences, only very few studies have applied genome sequences to date for new species descriptions (Moreira et al. 2014a, b).

Senior scientists who contributed to polyphasic taxonomy realize that the principles and practices of present-day polyphasic taxonomy should be questioned and microbial taxonomy be rethought. The role of microbial taxonomy is to provide a framework for reliable identification of organisms in order to learn about their functional role in a particular environment. The need to revisit polyphasic taxonomy has been articulated by Vandamme and Peeters (2014) using the taxonomy of the *Burkholderia* complex as an example. These authors state that "DDH had been historically introduced to approach whole-genome sequence (WGS)-derived information as closely as possible (Wayne et al. 1987) and now that we have direct access to WGS information, we want it to mimic the results obtained through (physical–chemical) DDH experiments." This statement exemplifies the paradox of keeping DDH as a standard where attempts are being made to translate the old DDH species threshold into new WGS-based thresholds even though the information derived from the latter techniques is superior to DDH.

Approximately 600 new bacterial and archaeal species are described each year using polyphasic taxonomy (Konstantinidis and Stackebrandt 2013), and at such pace, it will take centuries to describe even a small fraction of the novel species present in the biosphere. It is therefore clear that on purely pragmatic grounds, we can no longer proceed with the present-day orthodox polyphasic microbial taxonomy as defined by the comprehensive guidelines. We can no longer be "keeping bacterial taxonomy as the playground of a few privileged with full access to a battery of phenotypic, genotypic and chemotaxonomic tools" (Vandamme and Peeters 2014).

Another reason for the failure of polyphasic taxonomy is that the standards for species descriptions are still based on aged approaches that are not appropriate for many of the species that are currently being described. Many of the tests herald from medical microbiology introduced in the late nineteenth and early twentieth century but are still applied to environmental isolates. Among the many examples that might be given to illustrate that polyphasic taxonomy is failing in the description of biodiversity are the cases of *Burkholderia* (Vandamme and Peeters 2014), *Wolbachia* (Ellegaard et al. 2013), and *Pseudomonas* (Alvarez-Pérez et al. 2013). For *Wolbachia*, two genetically distinct and irreversibly separated clades were distinguished (Ellegaard et al. 2013), but these cannot be described as species. One of the biggest problems is that polyphasic taxonomy is unable to deal with uncultivated microbes. In the ubiquitous SAR 11 (*Pelagibacter*) clade, which is thought to be the most abundant bacterial group in the world's oceans, a number of phylotypes are recognized but poorly characterized by cultivation (Giovannoni et al. 2005; Brown et al. 2012). The most extreme case is new biodiversity described by single-cell sequencing to generate reference genomes of uncultured taxa from the marine bacterioplankton, e.g., two uncultured flavobacteria described by Woyke et al. (2009). In these and many other cases, polyphasic taxonomy is of little help in describing novelty. The prokaryotic code should include the description of uncultured organisms based on whole-genome sequences, particularly now with the advent of new technologies such as the single-cell genomics.

Whereas monoculture experimental standards and rules have guided the description of bacterial and archaeal species in the past, several colleagues have stressed that the time has come to integrate genomics as a reliable and reproducible standard into the taxonomy of the bacteria and archaea (Lan and Reeves 2000; Doolittle and Papke 2006; Fraser et al. 2009; Whitman 2009; Staley 2009; Klenk and Göker 2010; Zhi et al. 2012; Ellegaard et al. 2013; Chun and Rainey 2014). However, simply incorporating genome sequence data into polyphasic taxonomy as proposed by Ramasamy et al. (2014) might not be sufficient. Indeed, adding genome sequences to the list of key elements defined by Tindall et al. (2010) will not rejuvenate microbial taxonomy. We believe that taxonomists share together with ecologists and phylogenists the responsibility for a description of the microbial world. In fact, with the available genomic technology and sufficient metadata, we can construct the necessary standards and rules to develop robust and fast tools that describe and order microbial diversity.

The re-examination of the microbial species definition

A further and more fundamental failure of polyphasic taxonomy is that it uses a very broad species definition that is not based on an evolutionary species concept (Stackebrandt et al. 2002; Fraser et al. 2009). Recent progress in environmental microbiology has shown that classically described species often comprise assemblages of ecologically and genomically distinct populations. In fact, a species cutoff of 70 % as used in DDH leads to underspeciation within prokaryotes. The 16S rRNA gene on the other hand lacks resolution at the species level, even at the 98.7 % level. Universal cutoff levels to delineate species do not make sense since speciation is a dynamic process leading to sister taxa that is separated by variable sequence space (Shapiro and Polz 2014).

Although there is currently no consensus on a species concept for bacteria and archaea (cohan 2001; Rosselló-Mora and Amann 2001; de Queiroz 2005; Dykhuizen 2005; Nesbø et al. 2006; Staley 2006; Fraser et al. 2007; Achtman and Wagner 2008), taxonomy may nonetheless benefit from an evolutionary framework to order bacterial, archaeal,

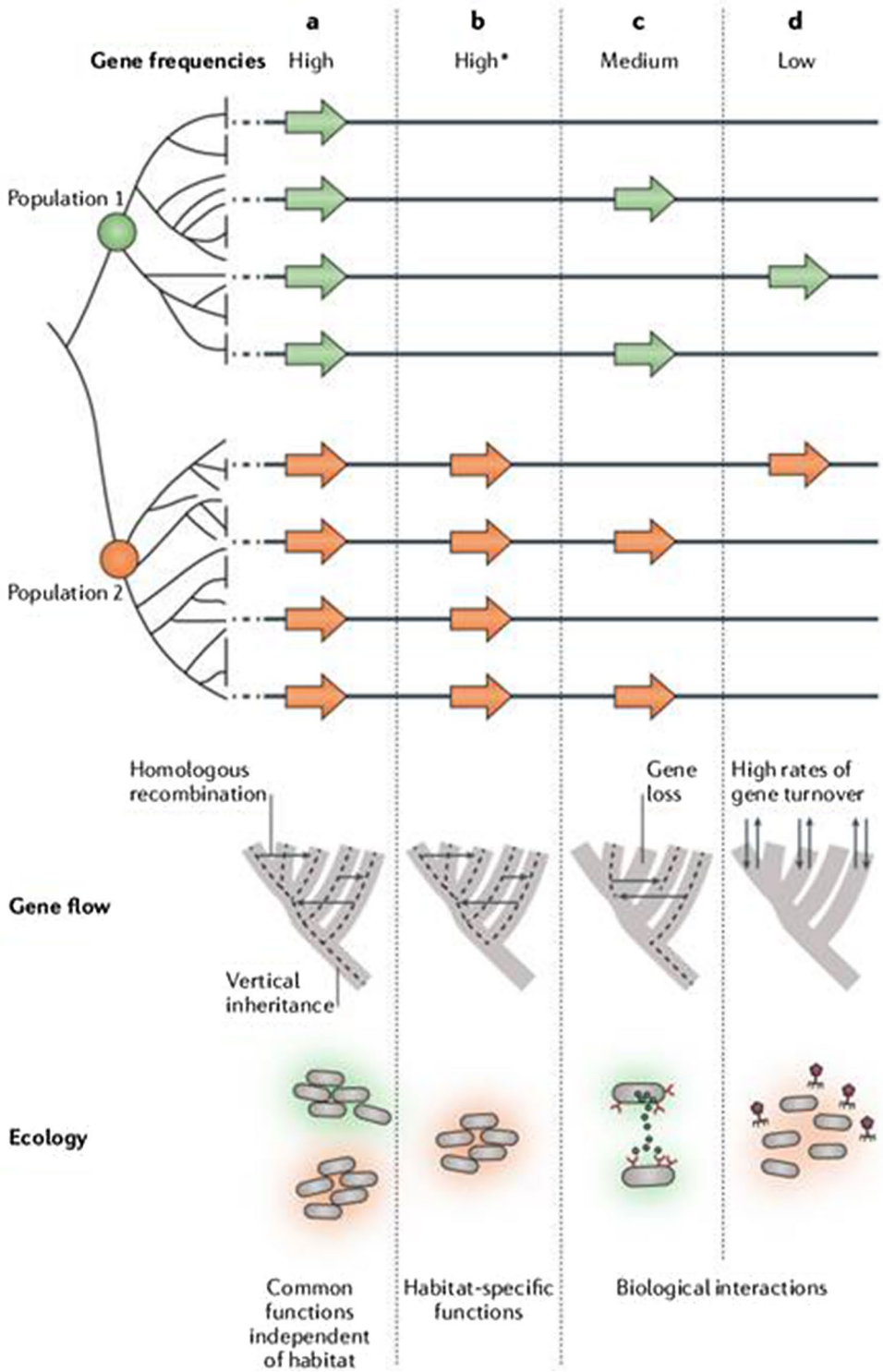
and eukaryotic microbial diversity into more natural units (Fraser et al. 2009). This framework has been provided by WGS, which allows identification of sequence clusters at high genotypic resolution based on variation in protein-coding genes distributed across the genomes. Importantly, such clusters are consistent with the vernacular notion of a species as a group of organisms that is more similar to each other than to any other species (Polz et al. 2006; Fraser et al. 2009). The discovery of clusters also offers a practicable solution to the species dilemma, i.e., to sidestep it for the moment and to continue with the pragmatic definition of species that emphasizes the existence and description of clusters of coexisting strains that are consistently similar on a genetic and phenotypic basis. This approach is not so much different from the present one, just shifting emphasis to molecular data. Such clusters may be defined from multiple different sources of genetic data (core gene sequences, microarrays or whole genomes) and form tractable units to address evolutionary and ecological questions.

The focus on sequence (phylogenetic) clusters as more natural units of organization for bacteria, archaea, and eukaryotic microbes is motivated by the following considerations. First, analyses of environmental isolates and metagenomes have shown that microbial communities consist of genotypic clusters of closely related organisms, with mounting evidence that these clusters display cohesive environmental associations and dynamics that differentiate them from other such clusters coexisting in the same samples (Hunt et al. 2008; Konstantinidis and DeLong 2008; Denef et al. 2010; Caro-Quintero and Konstantinidis 2012; Kashtan et al. 2014), and recent work has shown that it is possible to construct genomic backbone scaffolds for several hundred 'species' from a series of metagenomic samples and to then use this for template-based assembly of genomes from individual samples (Nielsen et al. 2014; Mick and Sorek 2014). Second, recent modeling and whole-genome analysis of clusters in the very early stages of divergence has suggested that, in spite of potential for horizontal gene transfer (HGT), selection is required for cluster formation in sympatry (recently reviewed in Polz et al. 2013; Shapiro and Polz 2014). But even if clusters form in allopatry, they are free to diverge ecologically because specific alleles or genes can spread in a population (i.e., location)-specific manner, as seen in individuals from large-scale metagenomic studies (Nielsen et al. 2014).

A fact that any species definition has to contend with is that bacteria and archaea can share genes across any species boundary imposed by taxonomists via HGT (Doolittle and Zhaxybayeva 2009). At face value, this violates the biological species concept as formulated by Mayr (1942). However, there is mounting evidence that many eukaryotes speciate by hybridization and that such events occur frequently (but have somewhat low probability of survival) (Mallet 2008). Moreover, recent population genomic analyses of clusters in the early stages of divergence have shown that although HGT occurs frequently, gene flow discontinuities exist between clusters even if they remain closely related (Cadillo-Quiroz et al. 2012; Shapiro et al. 2012). At least in one case, it was also demonstrated that these gene flow discontinuities are sufficient for adaptive alleles and genes to spread in a cluster-specific manner (Shapiro et al. 2012). It is important to realize that speciation events can be transient and need not necessarily lead to species (Mallet 2008; Wiedenbeck and Cohan 2011; Shapiro and Polz 2014). Hence, it will be a challenge for microbial taxonomists to delineate species that appear to have at least some permanence in the evolutionary spectrum.

A more natural definition of microbial species as proposed in the present text also solves the problem of the frequent observation that even closely related genomes can have high gene content variation that gives rise to at least some level of phenotypic variation. If, as argued above, clusters are gene flow units within which selection acts on gene frequencies, then it is possible that gene content variation arises due to frequency-dependent selection where the fitness of a genotype within a population depends on its frequency (Fig. 1; Cordero and Polz 2014). In fact, genes at low and intermediate frequency may be involved in niche complementarity, social interactions and predator–prey interactions (Cordero and Polz 2014). It has been argued previously that many genes occurring at low frequency within genomes are involved in predation evasion by varying surface antigenicity (Rodriguez-Valera et al. 2009; Cordero and Polz 2014).

Fig. 1 Gene frequencies and the evolutionary and ecological processes, extracted from Cordero and Polz (2014). Populations are recognized as genotypic clusters separated by gene flow boundaries and can have distinct habitats. **a** High-frequency genes (*green* and *orange arrows*; also represented by *short black lines* in the gene flow map) are primarily maintained by vertical inheritance and homologous recombination. These genes are observed across multiple ecological populations and typically encode core metabolic and housekeeping functions that are independent of the different environments. **b** High-frequency genes (High*) can also segregate ecological populations. After being gained or lost in a population-specific manner, these genes could follow similar patterns of gene flow as other core genes. They are potentially involved in habitat-specific functions (for example, the adaptation to use either the *orange* or *green* substrates as a nutrient source). **c** Medium-frequency genes flow by vertical inheritance, homologous recombination, and gene loss. As illustrated in the figure, without considering population structure (in other words, that the *green* and *orange* genes are derived from two distinct populations), the frequency of these genes would be indistinguishable from that of the High* genes (50 %). Recent studies suggest that some of these genes might be involved in local biological interactions (such as those that are mediated by public goods), which create frequency-dependent selection. **d** Low-frequency genes reflect extremely high rates of gene turnover, which represents an evolutionary strategy to diversify, often precipitated by negative frequency-dependent selection emerging from interactions with predators (such as phage) or with the immune system (color figure online)



Moreover, intermediate frequency genes may be involved in frequency-dependent interactions such as public good production and cheating as well as niche-complementation (Cordero and Polz 2014). This may also explain some phenotypic variation frequently observed among closely related genotypes. In the context of taxonomy, it will be important to recognize that some traits may be patchily distributed within a population. For example, any excreted enzyme may act as a public good and invite cheating within the same population or species (Cordero et al. 2012). Phenotypic variation among strains of the same species is a well-known example of possible cheating (Moreira et al. 2014a, b).

In summary, although we do not have an agreed upon species definition for bacteria and archaea, we propose that genotypic (phylogenetic) clusters can serve to easily and quickly formulate hypotheses of species (or populations). The properties of these units can then be further explored by genomics as outlined in the next section. However, we also note that it is often not easy to recognize the exact boundaries of clusters. This is because the extensive history of gene transfer may create "fuzzy" boundaries and nested structure of clusters when, as is typically the practice, analyzing phylogenetic structure using trees of concatenated genes (or genomes) (Hanage et al. 2005; Hanage 2013). A challenge for the future will therefore be to develop robust techniques that, we believe, should be based on analysis of patterns of contemporary gene flow rather than sequence similarity-based clustering.

Paradigm shift

Taxonomy must adjust to the genomics era, addressing the needs of its users in microbial ecology and clinical microbiology (Preheim et al. 2011), in a new paradigm of open-access genomic taxonomy (Thompson et al. 2013a). We witness already the tremendous efforts put into initiatives on prokaryote genomics, such as the Genomic Encyclopedia of Bacteria and Archaeae—GEBA (Wu et al. 2009; Klenk and Göker 2010), Genomes OnLine Database—GOLD (Kyrpides 1999; Pagani et al. 2012), and the Integrated Microbial Genomes—IMG (Markowitz et al. 2006, 2014).

Whereas the actual divorce between classical taxonomy, evolution, and ecology is hampering progress, the new paradigm of genomic taxonomy provides rapid diagnostics of microbial phenotypes and niches in an open-access manner. The open-access genomic taxonomy embraces the classification of species builds on many established genomic tools. Examples include genome signatures (e.g., genome-to-genome distance (GGD); Auch et al. 2010), average amino acid identity (AAI) (Rohwer and Edwards 2002), average nucleotide identity (ANI) (Konstantinidis and Tiedje 2005), Karlin genomic signature (Karlin and Burge 1995), supertrees analysis (Brown et al. 2001), codon usage bias (Wright 1990), metabolic pathway content, core-genome analysis, pan genome family trees (Snipen and Ussery 2010), and in silico proteome analysis, genotype-to-phenotype-to-genotype-derived metabolic features, including those features that may inform ecology (e.g., host–microbe interactions, and energy/nutrient cycling) and evolution (Dutilh et al. 2013, 2014; Amaral et al. 2014). Only recently species descriptions have began to include some measurements of genome-derived measurements of genetic relatedness based on, e.g., AAI/ANI, always with supporting DDH data, indicating that genomic taxonomy is not yet recognized by major journals as standards in species descriptions. Also none of these methods are included in minimal standards of species description. It also embraces the identification of strains based on diagnostic features disclosed in the new species descriptions. The application of genomic taxonomy is providing a predictive operational framework for reliable identification and classification. We argue for an open-access catalog of taxonomic descriptions with prototypes, diagnostic tables, links to culture collections, to genome and gene sequences, and to other phenotypic and ecological databases. Ideally, the open-access taxonomy is based solely on genome sequences that allow both the phylogenetic allocation of new strains and species in the taxonomic space and the phenotypic/metabolic characterization in open online databases.

A new species description needs to be based, first of all, on at least one complete genome (Thompson et al. 2013a). In this way, the genomic landscape of the novel bacterium becomes available to microbiologists. Ideally, additional representative genomes of strains belonging to the new species will be included in order to provide information on the intraspecies genomic and phenotypic variation. The species description process needs to be automated and openly available to all, i.e., open access. Genomic taxonomy has already been successfully applied as an alternative for the more traditional species description and re-classification (Thompson et al. 2009; Haley et al. 2010; Thompson et al. 2011a, 2013b; Moreira et al. 2014a, b). For example, the genus *Listonella* was reclassified as a later heterotypic synonym of the genus *Vibrio* (Thompson et al. 2011b), and a new taxonomic framework for the genus *Prochlorococcus* was proposed with the descriptions of new species (Thompson et al. 2013c).

The genome sequence of the new taxa can be used for automatic identification of a microbial species through open-access tools available in a web-based portal. The genome sequences can also allow for the rapid identification of major phenotypic features associated with that organism, and translation of genomic information into phenotype will be increasingly precise with more genomes being annotated. We argue that ultimately the analyses of genes coding for the specific proteins involved in the metabolic pathways responsible for diagnostic features (e.g., Voges–Proskauer reaction, indole production, arginine dihydrolase, ornithine decarboxylase, utilization of myo-inositol, sucrose and L-leucine, and fermentation of D-mannitol, D-sorbitol, L-arabinose, trehalose, cellobiose, D-mannose and D-galactose) may be an alternative to the time-consuming phenotypic characterization using the standard biochemical tests (Karp et al. 2005; Romero et al. 2005; Dutilh et al. 2013; Amaral et al. 2014). Diagnostic phenotypic data are very hard to retrieve and lack portability (see, e.g., Bergey's Manual, The Prokaryotes).

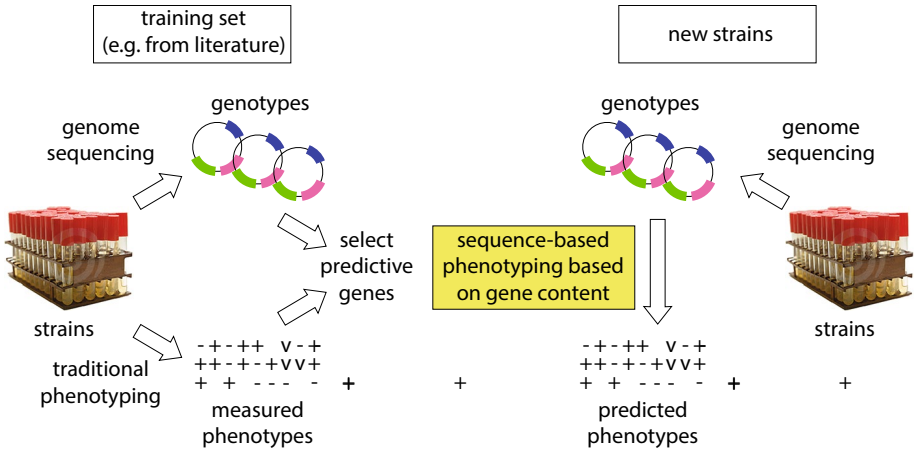


Fig. 2 Genotype-to-phenotype approach in genomic prokaryotic taxonomy. A training set (type and reference strains) is subjected to whole-genome sequencing and gene content (including genes coding for the specific enzymes of a given metabolic pathway and the regulator proteins) analysis. Measured phenotypic features of the training set are obtained from the literature (e.g., Bergey's manual and the Prokaryotes) and compared with the gene content in order to predict phenotypes. The phenotype of new strains is obtained by whole-genome sequencing using the diagnostic gene content defined in the training set (color figure online)

Huge amount of valuable phenotypic data are simply out of reach because they are available only in the species description papers, manuals or handbooks. On the other hand, researchers need electronic portable data in order to push forward different fields of microbiology. By using the genotype-to-phenotype strategy (Fig. 2), it will be possible to leverage genome information to overcome this serious shortcoming of current microbial taxonomy in microbial ecology and clinical microbiology. In addition, relating the wealth of resource-associated data to cultures deposited in microbial Biological Resource Centers will foster academic research and drive innovation in the bio-economy.

The manner in which phenotypic information is retrieved and presented in new species description and identification schemes will need to change in order to allow for open access of taxonomic data. Metabolic data are of paramount importance to link genomes and phenotype, but data accessibility needs also to be considered. We foresee two quite distinct situations in the process of open-access genomic taxonomy targeting biodiversity characterization. In the case of totally new taxa belonging to, e.g., a new phylum or class, for which metabolic data are scarce or even unknown and the genomic landscape is poorly known, significant efforts will be needed in order to provide experimental in vitro work to underpin the new descriptions. Of course, these are the most interesting cases in the context of biodiversity discovery. On the other hand, in cases of a new species description within a well-studied phylum (e.g., *Proteobacteria* and *Firmicutes*), phenotypes may be readily obtained by the genotype-to-phenotype-to-genotype approach. In this context, former microbial taxonomy studies (and the enormous phenotypic information available) performed in the last century will underpin the genotype-to-phenotype-to-genotype strategy. Efforts will be required to implement database-based high-throughput phenotypic methods that provide portable open-access data (Fig. 2). Methods of particular interest are those that reveal the amino acid sequences and the structure of molecules (e.g., secondary metabolites products, virulence factors). In comparison with phenotypic methods, the major advantage of new generation sequencing is the high throughput, relatively low cost, high information content, high data quality, and portability of data. The data can be easily checked for quality in different stages of taxa description and can also be deposited in public databases. International initiatives such as GEBA are already working on this goal by whole-genome sequencing of all type strains of known species (more than eleven thousand genomes) and by innovative technologies such as single-cell genomics of uncultured microbes for discovery of new biodiversity (Wu et al. 2009; Rinke et al. 2013). In spite of these large ongoing initiatives, most of the current species descriptions in the major specialized journals still use the polyphasic approach, because of the insistence that in vitro DDH and massive phenotyping remain the cornerstones of contemporary Microbial taxonomy. The majority of the known type and reference strains still have no genome sequence. Vandamme and Peeters (2014) have proposed a species description based on the full genome sequence and a minimal description of phenotypic characteristics, to be considered sufficient, cost-effective, and appropriate. The importance of increasing the rate of species descriptions is exemplified by the pace at which microbiome projects are advancing the study of culture-independent biodiversity of the most diverse environments and hosts which leads to the generation of Terabytes of DNA sequence in a matter

of days (Huang et al. 2014; Franzosa et al. 2014; Li et al. 2014; Nielsen et al. 2014). As the ongoing microbiome projects advance, there will be a growing gap between the field of microbial community diversity and Microbial taxonomy. We argue that the open-access genomic taxonomy can help to close this gap by establishing a stable, reproducible, and informative framework. Taxonomy needs also to be affordable. The cost for a new species description based on the genome sequences will be considerably less expensive and quicker than based on the polyphasic taxonomy.

In silico phenotyping

To distinguish different strains within a bacterial species, or different species within a genus, the field of bacterial taxonomic classification has developed sets of phenotypic tests. Examples of phenotypes that may be measured include metabolism of specific organic compounds, resistance to antibiotics, phage sensitivity. Specific phenotypic tests suitable for classification can be developed for each taxonomic group. Because microbial phenotypes are the result of metabolic pathways or functions encoded on the genomes of the bacterial strains, the phenotype is a proxy for phylogenetic classification.

In the past decade, great advances have been made in DNA-sequencing technologies. Several competing companies now provide the necessary equipment and chemistry to obtain high-quality draft genome sequences of bacterial strains at affordable prices. Third-generation sequencing will soon allow for sequencing of bacterial genomes in a few hours for a few dollars (Didelot et al. 2012). These genomes contain a wealth of genetic information and enable direct classification with respect to all other sequenced genomes, i.e., without the use of a phenotype as a proxy. Moreover, bioinformatic advances now enable mining of these genome sequences to predict the phenotype of the sequenced strain, known as in silico phenotyping, avoiding costly experimental phenotypic screens that need to be performed in the laboratory. We have recently proposed an approach for in silico genomic phenotyping based on gene content screens (Fig. 2) (Amaral et al. 2014). In this study, genes involved in the molecular pathways leading to the phenotypes were selected and genome sequences screened for the presence of these genes. This allowed us to confidently predict phenotypic classifications to each of the genomes (Amaral et al. 2014) that can be tested experimentally. A large collection of phenotypes and the associated genes is contained in the SEED database (Overbeek et al. 2014). This database contains hundreds of expert-annotated, manually curated subsystems that can be rapidly projected onto new genome sequences, providing an automated approach for in silico prediction of phenotypes.

Identifying or predicting the genes that are involved in each phenotype is known as gene-trait matching. Recently, a complete in silico pipeline was outlined for the consistent annotation of bacterial genomes followed by automated gene-trait matching (Dutilh et al. 2013). Condition for this approach is that the trait is consistently measured for all sequenced genomes. By using this approach-dubbed "genome-wide association study for microbes" (GWAS-M), candidate genes contributing to the trait can be obtained. The approach employs a machine-learning tool, and by analyzing a training set of bacteria that differ with respect to the trait, it identifies which genomic variables best explain the trait variation. These genomic variables can then be used to infer the phenotype of a strain based on its genome sequence.

Advances in genome sequencing fuel the young field of bioinformatic gene-trait matching, and a few applications have been published thus far. An early example of this approach was based on a comparative genome hybridization (CGH) array, and involved the identification of genes associated to growth on sugars and nitrogen dioxide production in *Lactobacillus plantarum* (Bayjanov et al. 2012). More recently, a large collection of 274 *Vibrio cholerae* genomes was mined for genomic variables that explained not phenotypes, but the occurrence of the isolates in three niche dimensions, including space, time, and habitat (Dutilh et al. 2014). This study revealed that mobile genetic elements explained most of the variation in all these niche dimensions and may be used to classify the genomes. These examples illustrate the versatility of gene-trait matching and its power for identifying genes associated with specific bacterial traits.

Genome sequencing is not without its drawbacks. Next-generation or 'second-generation' sequencing has removed many of the biases of cloning that plagued earlier genome sequences, but whole-genome assembly is often complicated by short reads and the miriad of repeat regions in the microbial genome. Ribosomal RNA operons are frequently present in multiple exact copies, and phage genes, transposons, and insertion elements all contribution to computational confusion during the assembly process. Finishing genomes completely—so that every base is known and error free—is both expensive and time-consuming, typically requiring PCR walking across repeat regions. Consequently, most microbial genomes are only sequenced to "high-quality draft status" typically meaning <100 contigs. Third-generation sequencing technologies, such as Pacific Biosciences and Oxford Nanopore, have the advantage of long reads (10,000 bp or longer), although currently their throughput and base calling accuracy is lower than the second-generation machines. However, many bacterial genomes have been sequenced and assembled with a single run on Pacific Biosciences

machines (Doi et al. 2014; Forde et al. 2014; Shiwa et al. 2014).

Genome annotation is generally based on similarity between predicted proteins in the genome and annotated proteins in the database. Of course, similarity-based annotation systems require a homolog of the predicted protein be known. Ideally, protein functions should be experimentally verified, but the function of very few proteins has been confirmed in the laboratory. Automated genome annotation therefore is susceptible to errors from missing information.

Genes coding for the proteins responsible for diagnostic phenotypic features can be retrieved using the RAST program and the KEGG metabolic database (http://www.genome.jp/kegg/). The BLASTP algorithm can then be used to identify genes associated with the biochemical pathways. The program ExPASy translate (ExPASy Bioinformatics Resource Portal) was used to analyse protein sequences. To automate searches for genes related to phenotypes of interest, specific programs and databases related to different taxonomic groups will need to be developed (73). For instance, amino acid FASTA files with coding sequences of a target phenotypic feature can be used as input in order to verify whether hits are found for the gene (enzyme) being searched in a specific database. Orthologs genes will have the greater BLAST scores and identity will be >40 % in this type of search. Gene sequence length normally needs to be >70 % of the query length. After these steps, if all the genes (enzymes) involved in a metabolic pathway are present in the genome, the organism is considered positive for a given phenotype, or if one or more genes (enzymes) in a metabolic pathway are absent, the organism is considered negative. It is also important to evaluate regulatory genes, global regulators of the different diagnostic phenotypic features/metabolic pathways, presence of indels in the gene sequences, sRNA regulation, and promoter sequences.

Despite sources of error (e.g., incomplete DNA sequencing and inaccurate annotations), our knowledge of microbial metabolism encoded in the databases is thorough. For example, in a recent study, we sequenced the genome of *Citrobacter sedlakii*, a previously unsequenced organism. At 320 contigs, our assembly was low-quality draft, but using Rapid Annotation using Subsystem Technology—RAST (Aziz et al. 2008), we annotated 1,399 reactions performed by enzyme complexes encoded in the genome. Only five genes were missed due to low sequencing coverage, and six genes were missed due to problems with the assembly and annotation (but we present in the genome upon further inspection; Cuevas et al. in preparation). This suggests that even genomes with a relatively low sequence coverage can be used to predict the metabolism that an organism performs which can then be used in taxonomic assignments.

Statements arguing in favor of a genomic microbial taxonomy

- Microbial taxonomy is moving from polyphasic taxonomy into a new open-access genomic microbial taxonomy with a set of standardizes tools used on a genome sequence. Mere translation of thresholds of polyphasic taxonomy will not contribute to it (Kämpfer and Glaeser 2012; Vandamme and Peeters 2014).
- The highest priority of a rejuvenated genomic Microbial taxonomy is to help describe better microbial diversity and to serve better the medical and environmental microbiologists and epidemiologists.
- As scientists, it is our duty to question the basis of taxonomy, both theory and practice, as well as the validity of the schemes that we produce. Incorporating ecological, phylogenetic, and evolutionary dimensions is needed to define a biologically coherent species concept. Re-establishing the link between phylogenetics and taxonomy will allow a better understanding of microbial speciation (Zhi et al. 2012).
- It will take time to develop a new coherent prokaryote species concept. A rush for a new species concept is not needed and would be counterproductive. International meetings on the topic might help to open up the discussion. Fortunately, we have the chance to welcome newcomers in the field, such as computer scientists, microbial ecologists, and evolutionary microbiologists. Microbial taxonomy seems to be in excellent shape, particularly in the Asian countries (Tamames and Rosselló-Móra 2012). The challenge now at stake for genomic Microbial taxonomy is to examine how the existing genomic databases, bioinformatics tools, and access facilities may be further developed into prototypes to be further tested and discussed. Automated methods such as the ones benchmarked by Larsen et al. (2014) will enable the use of WGS for higher resolution and more phylogenetically accurate classifications. It has been noticed that to date, microbial taxonomy has barely taken the wealth of information contained in completed sequenced genomes into account (Klenk and Göker 2010). It allows to incorporate taxonomy and typing in a high-resolution, reproducible, and portable scheme. The developments are expected to take place in parallel with the ongoing conservative practice of polyphasic taxonomy.
- We propose the following general steps as a roadmap for species description within known genera: First, perform whole-genome sequence of the novel type and reference strains and calculate genome similarity within species and toward the closest known species by means of MLSA, GGD, and AAI; second, check in the published literature (i.e., species descriptions, Bergeys Manual, and The

Prokaryotes) the list of useful discriminatory phenotypic features to be searched for in the genome sequences; third, apply the genotype-to-phenotype approach and define the presence of diagnostic phenotypes on the basis of the presence of the gene sequences, trying to obtain the maximum number of phenotypes based on genome sequences; fourth, perform the most basic phenotypic characterization of the novel strains in vitro, such as cell and colony morphology, growth at different ranges of temperature, pH, and salinity. Avoid doing, e.g., FAME, MALDI–TOF, AFLP, and other non-portable fingerprinting techniques; fifth, deposit the genome sequences of the novel type and reference strains in public open-access databases and the cultures in public collections; sixth, write concise text reporting the major findings obtained in the steps 1–5, in a manner that can be readily assessable by machines. Automation in the production of texts dealing with descriptions and updates of databases will be a plausible development. Analytical work and bioinformatics are also needed in order to use phenotypic information available in genome sequences. The new system clearly needs new tools to gain information from the genotype to the phenotype and back to the genotype.

- Specialized journals, e.g., IJSEM and SAM are starting to get involved in an open scientific discussion on Microbial genomic taxonomy and offer a tribune for it (Sen et al. 2014; Chun and Rainey 2014; Ramasamy et al. 2014). This will attract bright young scientists, needed for the remolding the theory and practice of genomic microbial taxonomy.
- It is necessary to emphasize that novel strains or strains with novel properties should be deposited in public collections (Stackebrandt et al. 2014). Genome databases are sadly full of sequences without a deposited culture in a recognized Culture Collection (Tamames and Rosselló-Móra 2012).

Acknowledgments We thank CAPES, CNPq, FAPERJ, and NSF for funding. R.E. is supported by NSF Grants DEB-1046413 and CNS-1305112. M.F.P. acknowledges funding by NSF Grants DEB 0918333 and OCE 1441943, and the Gordon and Betty Moore Foundation. D.U is supported by internal funding from Oak Ridge National Labs, and from grants from the Office of Biological and Environmental Research in the DOE Office of Science.

References

Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6:431–440. doi:10.1038/nrmicro1872

Alvarez-Pérez S, de Vega C, Herrera CM (2013) Multilocus sequence analysis of nectar pseudomonads reveals high genetic diversity and contrasting recombination patterns. PLoS One 8:e75797. doi:10.1371/journal.pone.0075797

Amaral GRS, Dias GM, Wellington-Oguri M et al (2014) Genotype to phenotype: identification of diagnostic vibrio phenotypes using whole genome sequences. Int J Syst Evol Microbiol 64:357–365. doi:10.1099/ijs.0.057927-0

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. doi:10.4056/sigs.531120

Aziz RK, Bartels D, Best AA et al (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75. doi:10.1186/1471-2164-9-75

Bayjanov JR, Molenaar D, Tzeneva V et al (2012) PhenoLink—a web-tool for linking phenotype to ~omics data for bacteria: application to gene-trait matching for Lactobacillus plantarum strains. BMC Genomics 13:170. doi:10.1186/1471-2164-13-170

Brown JR, Douady CJ, Italia MJ et al (2001) Universal trees based on large combined protein sequence data sets. Nat Genet 28:281–285. doi:10.1038/90129

Brown MV, Lauro FM, DeMaere MZ et al (2012) Global biogeography of SAR11 marine bacteria. Mol Syst Biol 8:595. doi:10.1038/msb.2012.28

Cadillo-Quiroz H, Didelot X, Held NL et al (2012) Patterns of gene flow define species of thermophilic Archaea. PLoS Biol 10:e1001265. doi:10.1371/journal.pbio.1001265

Caro-Quintero A, Konstantinidis KT (2012) Bacterial species may exist, metagenomics reveal. Environ Microbiol 14:347–355. doi:10.1111/j.1462-2920.2011.02668.x

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. doi:10.1099/ijs.0.054171-0

Coenye T, Vandamme P (2003) Extracting phylogenetic information from whole-genome sequencing projects: the lactic acid bacteria as a test case. Microbiology 149:3507–3517. doi:10.1099/mic.0.26515-0

Coenye T, Gevers D, Van de Peer Y et al (2005) Towards a prokaryotic genomic taxonomy. FEMS Microbiol Rev 29:147–167

Cohan F (2001) Bacterial species and speciation. Syst Biol 50:513–524

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

Cordero OX, Polz MF (2014) Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol 12:263–273. doi:10.1038/nrmicro3218

Cordero OX, Ventouras LA, DeLong EF, Polz MF (2012) Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. Proc Natl Acad Sci USA 109:20059–20064. doi:10.1073/pnas.1213344109

Cuevas DA, Garza D, Sanchez SE et al (2014) Elucidating genomic gaps using phenotypic profiles [v1; ref status: approved with reservations 1, http://f1000r.es/488]. F1000Research 3:210. doi:10.12688/f1000research.5140.1

De Ley J (1970) Reexamination of the association between melting point, buoyant density, and chemical base composition of deoxyribonucleic acid. J Bacteriol 101:738–754

De Queiroz K (2005) Ernst Mayr and the modern concept of species. Proc Natl Acad Sci USA 102(Suppl1):6600–6607. doi:10.1073/pnas.0502030102

Denef VJ, Mueller RS, Banfield JF (2010) AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. ISME J 4:599–610. doi:10.1038/ismej.2009.158

Didelot X, Bowden R, Wilson DJ et al (2012) Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 13:601–612. doi:10.1038/nrg3226

Doi Y, Hazen TH, Boitano M, et al (2014) Whole genome assembly of *Klebsiella pneumoniae* co-producing NDM-1 and OXA-232

carbapenemases using single-molecule, real-time sequencing. Antimicrob Agents Chemother 58(10):5947–5953. doi:10.1128/AAC.03180-14
Doolittle WF, Papke RT (2006) Genomics and the bacterial species problem. Genome Biol 7:116. doi:10.1186/gb-2006-7-9-116
Doolittle WF, Zhaxybayeva O (2009) On the origin of prokaryotic species. Genome Res 19:744–756. doi:10.1101/gr.086645.108
Dutilh BE, Backus L, Edwards RA et al (2013) Explaining microbial phenotypes on a genomic scale: GWAS for microbes. Brief Funct Genomics 12:366–380. doi:10.1093/bfgp/elt008
Dutilh BE, Thompson CC, Vicente AC et al (2014) Comparative genomics of 274 *Vibrio cholerae* genomes reveals mobile functions structuring three niche dimensions. BMC Genomics 15:654. doi:10.1186/1471-2164-15-654
Dykhuizen D (2005) Species numbers in bacteria. Proc Calif Acad Sci 56:62–71
Ellegaard KM, Klasson L, Näslund K et al (2013) Comparative genomics of Wolbachia and the bacterial species concept. PLoS Genet 9:e1003381. doi:10.1371/journal.pgen.1003381
Forde BM, Ben Zakour NL, Stanton-Cook M et al (2014) The complete genome sequence of *Escherichia coli* EC958: a high quality reference sequence for the globally disseminated multidrug resistant *E. coli* O25b:H4-ST131 clone. PLoS One 9:e104400. doi:10.1371/journal.pone.0104400
Franzosa EA, Morgan XC, Segata N et al (2014) Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci USA. doi:10.1073/pnas.1319284111
Fraser C, Hanage WP, Spratt BG (2007) Recombination and the nature of bacterial speciation. Science 315:476–480. doi:10.1126/science.1127573.Recombination
Fraser C, Alm EJ, Polz MF et al (2009) The bacterial species challenge: ecological diversity. Science 323:741–746
Gevers D, Cohan FM, Lawrence JG et al (2005) Re-evaluating prokaryotic species. Nat Rev Microbiol 3:733–739
Giovannoni SJ, Tripp HJ, Givan S et al (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309:1242–1245. doi:10.1126/science.1114057
Haley BJ, Grim CJ, Hasan NA et al (2010) Comparative genomic analysis reveals evidence of two novel Vibrio species closely related to *V. cholerae*. BMC Microbiol 10:154. doi:10.1186/1471-2180-10-154
Hanage WP (2013) Fuzzy species revisited. BMC Biol 11:41. doi:10.1186/1741-7007-11-41
Hanage WP, Fraser C, Spratt BG (2005) Fuzzy species among recombinogenic bacteria. BMC Biol 3:6. doi:10.1186/1741-7007-3-6
Huang K, Brady A, Mahurkar A et al (2014) MetaRef: a pan-genomic database for comparative and community microbial genomics. Nucleic Acids Res 42:D617–D624. doi:10.1093/nar/gkt1078
Hunt DE, David LA, Gevers D et al (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320:1081–1085. doi:10.1126/science.1157890
Kämpfer P, Glaeser SP (2012) Prokaryotic taxonomy in the sequencing era—the polyphasic approach revisited. Environ Microbiol 14:291–317. doi:10.1111/j.1462-2920.2011.02615.x
Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11:283–290
Karp PD, Ouzounis CA, Moore-Kochlacs C et al (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res 33:6083–6089. doi:10.1093/nar/gki892
Kashtan N, Roggensack SE, Rodrigue S et al (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science 344:416–420. doi:10.1126/science.1248575
Klenk H-P, Göker M (2010) En route to a genome-based classification of Archaea and Bacteria? Syst Appl Microbiol 33:175–182. doi:10.1016/j.syapm.2010.03.003
Konstantinidis KT, DeLong EF (2008) Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J 2:1052–1065. doi:10.1038/ismej.2008.62
Konstantinidis KT, Stackebrandt E (2013) Defining taxonomic ranks. In: Rosenberg E, DeLong EF, Lory S, et al (eds) Prokaryotes (4th ed). Prokaryotic Biol. Symbiotic Assoc., 4th edn p 229
Konstantinidis KT, Tiedje JM (2005) Towards a genome-based taxonomy for prokaryotes. J Bacteriol 187:6258–6264. doi:10.1128/JB.187.18.6258
Kyrpides NC (1999) Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. Bioinformatics 15:773–774
Lan R, Reeves PR (2000) Intraspecies variation in bacterial genomes: the need for a species genome concept. Trends Microbiol 8:396–401
Larsen MV, Cosentino S, Lukjancenko O et al (2014) Benchmarking of methods for genomic taxonomy. J Clin Microbiol 52:1529–1539. doi:10.1128/JCM.02981-13
Li J, Jia H, Cai X et al (2014) An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. doi:10.1038/nbt.2942
Mallet J (2008) Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. Philos Trans R Soc Lond B Biol Sci 363:2971–2986. doi:10.1098/rstb.2008.0081
Markowitz VM, Korzeniewski F, Palaniappan K et al (2006) The integrated microbial genomes (IMG) system. Nucleic Acids Res 34:D344–D348. doi:10.1093/nar/gkj024
Markowitz VM, Chen I-MA, Palaniappan K et al (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res 42:D560–D567. doi:10.1093/nar/gkt963
Mayr E (1942) Systematics and the origin of species—from the viewpoint of a zoologist. Harvard Univ. Press, Cambridge
Mick E, Sorek R (2014) High-resolution metagenomics. Nat Biotechnol 32:750–751. doi:10.1038/nbt.2962
Moreira APB, Duytschaever G, Tonon LAC et al (2014a) Photobacterium sanctipauli sp. nov. isolated from bleached *Madracis decactis* (Scleractinia) in the St Peter & St Paul Archipelago, Mid-Atlantic Ridge, Brazil. Peer J 2:e427. doi:10.7717/peerj.427
Moreira APB, Duytschaever G, Tonon LAC et al (2014b) Vibrio madracius sp. nov. isolated from *Madracis decactis* (Scleractinia) in St Peter & St Paul Archipelago, Mid-Atlantic Ridge, Brazil. Curr Microbiol 2:e427. doi:10.1007/s00284-014-0600-1
Nesbø CL, Dlutek M, Doolittle WF (2006) Recombination in Thermotoga: implications for species concepts and biogeography. Genetics 172:759–769. doi:10.1534/genetics.105.049312
Nielsen HB, Almeida M, Juncker AS et al (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. doi:10.1038/nbt.2939
Overbeek R, Olson R, Pusch GD et al (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res 42:D206–D214. doi:10.1093/nar/gkt1226
Pagani I, Liolios K, Jansson J et al (2012) The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 40:D571–D579. doi:10.1093/nar/gkr1100
Polz MF, Hunt DE, Preheim SP, Weinreich DM (2006) Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. Philos Trans R Soc Lond B Biol Sci 361:2009–2021. doi:10.1098/rstb.2006.1928
Polz MF, Alm EJ, Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet 29:170–175. doi:10.1016/j.tig.2012.12.006

Preheim SP, Timberlake S, Polz MF (2011) Merging taxonomy with ecological population prediction in a case study of Vibrionaceae. Appl Environ Microbiol 77:7195–7206. doi:10.1128/AEM.00665-11

Ramasamy D, Mishra AK, Lagier J-C et al (2014) A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol 64:384–391. doi:10.1099/ijs.0.057091-0

Rinke C, Schwientek P, Sczyrba A et al (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437. doi:10.1038/nature12352

Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B et al (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7:828–836. doi:10.1038/nrmicro2235

Rohwer F, Edwards R (2002) The phage proteomic tree: a genome-based taxonomy for phage. J Bacteriol 184:4529–4535. doi:10.1128/JB.184.16.4529

Romero P, Wagg J, Green ML et al (2005) Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 6:R2. doi:10.1186/gb-2004-6-1-r2

Rosselló-Móra R (2012) Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. Environ Microbiol 14:318–334. doi:10.1111/j.1462-2920.2011.02599.x

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

Sen A, Daubin V, Abrouk D, Gifford I, Berry AM, Normand P (2014) Phylogeny of the class Actinobacteria revisited in the light of complete genomes. The orders 'Frankiales' and Micrococcales should be split into coherent entities: proposal of Frankiales ord. nov., Geodermatophilales ord. nov., Acidothermales ord. nov. and Nakamurellales ord. nov. Int J Syst Evol Microbiol 64:3821–3832. doi:10.1099/ijs.0.063966-0

Shapiro BJ, Polz MF (2014) Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol 22:235–247. doi:10.1016/j.tim.2014.02.006

Shapiro BJ, Friedman J, Cordero OX et al (2012) Population genomics of early events in the ecological differentiation of bacteria. Science 336:48–51. doi:10.1126/science.1218198

Shiwa Y, Yanase H, Hirose Y et al (2014) Complete genome sequence of *Enterococcus mundtii* QU 25, an efficient l-(+)-Lactic Acid-producing bacterium. DNA Res 21:369–377. doi:10.1093/dnares/dsu003

Sneath PHA, Sokal RR (1973) The principles and practice of numerical classification. Numer. Taxon

Snipen L, Ussery DW (2010) Standard operating procedure for computing pangenome trees. Stand Genomic Sci 2:135–141. doi:10.4056/sigs.38923

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

Stackebrandt E, Frederiksen W, Garrity GM et al (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. doi:10.1099/ijs.0.02360-0.02360

Stackebrandt E, Smith D, Casaregola S et al (2014) Deposit of microbial strains in public service collections as part of the publication process to underpin good practice in science. Springerplus 3:208. doi:10.1186/2193-1801-3-208

Staley JT (2006) The bacterial species dilemma and the genomic–phylogenetic species concept. Philos Trans R Soc Lond B Biol Sci 361:1899–1909. doi:10.1098/rstb.2006.1914

Staley JT (2009) Universal species concept: Pipe dream or a step toward unifying biology? J Ind Microbiol Biotechnol 36:1331–1336. doi:10.1007/s10295-009-0642-8

Tamames J, Rosselló-Móra R (2012) On the fitness of microbial taxonomy. Trends Microbiol 20:514–516. doi:10.1016/j.tim.2012.08.012

Thompson CC, Vicente ACP, Souza RC et al (2009) Genomic taxonomy of vibrios. BMC Evol Biol 9:258. doi:10.1186/1471-2148-9-258

Thompson C, Vieira NM, Vicente A, Thompson F (2011a) Towards a genome based taxonomy of Mycoplasmas. Infect Genet Evol 11:1798–1804. doi:10.1016/j.meegid.2011.07.020

Thompson FL, Thompson CC, Dias GM et al (2011b) The genus Listonella MacDonell and Colwell 1986 is a later heterotypic synonym of the genus *Vibrio Pacini* 1854 (Approved Lists 1980)—a taxonomic opinion. Int J Syst Evol Microbiol 61:3023–3027. doi:10.1099/ijs.0.030015-0

Thompson CC, Chimetto L, Edwards RA et al (2013a) Microbial genomic taxonomy. BMC Genomics 14:913. doi:10.1186/1471-2164-14-913

Thompson CC, Emmel VE, Fonseca EL et al (2013b) Streptococcal taxonomy based on genome sequence analyses. F1000Res 2:67. doi:10.12688/f1000research.2-67.v1

Thompson CC, Silva GZ, Vieira NM et al (2013c) Genomic taxonomy of the genus Prochlorococcus. Microb Ecol 66:752–762. doi:10.1007/s00248-013-0270-8

Tindall BJ, Rosselló-Móra R, Busse H-J et al (2010) Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst Evol Microbiol 60:249–266. doi:10.1099/ijs.0.016949-0

Vandamme P, Peeters C (2014) Time to revisit polyphasic taxonomy. Antonie Van Leeuwenhoek 106:57–65

Wayne LG, Brenner DJ, Colwell RR, et al (1987) International Committee on Systematic Bacteriology announcement of the report of the ad hoc Committee on Reconciliation of Approaches to Bacterial Systematics. Int J Syst Bacteriol 463–464

Whitman WB (2009) The modern concept of the procaryote. J Bacteriol 191:2000–2005. doi:10.1128/JB.00962-08 Discussion 2006–2007

Wiedenbeck J, Cohan FM (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35:957–976. doi:10.1111/j.1574-6976.2011.00292.x

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

Woyke T, Xie G, Copeland A et al (2009) Assembling the marine metagenome, one cell at a time. PLoS One 4:e5299. doi:10.1371/journal.pone.0005299

Wright F (1990) The effective number of codons used in a gene. Gene 87:23–29

Wu D, Hugenholtz P, Mavromatis K et al (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462:1056–1060. doi:10.1038/nature08656.A

Yarza P, Richter M, Peplies J et al (2008) The all-species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol 31:241–250

Zhi X-Y, Zhao W, Li W-J, Zhao G-P (2012) Prokaryotic systematics in the genomics era. Antonie Van Leeuwenhoek 101:21–34. doi:10.1007/s10482-011-9667-x