Antonie van Leeuwenhoek (2012) 101:21–34
DOI 10.1007/s10482-011-9667-x

PERSPECTIVE

# Prokaryotic systematics in the genomics era

Xiao-Yang Zhi · Wei Zhao · Wen-Jun Li ·
Guo-Ping Zhao

Received: 30 August 2011 / Accepted: 24 October 2011 / Published online: 25 November 2011

**Abstract** As an essential and basic biological discipline, prokaryotic systematics is entering the era of genomics. This paradigmatic shift is significant not only for understanding molecular phylogeny at the whole genome level but also in revealing the genetic or epigenetic basis that accounts for the phenotypic criteria used to classify and identify species. These developments provide an opportunity and a challenge for systematists to reanalyze the molecular mechanisms underlying the taxonomic characteristics of prokaryotes by drawing the knowledge from studies of genomics and/or functional genomics employing platform technologies and related bioinformatics tools. It is expected that taxonomic books, such as Bergey's Manual of Systematic Bacteriology may evolve into a systematics library indexed by phylogenomic information with an comprehensive understanding of prokaryotic speciation and associated increasing knowledge of biological phenomena.

**Keywords** Taxonomy · Genomics · Prokaryotic systematics · Molecular phylogeny

Xiao-Yang Zhi and Wei Zhao contribute equally to the article.

X.-Y. Zhi · W.-J. Li
Key Laboratory of Microbial Diversity in Southwest China, Ministry of Education and the Laboratory for Conservation and Utilization of Bio-Resources, Yunnan Institute of Microbiology, Yunnan University, Kunming 650091, China

W. Zhao · G.-P. Zhao (✉)
Key Laboratory of Synthetic Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China
e-mail: gpzhao@sibs.ac.cn

W. Zhao · G.-P. Zhao
State Key Laboratory of Genetic Engineering, Department of Microbiology, School of Life Sciences and Institute of Biomedical Sciences, Fudan University, Shanghai 200433, China

W. Zhao
China HYK Gene Technology Company Ltd., Shenzhen, Guangdong 518057, China

G.-P. Zhao
Shanghai-MOST Key Laboratory of Disease and Health Genomics, Chinese National Human Genome Center at Shanghai, Shanghai 201203, China

G.-P. Zhao
Department of Microbiology and Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong, SAR, China

## Current ground rules in prokaryotic taxonomy

Microbial systematics is the scientific study of the kinds and diversity of microorganisms and of relationships between them (Goodfellow and O'Donnell 1993). It is a basic scientific discipline that encompasses classification, nomenclature and identification and includes studies on genetic mechanisms, which underpin evolutionary processes and phylogeny. The first step, classification, involves the generation of an orderly and reliable framework for accommodating individual strains based on similarities and differences of their characters, though there has been a tendency to give more weight to differences in practice. The next stage, nomenclature, deals with the terms used to recognize ranks in the taxonomic hierarchy (e.g. genera and species) and with the important practice of giving the correct, internationally recognized names to taxonomic groups by following the rules laid out in the International Code of Nomenclature of Bacteria (Sneath 1992). The final step, identification, is both the act and result of establishing whether strains belong to established and validly published taxa. This involves determining the key characteristics of unknown isolates by using standard methods and criteria. Isolates found outside known groups should be described and classified as new taxa. It should be noted that the terms classification and taxonomy are not synonymous, the latter denotes the theoretical study of classification, including its bases, principles and roles (Simpson 1961).

Classification is the basis to other sciences, but at the same time is dependant on them for the acquisition of new data derived from technological advances. However, the basic unit of classification (and identification) of the species through a universally accepted definition of species in prokaryotic systematics is still a highly charged issue (Goodfellow et al. 1997; Schleifer 2009). In contrast, it is well known that the classification of prokaryotic groups passes through three steps, alpha (analytical phase), beta (synthetic phase), and gamma (biological phase) taxonomy.

The first taxonomy stage, i.e., the alpha taxonomy is the level at which species are classified, named and identified. Then, the beta taxonomy covers the assignment of species to natural classifications; these may be based on either phenetic or phylogenetic criteria (Goodfellow and O'Donnell 1993). Phylogenetic classifications are often considered to be the most theoretically sound (Doolittle 1999) and most beautiful in nature (Pace 2009). Phylogenetic criteria, notably 16S rRNA sequence variations in archaea and bacteria, are seen to provide the backbone for the classification of prokaryotes (Vandamme et al. 1996; Tindall et al. 2010). However, current approaches to the classification of prokaryotes rest on the integrated use of genotypic and phenotypic features acquired through the application of chemotaxonomic, molecular systematic and numerical taxonomic procedures (Goodfellow and O'Donnell 1993; Vandamme et al. 1996; Tindall et al. 2010). This practice, known as polyphasic taxonomy was introduced by Colwell (1970) to encompass successive or simultaneous studies on groups of prokaryotes using methods chosen to yield high quality genotypic and phenotypic data. The extensive application of polyphasic taxonomy has led to marked improvements in the classification of prokaryotes that in turn has provided a sound basis for stable nomenclature and improved identification, as exemplified in the present edition of *Bergey's Manual of Systematic Bacteriology* (de Vos et al. 2009; Krieg et al. 2010; Goodfellow et al. 2011). Nevertheless, the polyphasic approach to classification is essentially utilitarian and it does not address the need to generate a theory-derived classification based on phylogenetic/evolutionary concepts (Schleifer 2009).

The final stage, gamma taxonomy, covers intraspecific categories such as subspecies, ecotypes and polymorphisms and concerns over biological aspects of taxa. The analysis of intraspecific variation and related evolutionary processes is critical in revealing the underlining mechanism of speciation, an important aspect of systematics. However, most studies in this area are carried out by scientists working in ecology (environmental biology) (Lucker et al. 2010; Mira et al. 2010) and epidemiology (medical biology) (Morschhauser et al. 2000; Morelli et al. 2010) rather than by taxonomists.

Classification is a prerequisite for identification (Priest and Williams 1993). The development of both disciplines depended heavily on the innovations in technology (Klenk and Goker 2010). In general, the kinds of characters used to describe "similarities" and "differences" between microbial taxa depend on which techniques and associated tools are used. Reliance on microscopy and pure cultures led Ferdinand Cohn (1872) to classify bacteria into six genera based on morphological properties, a study that started an era whereby microbiologists began to reveal the tremendous diversity of microorganisms. Initially, the most important taxonomic markers used in this classical approach were limited to morphology, growth requirements, and pathogenic potential. Later, serological traits and other physiological characters were used to distinguish among different bacteria, notably pathogens and their subtypes, a skew that is still evident today. In the first half of the last century, more and more biochemical data enriched our knowledge of enzymology and metabolism, thereby further facilitating the recognition of different kinds of microbes (Buchanan 1955). This chemotaxonomic approach significantly improved the resolution of classifications when compared to those based on morphological features (Schleifer and Stackebrandt 1983) and thereby leading to a step forward in microbial systematics.

Chemical composition of genomic DNA (GC content) was one of the important chemotaxonomic characters to be widely used in classification. Then, microbial systematists realized the significance of DNA sequence information during the early days of DNA–DNA hybridization and later the importance of rRNA sequence studies. The emergence of phylogenetic inference based on the sequence of small subunit ribosomal RNA not only led to the recognition of the *Archaea* as a separate kingdom (Woese and Fox 1977), but also moved prokaryotic systematics into a new era. In this respect, it has to be emphasized that, in the final analysis, the genome is the ultimate record of the evolutionary history of life (Zuckerkandl and Pauling 1965; Boussau and Daubin 2010). It now needs to be recognized that fast developing sequencing techniques provide a new key that will lead to the classification of prokaryotes based on genomic data (Wu et al. 2009; Metzker 2010). The formerly implausible possibility of using data from entire genome sequences in prokaryotic classification is becoming, or will soon, become a reality. In fact, a few distinct but related phylogenetic systems have been developed based on genotypic information derived from DNA structural information (Wu et al. 2009).

## Genomic information for prokaryotic systematics

The availability of ever increasing whole-genome data (see: http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) and functional genomic analyses have significantly improved our understanding of the biochemistry, genetics, physiology and evolution of microorganisms (Wu et al. 2009). The explosion of genomic information provides unprecedented opportunities for assessing taxonomic relationships between microorganisms, thereby allowing the generation of molecular phylogenies. Comparison of related genomes and inferences drawn from ancestral ones will allow description of species by characterizing genetic events, such as gene duplication, gene decay, horizontal gene transfer, as well as indels (insertions and deletions) and single nucleotide polymorphisms (SNPs), at chromosomal and gene levels. So far a few models have been proposed to address fundamental evolutionary questions; some of which have demonstrated power with accuracy (Coenye et al. 2005; Konstantinidis and Tiedje 2005; Boussau and Daubin 2010).

Gene duplication is important as it influences genetic adaptation of microorganisms to changing environments, notably by genome expansion, thereby promoting species diversity in nature (Hooper and Berg 2003; Hittinger and Carroll 2007; Innan and Kondrashov 2010). The analysis of paralogous genes within whole genomes shows that more than 40% of the coding capacity of a bacterial genome may have originated through gene duplication (Jordan et al. 2001; Gevers et al. 2004). Initially, it was proposed that bacterial genomes might have evolved from a small ancestral genome through several gene duplications (Kunisawa 1995) but inferences drawn from currently available genomes indicate that gene duplication has a modest effect on genome evolution (Kolsto 1997). On the other hand, it is worth mentioning that genes involved in environmental adaptation are retained after duplication (Gevers et al. 2004) suggesting that there is a role for gene duplication in microbial evolution.

As an opposite force, gene decay leads to the contraction of genomes. Complete or partial nucleotide deletions in functional genes may lead to inactive genes or pseudogenes, respectively (Andersson and Andersson 2001). The influence of gene decay is variable within different bacterial lineages, it is particularly apparent in some bacterial groups with a host-associated lifestyle, such as *Mycobacterium leprae* (3.2 Mb). The genome of this organism is less than half of that of the nonpathogenic *Mycobacterium smegmatis* (7.0 Mb), as it contains 1,116 pseudogenes and inactive genes (Cole et al. 2001; Monot et al. 2009)

Prokaryotes have evolved other mechanisms for rapid adaptation to new environmental niches. The introduction of novel genes into prokaryotes by horizontal gene transfer (HGT) may lead to diversification and speciation (Lawrence and Retchless 2009; Ochman et al. 2000). Taking this concept to an extreme, it can be claimed that two taxa are more similar to one another than to a third one not because they share a more recent ancestor but because they exchange genes more frequently (Gogarten et al. 2002). The estimated frequency of HGT genes in whole genomes of prokaryotes is usually low (Kunin and Ouzounis 2003), however, HGT may play a significant biological role in their evolution through, for instance, the acquisition of antibiotic resistance or pathogenic properties. In antibiotic producing actinomycetes, HGT is usually observed in non-conserved regions of the genome, i.e., the non-core regions, indicating the effect of more recent events (Bentley et al. 2002; Philippe and Douady 2003).

Chromosomal rearrangement is a genetic event that tends to influence whole genome organization more than genome content. Its occurrence largely depends on the presence of repeats and mobile elements, such as insertion sequences, transposons and prophage sequences (Kolsto 1997; Bennett 2004). Large-scale chromosomal rearrangement may lead to a huge inversion of DNA segments manifesting as an X-shaped pattern in alignments of two complete genomes. Generally, more distantly related bacterial taxa show a higher level of chromosomal rearrangement, and consequently, a more irregular gene order (Rocha 2004).

In contrast to the large-scale genomic plasticity described above, single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) are “small” genetic variations, which present at a higher rate in the history of the genome revolution (Gupta and Griffiths 2002; Gao and Gupta 2012, in press). Basically, SNPs can occur at any nucleotide within the genome resulting in base substitution or gene truncation mutations while small indels may lead to truncation, deletion or frameshifts (TDF) of the affected genes. However, analysis of whole genomes has shown that the presence of SNPs and indels is not stochastic as such changes are either preserved or lost depending upon adaptations to the environment (Pearson et al. 2009). Because the characteristic status of SNPs is limited to four possible nucleotides, and for its constant mutation rate, SNPs can correlate the samples’ discrepancy time from their ancestor, they have been widely and efficiently used for phylogenetic tree construction covering decades or centuries of microevolution, a task of gamma taxonomy (Gupta 2001; Morelli et al. 2010; Pearson et al. 2009). However, since many SNPs and indels may occur among closely related bacteria, it is usually difficult to identify the genetic divergences that account for the phenotypic differences among them, unless some non-synonymous mutations or TDFs are found to be responsible for phenotypic variations critical in taxonomy (Zhao et al.2010).

The universally accepted DNA sequence-based method currently used is based on analyses of 16S rRNA gene sequences. The conservation of the 16S rRNA gene made it one of the best candidates for low cost PCR and sequencing studies. In addition, the proportion of information content to length is relatively high thereby providing high resolution and well-supported phylogenetic trees that show relationships from genera to phyla (Ludwig and Schleifer 1994). On the other hand 16S rRNA sequences often contain insufficient information to show relationships at lower taxonomic ranks, particularly at the level of species and subspecies (Stackebrandt et al. 2002). In addition, nucleotide variations within multiple rRNA operons in a given genome (Harrington and On 1999; Pei et al. 2010), as well as the possibility of 16S rRNA genes derived from HGT (Ueda et al. 1999; Schouls et al. 2003), may distort relationships between taxa in phylogenetic trees.

Since Darwin’s era, systematists have been classifying individual species in order to reflect their inferred evolutionary relationships. The availability of complete genomes and the types of genomic variations reviewed above, make it possible to reconstruct phylogenies based on a much larger data set, from which more reliable and accurate trees of life can be built as shown in Table 1. So far, most studies on prokaryotic classifications based on genomic sequences have been focused on one or a few methods. It is critical to understand that each method has a limited resolution covering part, but not all levels of taxonomic information (Coenye et al. 2005). Recently, large-scale studies have integrated two or more methods based on genome content and chromosomal

**Table 1** Approaches for assessing taxonomic relationships based on whole-genome sequences

| Approach | Description | Reference |
|---|---|---|
| Trees based on genome content analysis (generally, identification and comparison of orthologous genes) | 1. Number of COGs between two genomes depends on evolution distance between them | Bansal and Meyer (2002); Brown et al. (2001) |
| | 2. Number and category of COGs selected for comparison and phylogenetic tree construction is important | Coenye and Vandamme (2003), Huynen and Bork (1998); Wolf et al. (2002) |
| | 3. Reasonably good congruence to 16S rRNA sequence-based species relationships | Huson and Steel (2004); Huynen and Bork (1998); Snel et al. (1999). |
| | 4. HGT effect has yet to be thoroughly analyzed. Phylogenetically closely related species may not necessarily share more COGs, while species adaptation to similar niches does | Dutilh et al. (2004); Zhao et al. (2010) |
| Trees based on presence and absence of genes | 1. An alternative approach aiming at analyzing genome content | Fitz-Gibbon and House (1999); House and Fitz-Gibbon (2002); Wolf et al. (1999) |
| | 2. Initial results show good congruence with 16S rRNA trees | Lin and Gerstein (2000) |
| | 3. There are exceptions to point 2 | |
| Trees based on indels or SNPs in conserved genes | 1. Most widely distributed at the genomic level reflecting fine evolutionary differences | Gupta et al. (2003)<br>Foster et al. (2009); Pearson et al. (2009) |
| | 2. The relatively short evolutionary history of SNPs and indels leaves less time to evolve reversals or convergent mutations | Chuang et al. (2010)<br>Snel et al. (2005); Ventura et al. (2007) |
| | 3. This approach is especially useful for defining relationships between species | |
| | 4. Currently limited by the numbers of taxa with entire genomes for all members of a lineage. Still prohibitively expensive for sequencing deeply | |
| Trees based on chromosomal gene order | 1. For a large part, gene order depends on gene content. This approach needs a large-scale comparison of COG genes | Korbel et al. (2002); Snel et al. (2005); Wolf et al. (2001) |
| | 2. Gene order evolves faster than gene content. It is more suited for resolving the phylogeny of closely related species | Coenye et al. (2005); Huynen and Bork (1998) |
| | 3. It is sensitive to chromosomal rearrangement | Suyama and Bork (2001) |
| Trees based on metabolic pathway content analysis | 1. A method analyzing metabolic pathways with physiological significance at genomic level | Hong et al. (2004); Ma and Zeng (2004) |
| | 2. The major results of metabolic network based phylogenetic trees show good congruence with 16S rRNA gene trees and gene content trees | Snel et al. (2005); Sutcliffe (2010); Zhao et al. (2010) |
| | 3. Could be limited by a lack of understanding of metabolic pathway reactions or by dearth of functional genomic data | Okura et al. (2008); Sun et al. (2011); Wang et al. (2010) |

organization (Kunin et al. 2005; Mira et al. 2010). Historically, systematists have been seeking a tree that is “fairly true for each great kingdom of Nature”, and which represents a “truly evolution history of Life”. However, there are still gaps between classification and phylogeny with respect to understanding the evolution of life based on genomic sequences, even with complete genome sequences. From this perspective, four issues must be taken into consideration.

(1) *Artifacts may result from the selection of unrepresentative samples*. Although a phylogeny-driven GEBA (Genomic Encyclopedia of *Bacteria* and *Archaea*) program was successfully initiated (Wu et al. 2009), the increasing numbers of genomes available from databases such as NCBI remain biased towards organisms of biotechnological and medical importance. Furthermore, it is estimated that more than 99% of all microorganisms present in natural ecosystems cannot be cultured using routine techniques (Hugenholtz et al. 1998). Therefore, selectively sequencing genomes of representative samples of environmental diversity (like metagenomics) will become increasingly important for taxonomic research, especially when compared with traditional methods.

(2) *Artifacts may result from the use of unsophisticated mathematical methods in the construction of phylogenetic trees*. The use of appropriate mathematical models will become increasingly important as more and more genome datasets become available. The use of more reliable mathematical models will lead to improved precision but not necessarily to improved accuracy if systematic biases cannot be resolved by the analytical methods (Rannala and Yang 2008).

(3) *The increasing number of genes used in tree building, may result in fewer common characters (genes) left in genomes that can be used as phylogenetic signals***.** The common characters within a group are not usually shared among sister groups hence only a few genes, most of which encode for ribosomal proteins (Wu et al. 2009), can be used to reconstruct the tree of life. This might explain why the phylogeny of ribosomal RNA genes are usually consistent with the tree of life simulated by concatenated conserved gene sets. On the other hand, although there are far greater numbers of genes encoding indispensable metabolic processes for free living cells than the number of ribosomal proteins, the variations among the orthologous genes of distant species are too high to be identified exactly, completely and easily with respect to techniques, i.e., the commonly employed bi-directional best hit method (Tatusov et al. 1997). Hence, the identification of conserved orthologous groups (COG) determines the proportion of genomic information used in subsequent analyses.

(4) *Another crucial issue for phylogenomic analysis is how to understand and correlate results generated from analyses of incongruent genome features.* This problem is especially serious when the theoretical bases of methods are completely different, such as phylogenies based on supermatrix and on string frequency. In other words, whether it is an artifact or realistic, the result we expect to obtain is a tree that is in accordance with trees of individual genes and species, that is, the biological information needed to understand and explain the tree. However, improvements are usually focused on aspects of mathematical methods without reference to the biological significance behind them. Consequently, integrated approaches based on biochemistry, genetics and physiology, will provide an opportunity to the evolutionary understanding of systematics at the genome level. This approach should hopefully develop into a trend for taxonomic research in future.

## Understanding speciation of prokaryotes and their biological impact with genomic information

There is a continuing debate about the concept of prokaryotic species (Table 2) though both systematists and evolutionary biologists believe that closely related species have some fundamental dynamic properties, albeit with a boundary amongst them (Achtman and Wagner 2008; Doolittle and Zhaxybayeva 2009; Ereshefsky 2010; Lawrence and Retchless 2010). However, more thought needs to be given to whether boundaries do exist and if so how they can be found avoiding drawing conclusions merely based on phenotypic criteria used to circumscribe so called taxonomic ‘species’ (Ereshefsky 2010). The phylophenetic species concept (Stackebrandt and Goebel 1994), which is based on three independent approaches (genomic boundaries determined by DNA–DNA hybridization; phenotype descriptions; and relationships based on the phylogeny of 16S rRNA genes), has been considered to be the most universally applicable in the delineation of prokaryotic species

(Rosselló-Mora and Amann 2001). Pragmatically, this approach defined a series of standards for the taxonomic characterisation of groups that could also be replicated between different laboratories. Its application provided stable and predictable classifications although some serious problems and drawbacks were evident (Coenye et al. 2005; Schleifer 2009). Strictly speaking, this approach provides an arbitrary and anthropocentric definition of prokaryotic species.

Macrobial systematists have attempted to fit their cluster-based demarcations in accordance with a theory, that is, successful interbreeding within animal and plant species. In contrast, since microorganisms have unparalleled diversity and population sizes, it is difficult to understand speciation processes based only on one of the various models proposed for a theory-based concept of species (Table 2). One so-called theory-based concept of microbial species is The Evolutionary Species Concept. However, due to the nature of prokaryotes and the difficulties in observing evolutionary tendencies amongst them, the application of this concept is not yet possible (Rosselló-Mora 2003). Recently, James Staley proposed a phylogenomic species concept (Staley 2009) that drew more information from genome sequences for phylogenetic reconstruction, mainly by multilocus sequence analysis (MLSA). This approach is not only theory-based but also pragmatic. However, more comprehensive phylogenetic signals should be generated from gene content, gene order, and other whole-genome features (Delsuc et al. 2005) and properties, which will soon be available from ongoing extensive prokaryotic genomic sequencing studies.

## Speciation based on multiplex variability of prokaryotic genomic evolution

Prokaryotic chromosomes have been sculptured more by various kinds of large DNA alterations than by mutations in single gene sequences (Mira et al. 2002). The CRISPR-Cas (clustered regularly inter-spaced short palindromic repeats-CRISPR-associated proteins) modules recently characterized in archaea and bacteria (Cui et al. 2008; Makarova et al. 2011) have revealed a high degree of evolutionary plasticity in prokaryotic genomes indicating that there are more processes giving rise to genetic novelty than previously thought. According to the different donors of genetic material, this evolutionary process can be divided into two categories: vertical and lateral inheritances.

Vertical inheritances provided sufficient evidence for recapitulating the Darwinian-Mendelian model of parent-to-offspring gene flow. However, this concept has been severely challenged by the quantitative and qualitative importance of genetic transfers between lineages, notably between prokaryotic species (Charlebois et al. 2003), though such phenomena have significant implications for the generation of a universal tree of life. Given the genetic connections, the topology of the evolutionary history of life becomes more reticulate than tree-like (Lopez and Bapteste 2009). The paradigmatic shift from a

**Table 2** Prokaryotic species concepts

| Concepts | Description | Reference |
|---|---|---|
| Cohesion species | A group of organisms whose divergence is capped by one or more forces of cohesion | Meglitsch (1954) |
| Biological species | Frequently genetic exchange occurs among organisms within a species | Mayr (1970) |
| Recombination species | Demarcated species as groups of microbes whose genomes can recombine | Dykuizen and Green (1991) |
| Phonetic species | A similarity concept based on statistically co-varying characteristics that are not necessarily universal among members of the taxon | Hull (1997) |
| Evolutionary species | A lineage concept that is explicitly temporal, treating these units as lineages extended in time | Mayden (1997) |
| Ecological species | Species seen as an evolutionary lineage bound by ecotype-periodic selection | Cohan (2001, 2002) |
| Adaptive divergence species | Explicitly based on evolutionary theory, specifically the stable ecotype model; it incorporates the processes of ecological adaptation, evolutionary descents and homologous recombination | Vos (2011) |

monistic to a pluralistic understanding of evolutionary processes is reflected by a graph-theoretical shift, from trees (i.e., connected acyclic graphs) to networks (i.e., connected graphs that may contain reticulations, Bapteste et al. 2009).

However, when HGT happened its effect as a disruptive force might influence the phylogenic construction of related organisms. HGT acquired ancient genes are more likely to be retained in all descendants, such as those encoding ATPases and aminoacyl-tRNA synthetases, though they could be differentially lost and/or secondarily transferred (Huang and Gogarten 2006). However, more ancient HGT is difficult to identify based on similarities or phylogenetic analyses. This means that the complication of evolutionary networks introduced by convoluted HGT should be limited to relative low-level taxonomic ranks. In other words, HGT occurs frequently amongst closely related individuals and species and rarely between genealogically distant relatives (Andam and Gogarten 2011).

A recent study revealed that the frequency of HGT was linearly correlated with similarities between donors and recipients in both genome and proteome sequences, with 86% of HGT occurring between pairs of organisms that had less than 5% difference in GC content (Popa et al. 2011). In addition, biased HGT has the possibility to generate evolutionary patterns similar to vertical inheritance, at least, the signal detected in the descendents with a common ancestor is difficult to be distinguished from the signal due to biased gene transfer (Andam et al. 2010). A case study comparing the level of incongruence in proteobacterial and eukaryotic genes indicated that HGT could not be considered as a major evolutionary process in these bacteria (Soria-Carrasco and Castresana 2008).

Even when the complication brought about by lateral inheritance was excluded, the phylogenetic incongruence of orthologous genes implied that they probably had a different evolutionary history (Bapteste et al. 2005); in particular the use of different tree reconstruction methods gave rise to a non-negligible statistically significant incongruence (Jeffroy et al. 2006). In practice, the congruence among the individual genes is usually confirmed by a two-step process. First, the candidate genes should be universally distributed; potentially incongruent genes should be excluded by statistical tests, e.g., the incongruence length difference (ILD) test (Farris et al.1994; Planet and Sarkar 2005). The retained genes should then concatenated to maximize the phylogenetic signal and enhance the statistical support for branches in the tree inferred by the large dataset. Next, the individual genes should be the subject of another test, such as the Kishino-Hasegawa (KH), Shimodaira-Hasegawa (SH) or the approximately unbiased (AU) test (Poptsova 2009) on the supposition that the super-tree is the best. Crucially, when many genes are used in an analysis, it is necessary to account for the fact that different genes undergo different selective pressures hence the rate heterogeneity within sites may vary from gene to gene (Bevan et al. 2007). However, if the heterogeneity of nucleotide frequencies among taxa is considered, this refers to the equality of the nucleotide frequency bias among species (Rosenberg and Kumar 2003), the analysis seems to go in a direction that cannot be easily controlled. Given that, we propose to establish a database containing the pre-built evolutionary model for each orthologous gene, and generate a standard method of phylogenetic analysis for the purpose of classification.

## Integrating the biological knowledge of taxa with prokaryotic systematics

Building classifications based on phylogenetic relationships between species is an essential facet of prokaryotic systematists. However, this is not an end in itself as it is also important to know the similarities or differences in the biological characteristics between diverse species. The ever-increasing genomic information will provide a great opportunity not only to delineate a more accurate and precise evolutionary history of prokaryotic species but will also raise our understanding of their distinctive biological properties.

In the era of chemotaxonomy, chemical characteristics of cellular components, particularly, cell wall and membrane constituents, were commonly used for prokaryotic classification and identification though the analysis and comparison of these chemical indices were laborious and time-consuming. However, the availability of whole-genome sequence data makes it a realistic proposition to gradually correlate chemotaxonomic phenotypes with the molecular genotypes of corresponding taxa, particularly at species and subspecies levels (Sutcliffe 2010; Zhao et al. 2010). Recently, by sequencing and comparing the first representative genome of the genus *Amycolatopsis* with model *Nocardia* and *Streptomyces* strains, the genetic basis of cell wall components of *Amycolatopsis mediterranei* U32 was intensively revealed (Zhao et al. 2010).

It should be noted that although the analysis mentioned above seemed straightforward, reliable results will only be obtained when the biosynthetic pathways and the enzymes catalyzing cell wall synthesis are thoroughly understood. Furthermore, although the function of an enzyme might be predicted by its evolutionary history (Eisen 1998), the genetic variations corresponding to the phenotypic differences may not be as simple as it was thought as observed discrepancies may be derived from different sources, such as multiple enzyme catalyzed reactions, and quantitative rather than qualitative differences in chemical components or enzyme activities. Consequently, it is essential to analyse the genomic variations at all levels and, where applicable, to determine epigenetic properties such as gene expression and protein modification. This means that the corresponding chemotaxonomic characters may need to be reanalyzed in a more quantitative or representative manner. In this context, sequence analysis of isoprenyl diphosphate synthases, which determine the chain length of menaquinones (MK) in actinomycetes may only distinguish between MK 7 and MK 8, not between longer chains (unpublished data, Zhao W et al.). Similarly, the molecular mechanism determining the percentage of different phospholipid components in cell membranes has still to be resolved (Barona-Gómez et al. 2012, in press). Nonetheless, it can be anticipated that the genetic basis, which accounts for traditional phenotypic properties will be identified in the near future thereby providing reliable data for the classification and identification of archaeal and bacterial taxa.

An understanding of the genetic basis of serotyping using whole genome sequence data is another prospective development at the subspecies/strains' level. Serotyping systems for *Escherichia coli* and *Salmonella* spp. are well established, and widely used to identify strains for epidemiological and surveillance purposes (Beutin et al. 2007; Switt et al. 2009). Compared to traditional technologies, genomic information provides a simpler and more convenient method for rapid serotyping by analysis of the gene clusters (or genes), which encode the synthesis of bacterial surface antigens (Liu et al. 2008). Recently, serotypes of several bacteria, such as *Cronobacter sakazakii*, *Proteus* and *Vibrio parahaemolyticus* were identified using this approach (Okura et al. 2008; Wang et al. 2010; Sun et al. 2011).

As we have emphasized, prokaryotic systematics is a fundamental biological discipline. The relationships among prokaryotic taxa should be based on their phylogenomic information attendant with biological knowledge encoded in genomes and expressed as their phenotypes. Comparative and functional genomic analyses need to be carried out in order to match up with corresponding phenotypes. Meanwhile, established relationships between biological knowledge and phylogenomic information can be expected to further facilitate biological research, not least with respect to uncultured bacteria where genome sequences can be derived from metagenomic sequencing (Petrosino et al. 2009; Mocali and Benedetti 2010).

## A perspective for the molecular systematics library of prokaryotes

Within a prokaryotic species, the gene reservoir available for inclusion in its pan-genome is vast. More genome-specific genes will continue to be identified following sequencing of hundreds of genomes (Tettelin et al. 2005). In contrast, the core genome of a species, including all genes responsible for its basic cellular functions, will not change dramatically, except in the case of some obligate bacterial symbionts (Moran 2003). In other words, the core genome shapes and maintains essential functions (Gil et al. 2004; Koonin 2003), while the peripheral genome contributes to species diversity and/or encodes accessory biochemical pathways and functions, which are not essential for bacterial growth but may confer selective advantages, such as adaptation to different niches or survival under stressful growth conditions (Medini et al. 2005).

The theoretical basis of the Ecological Species Concept emphasizes the aspect of pan-genome selectively outlined above (Koeppel et al. 2008; Pena et al. 2010). In contrast, attributes inherited from the last common ancestor of these ecotypes determine who they are, and where they came from. Nevertheless, the concept that 'everything is everywhere: but the environment selects' (O'Malley 2007) implies that before environmentally imposed selective pressures on strains, 'who they are' is of greater importance, especially for taxonomists. Consequently, we suggest that the phylogenomic backbone of prokaryotic systematics should be merged with knowledge on the biology of species (as it was for traditional taxonomy), including cellular structure (morphological traits), metabolism (biochemical traits) and development/differentiation (physiological traits) in order to understand their evolution along with their relationships within genera and/or within their ecological niche. Here, a molecular systematics library of prokaryotes based on cellular life is proposed to update the current taxonomic system.

As mentioned above, the present taxonomic system has been organized as a book or dictionary, with the phylogeny of 16S rRNA genes running through it. To date, this polyphasic approach has facilitated the classification of a remarkable diversity of prokaryotes (de Vos et al. 2009; Krieg et al. 2010; Goodfellow et al. 2011). On the other hand, the information included in this 'book' is restricted due to the limited information available on the genomic variation used to construct the framework, but also because finite phenogenetic characters were used to circumscribe the biological characteristics of prokaryotic species, especially in many cases where the phenotypes used for describing different taxa have yet to be correlated with their encoding genotypes.

All of the mismatches outlined above help account for the fact that classification nowadays is a more or less descriptive cataloging of natural history, this in turn leads to a superficial understanding of evolution and biology. In contrast, the plethora of knowledge to be gleaned from phylogenomic analyses of species through large scale sequencing efforts, will lead to the identification of critical biological traits (phenotypes), notably those revealed by genomic, functional genomic and/or proteomic analyses and related experimental studies. These developments will have a revolutionary impact on the way prokaryotes are classified and identified. These prospective changes are in their infancy as too few representative genomes are available, a situation that can be expected to change rapidly. Besides systems biology studies based on genomic information will continuously enrich our biological knowledge of individual species. In time, a molecular systematics library will be generated to accommodate all species, as an open source library in which phylogenies based on genomic sequences will be enriched by corresponding biological knowledge.

As we stated at the beginning, prokaryotic systematics is a fundamental biological discipline. However, the segregation of phylogeny and biology traits has made the subject more and more complex rather than providing a vehicle for explaining natural evolutionary and ecological systems. However, the developments outlined above should lead to a real understanding of the nature of *species* and why they are what they are, thereby moving prokaryotic systems away from merely recording similarities and differences between them. Similarly, with respect to single microorganisms the focus should be on understanding cellular processes by drawing from increasingly available phylogenomic information.

In summary, as the genomics era unfolds prokaryotic taxonomy and systematics, it should be remodelled so that taxa are defined by their biological nature. An attractive consequence of this development will be that systematics will no longer be seen as a laborious and lagging science but will become an exciting discipline based on ever increasing biological knowledge.

**Acknowledgments** This research was supported by the National Basic Research Program of China (Grant 2010CB833801 and 2012CB721102), the National Natural Science Foundation of China (Grants 30830002 and 31121001) and the Research Unit Fund of Li Ka Shing Institute of Health Sciences (Grant 7103506). We thank Prof. Mike Goodfellow for his helpful comments on our manuscript. And we also appreciate the comments made by Dr. Shuhua XU of the Institute of Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences.

## References

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

Andam CP, Gogarten JP (2011) Biased gene transfer in microbial evolution. Nat Rev Microbiol 9:543–555

Andam CP, Williams D, Gogarten JP (2010) Biased gene transfer mimics patterns created through shared ancestry. Proc Natl Acad Sci USA 107:10679–10684

Andersson JO, Andersson SG (2001) Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. Mol Biol Evol 18:829–839

Bansal AK, Meyer TE (2002) Evolutionary analysis by whole-genome comparisons. J Bacteriol 184:2260–2272

Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF (2005) Do orthologous gene phylogenies really support tree-thinking? BMC Evol Biol 5:33

Bapteste E, O'Malley MA, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L et al (2009) Prokaryotic evolution and the tree of life are two different things. Biol Direct 4:34

Barona-Gómez F, Cruz-Morales P, Noda-García L (2012). What can genome-scale metabolic network reconstructions do for prokaryotic systematics?. Antonie van Leeuwenhoek (in press)

Bennett PM (2004) Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement. Methods Mol Biol 266:71–113

Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD et al (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417:141–147

Beutin L, Miko A, Krause G, Pries K, Haby S, Steege K, Albrecht N (2007) Identification of human-pathogenic strains of Shiga toxin-producing *Escherichia coli* from food by a combination of serotyping and molecular typing of Shiga toxin genes. Appl Environ Microbiol 73:4769–4775

Bevan RB, Bryant D, Lang BF (2007) Accounting for gene rate heterogeneity in phylogenetic inference. Syst Biol 56:194–205

Boussau B, Daubin V (2010) Genomes as documents of evolutionary history. Trends Ecol Evol 25:224–232

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. Nat Genet 28:281–285

Buchanan RE (1955) Taxonomy. Annu Rev Microbiol 9:1–20

Charlebois RL, Beiko RG, Ragan MA (2003) Microbial phylogenomics: branching out. Nature 421:217

Chuang PC, Chen YM, Chen HY, Jou R (2010) Single nucleotide polymorphisms in cell wall biosynthesis-associated genes and phylogeny of *Mycobacterium tuberculosis* lineages. Infect Genet Evol 10:459–466

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

Coenye T, Gevers D, van de Peer Y, Vandamme P, Swings J (2005) Towards a prokaryotic genomic taxonomy. FEMS Microbiol Rev 29:147–167

Cohan FM (2001) Bacterial species and speciation. Syst Biol 50(4):513–524

Cohan FM (2002) What are bacterial species? Annu Rev Microbiol 56:457–487

Cohn F (1872) Untersuchungen űber Bakterien. Beitr Biol Pflanz 1875 1 (Heft 2):127–224

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N et al (2001) Massive gene decay in the leprosy bacillus. Nature 409:1007–1011

Colwell RR (1970) Polyphasic taxonomy of bacteria. In: Izuka H, Hasegawa T (eds) Culture collections of microorganisms. University of Tokyo Press, Tokyo, pp 421–436

Cui Y, Li Y, Gorge O, Platonov ME, Yan Y, Guo Z, Pourcel C, Dentovskaya SV et al (2008) Insight into microevolution of *Yersinia pestis* by clustered regularly interspaced short palindromic repeats. PLoS One 3:e2652

De Vos P, Garrity GM, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer K-H, Whitman WB (2009) Bergey's Manual of Systematic Bacteriology, 2nd Edn, Vol 3, The Firmacutes, Springer, New York

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361–375

Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124–2129

Doolittle WF, Zhaxybayeva O (2009) On the origin of prokaryotic species. Genome Res 19:744–756

Dutilh BE, Huynen MA, Bruno WJ, Snel B (2004) The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J Mol Evol 58:527–539

Dykuizen D, Green L (1991) Recombination in *Escherichia coli* and the definition of biological species. J Bacteriol 173:7257–7268

Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8:163–167

Ereshefsky M (2010) Microbiology and the species problem. Biol Philos 25:553–568

Farris JS, Källersjö M, Kluge AG, Bult C (1994) Testing the significance of incongruence. Cladistics 10:315–319

Fitz-Gibbon ST, House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res 27:4218–4222

Foster JT, Beckstrom-Sternberg SM, Pearson T, Beckstrom-Sternberg JS, Chain PS et al (2009) Whole-genome-based phylogeny and divergence of the genus *Brucella*. J Bacteriol 191:2864–2870

Gao B, Gupta RS (2012) Microbial Systematics in the Post-genomics Era. Antonie van Leeuwenhoek (in press)

Gevers D, Vandepoele K, Simillon C, van de Peer Y (2004) Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol 12:148–154

Gil R, Silva FJ, Pereto J, Moya A (2004) Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev 68:518–537

Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19:2226–2238

Goodfellow M, O'Donnell AG (1993) Handbook of new bacterial systematics. Academic Press, London

Goodfellow M, Manfio GP, Chun J (1997) Towards a practical species concept for cultivable bacteria. In: Claridge MD, Dawah HA, Wilson MR (eds) Species: the units of diversity. Chapman and Hall, London, pp 25–59

Goodfellow M, Kämpfer P, Busse HJ, Trujillo M, Suzuki K-I, Ludwig W, Whitman WB (2011). Bergey's Manual of Systematic Bacteriology, 2nd Edn, Vol 5, The *Actinobacteria*, Springer, New York (in press)

Gupta RS (2001) The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. Int Microbiol 4:187–202

Gupta RS, Griffiths E (2002) Critical issues in bacterial phylogeny. Theor Popul Biol 61:423–434

Gupta RS, Pereira M, Chandrasekera C, Johari V (2003) Molecular signatures in protein sequences that are characteristic of cyanobacteria and plastid homologues. Int J Syst Evol Microbiol 53:1833–1842

Harrington CS, On SL (1999) Extensive 16S rRNA gene sequence diversity in *Campylobacter hyointestinalis* strains: taxonomic and applied implications. Int J Syst Bacteriol 49:1171–1175

Hittinger CT, Carroll SB (2007) Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449:677–681

Hong SH, Kim TY, Lee SY (2004) Phylogenetic analysis based on genome-scale metabolic pathway reaction content. Appl Microbiol Biotechnol 65:203–210

Hooper SD, Berg OG (2003) On the nature of gene innovation: duplication patterns in microbial genomes. Mol Biol Evol 20:945–954

House CH, Fitz-Gibbon ST (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. J Mol Evol 54:539–547

Huang J, Gogarten JP (2006) Ancient horizontal gene transfer can benefit phylogenetic reconstruction. Trends Genet 22:361–366

Hugenholtz P, Goebel BM, Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol 180:4765–4774

Hull DL (1997) The ideal species concept-and why we can't get it. In: Claridge MF, Dawah HA, Wilson MR (eds) Species: the units of biodiversity. Chapman and Hall, London, pp 357–380

Huson DH, Steel M (2004) Phylogenetic trees based on gene content. Bioinformatics 20:2044–2049

Huynen MA, Bork P (1998) Measuring genome evolution. Proc Natl Acad Sci USA 95:5849–5856

Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11:97–108

Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? Trends Genet 22:225–231

Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Res 11:555–565

Klenk HP, Goker M (2010) En route to a genome-based classification of *Archaea* and *Bacteria*? Syst Appl Microbiol 33:175–182

Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP et al (2008) Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc Natl Acad Sci USA 105:2504–2509

Kolsto AB (1997) Dynamic bacterial genome organization. Mol Microbiol 24:241–248

Konstantinidis KT, Tiedje JM (2005) Towards a genome-based taxonomy for prokaryotes. J Bacteriol 187:6258–6264

Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 1:127–136

Korbel JO, Snel B, Huynen MA, Bork P (2002) SHOT: a web server for the construction of genome phylogenies. Trends Genet 18:158–162

Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, Ludwig W, Whitman WB (2010) Bergey's Manual of Systematic Bacteriology, 2nd Edition, Volume 4, The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gennatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae and Planctomycetes, Springer, USA

Kunin V, Ouzounis CA (2003) The balance of driving forces during genome evolution in prokaryotes. Genome Res 13:1589–1594

Kunin V, Ahren D, Goldovsky L, Janssen P, Ouzounis CA (2005) Measuring genome conservation across taxa: divided strains and united kingdoms. Nucleic Acids Res 33:616–621

Kunisawa T (1995) Identification and chromosomal distribution of DNA sequence segments conserved since divergence of *Escherichia coli* and *Bacillus subtilis*. J Mol Evol 40:585–593

Lawrence JG, Retchless AC (2009) The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. Methods Mol Biol 532:29–53

Lawrence JG, Retchless AC (2010) The myth of bacterial species and speciation. Biol Philos 25:569–588

Lin J, Gerstein M (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res 10:808–818

Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SN, Wang Q, Reeves PR, Wang L (2008) Structure and genetics of *Shigella* O antigens. FEMS Microbiol Rev 32:627–653

Lopez P, Bapteste E (2009) Molecular phylogeny: reconstructing the forest. C R Biol 332:171–182

Lucker S, Wagner M, Maixner F, Pelletier E, Koch H, Vacherie B, Rattei T, Damste JS, Spieck E, Le Paslier D, Daims H (2010) A *Nitrospira* metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. Proc Natl Acad Sci USA 107:13479–13484

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

Ma HW, Zeng AP (2004) Phylogenetic comparison of metabolic capacities of organisms at genome level. Mol Phylogenet Evol 31:204–213

Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S et al (2011) Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 9:467–477

Mayden RL (1997) A hierarchy of species concepts: the denouement in the saga of the species problem. In: Claridge MF, Dawah HA, Wilson MR (eds) Species: the units of biodiversity. Chapman and Hall, London, pp 381–382

Mayr E (1970) Populations, species and evolution. Harvard University Press, Cambridge

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15:589–594

Meglitsch PA (1954) On the nature of species. Syst Zool 3:491–503

Metzker ML (2010) Sequencing technologies: the next generation. Nat Rev Genet 11:31–46

Mira A, Klasson L, Andersson SG (2002) Microbial genome evolution: sources of variability. Curr Opin Microbiol 5:506–512

Mira A, Martin-Cuadrado AB, D'Auria G, Rodriguez-Valera F (2010) The bacterial pan-genome: a new paradigm in microbiology. Int Microbiol 13:45–57

Mocali S, Benedetti A (2010) Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. Res Microbiol 161:497–505

Monot M, Honore N, Garnier T, Zidane N, Sherafi D, Paniz-Mondolfi A, Matsuoka M et al (2009) Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. Nat Genet 41:1282–1289

Moran NA (2003) Tracing the evolution of gene loss in obligate bacterial symbionts. Curr Opin Microbiol 6:512–518

Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M et al (2010) *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet 42:1140–1143

Morschhauser J, Kohler G, Ziebuhr W, Blum-Oehler G, Dobrindt U, Hacker J (2000) Evolution of microbial pathogens. Philos Trans R Soc Lond B Biol Sci 355:695–704

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Okura M, Osawa R, Tokunaga A, Morita M, Arakawa E, Watanabe H (2008) Genetic analyses of the putative O and K antigen gene clusters of pandemic *Vibrio parahaemolyticus*. Microbiol Immunol 52:251–264

O'Malley MA (2007) The nineteenth century roots of 'everything is everywhere'. Nat Rev Microbiol 5:647–651

Pace NR (2009) Mapping the tree of life: progress and prospects. Microbiol Mol Biol Rev 73(4):565–576

Pearson T, Okinaka RT, Foster JT, Keim P (2009) Phylogenetic understanding of clonal populations in an era of whole genome sequencing. Infect Genet Evol 9:1010–1019

Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, Jin Z, Lee P et al (2010) Diversity of 16S rRNA genes within individual prokaryotic genomes. Appl Environ Microbiol 76:3886–3897

Pena A, Teeling H, Huerta-Cepas J, Santos F, Yarza P, Brito-Echeverria J, Lucio M et al (2010) Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting *Salinibacter ruber* strains. ISME J 4:882–895

Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J (2009) Metagenomic pyrosequencing and microbial identification. Clin Chem 55:856–866

Philippe H, Douady CJ (2003) Horizontal gene transfer and phylogenetics. Curr Opin Microbiol 6:498–505

Planet PJ, Sarkar IN (2005) mILD: a tool for constructing and analyzing matrices of pairwise phylogenetic character incongruence tests. Bioinformatics 21:4423–4424

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res 21:599–609

Poptsova M (2009) Testing phylogenetic methods to identify horizontal gene transfer. Methods Mol Biol 532:227–240

Priest FG, Williams ST (1993) Computer-assisted identification. In: Goodfellow M, O'Donnell AG (eds) Handbook of bacterial systematics. Academic Press, London, pp 362–381

Rannala B, Yang Z (2008) Phylogenetic inference using whole genomes. Annu Rev Genomics Hum Genet 9:217–231

Rocha EP (2004) Order and disorder in bacterial genomes. Curr Opin Microbiol 7:519–527

Rosenberg MS, Kumar S (2003) Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. Mol Biol Evol 20:610–621

Rosselló-Mora R (2003) Opinion: the species problem, can we achieve a universal concept? Syst Appl Microbiol 26:323–326

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

Schleifer KH (2009) Classification of *Bacteria* and *Archaea*: past, present and future. Syst Appl Microbiol 32:533–542

Schleifer KH, Stackebrandt E (1983) Molecular systematics of prokaryotes. Annu Rev Microbiol 37:143–187

Schouls LM, Schot CS, Jacobs JA (2003) Horizontal transfer of segments of the 16S rRNA genes between species of the *Streptococcus anginosus* group. J Bacteriol 185:7241–7246

Simpson GG (1961) Principles of animal taxonomy, Columbia University Press, New York

Sneath PHA (1992) International code of nomenclature of bacteria (bacteriological code 1990 revision). American Society of Microbiology, Washington

Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. Nat Genet 21:108–110

Snel B, Huynen MA, Dutilh BE (2005) Genome trees and the nature of genome evolution. Annu Rev Microbiol 59:191–209

Soria-Carrasco V, Castresana J (2008) Estimation of phylogenetic inconsistencies in the three domains of life. Mol Biol Evol 25:2319–2329

Stackebrandt E, Goebel BM (1994) Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol 44:846–849

Stackebrandt E, Frederiksen W, Garrity GM, Grimont PA, Kämpfer P, Maiden MC 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

Staley J (2009) The phylogenomic species concept. Microbiology Today, May 09:80–83

Sun Y, Wang M, Liu H, Wang J, He X, Zeng J, Guo X, Li K, Cao B, Wang L (2011) Development of an O-antigen serotyping scheme for *Cronobacter sakazakii*. Appl Environ Microbiol 77:2209–2214

Sutcliffe IC (2010) A phylum level perspective on bacterial cell envelope architecture. Trends Microbiol 18:464–470

Suyama M, Bork P (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet 17:10–13

Switt AI, Soyer Y, Warnick LD, Wiedmann M (2009) Emergence, distribution, and molecular and phenotypic characteristics of Salmonella enterica serotype 4, 5, 12:i:-. Foodborne Pathog Dis 6:407–415

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA 102:13950–13955

Tindall BJ, Rosselló-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

Ueda K, Seki T, Kudo T, Yoshida T, Kataoka M (1999) Two distinct mechanisms cause heterogeneity of 16S rRNA. J Bacteriol 181:78–82

Vandamme P, Pot B, Gillis M, de Vos P, Kersters K, Swings J (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev 60:407–438

Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF, van Sinderen D (2007) Genomics of *Actinobacteria*: tracing the evolutionary history of an ancient phylum. Microbiol Mol Biol Rev 71:495–548

Vos M (2011) A species concept for bacteria based on adaptive divergence. Trends Microbiol 19(1):1–7

Wang Q, Torzewska A, Ruan X, Wang X, Rozalski A, Shao Z, Guo X, Zhou H, Feng L, Wang L (2010) Molecular and genetic analyses of the putative *Proteus* O antigen gene locus. Appl Environ Microbiol 76:5471–5478

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

Wolf YI, Brenner SE, Bash PA, Koonin EV (1999) Distribution of protein folds in the three superkingdoms of life. Genome Res 9:17–26

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol 1:8

Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees and the tree of life. Trends Genet 18:472–479

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ et al (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462:1056–1060

Zhao W, Zhong Y, Yuan H, Wang J, Zheng H, Wang Y, Cen X, Xu F, Bai J, Han X et al (2010) Complete genome sequence of the rifamycin SV-producing *Amycolatopsis mediterranei* U32 revealed its genetic characteristics in phylogeny and metabolism. Cell Res 20:1096–1108

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theoret Biol 8:357–366