Antonie van Leeuwenhoek
DOI 10.1007/s10482-011-9663-1

PERSPECTIVE

Microbial systematics in the post-genomics era

Beile Gao · Radhey S. Gupta

Received: 1 September 2011 / Accepted: 15 October 2011

Abstract Microbial systematics and phylogeny should form the foundation and guiding light for a comprehensive understanding of different aspects of microbiology. However, there are many critical issues in microbial systematics that are currently not resolved. Some of these include: how to define and delimit a prokaryotic species; development of rationale criteria for the assignment of higher taxonomic ranks; understanding what unique properties distinguish species from different groups; and understanding the branching order and interrelationship among higher prokaryotic clades. The sequencing of genomes from large numbers of cultured as well as uncultured microbes covering prokaryotic diversity provides unique means to achieve these important objectives. Prokaryotic genomes are found to be very diverse and dynamic and horizontal gene transfers (HGTs) are indicated to have played important role in species/genome evolution. Although HGT adds a layer of complexity in terms of understanding the genomes and species evolution, it is contended that vast majority of genes and genetic characteristics that are distinctive characteristics of higher prokaryotic taxa are vertically inherited and based on them a solid foundation for microbial systematics can be developed. We describe two kinds of molecular markers consisting of conserved indels in protein sequences and whole proteins that are specific for different groups that are proving particularly valuable in defining different prokaryotic groups in clear molecular terms and in understanding their interrelationships. The genetic and biochemical studies on these taxa-specific molecular markers also open the way to discover novel biochemical and physiological characteristics that are unique properties of these groups.

Keywords Microbial phylogeny · Bacterial systematics · Molecular markers · Conserved indels · Conserved signature proteins · Higher taxonomic clades · Horizontal gene transfer

Introduction

Since the first bacterial genome of *Haemophilus influenzae* was published in 1995, the number of complete genomes has increased at an exponential pace (Fleischmann et al. 1995; NCBI database 2011). Even at the very beginning of genome sequencing projects, the concept of "post-genomics era" was anticipated, which indicated its expected big influence on biological research (Gershon 1997; Fraser-Liggett 2005). Although it is only 15 years since the first

B. Gao
Section of Microbial Pathogenesis, Yale University
School of Medicine, New Haven, CT 06536, USA

R. S. Gupta (✉)
Department of Biochemistry and Biomedical Sciences,
McMaster University, Hamilton L8N3Z5, Canada
e-mail: gupta@mcmaster.ca

Published online: 03 November 2011

genome was sequenced, the development of deep-sequencing technologies has flooded the genome databases with either pure-culture bacterial genomes that have taxonomic names or mixtures of cultured and uncultured microbes from certain environments/niches (Venter et al. 2004; Tringe et al. 2005; Delong et al. 2006; Xie et al. 2011). Hence, we are able to explore the microbial community more deeply and much faster from different perspectives including their ecological diversity, niche adaptation, ability to produce diverse natural products, pathogenic potential, presence in human microbiota, etc. (Dinsdale et al. 2008; Tringe et al. 2005; Pallen and Wren 2007; Arumugam et al. 2011). All of these studies depend upon and will greatly benefit from a sound framework of microbial classification (i.e., microbial systematics) and phylogeny.

As the genome sequencing data have accumulated, many studies have been carried out to investigate the phylogenetic relationships of different prokaryotes by comparative genomics approaches; the most popular of these methods include examining the gene order, shared gene content, construction of supertrees, etc. (Belda et al. 2005; Snel et al. 1999; Lathe et al. 2000; Beiko et al. 2005; Ding et al. 2008; Ciccarelli et al. 2006). These studies certainly facilitate our understanding of microbial systematics in the light of genome evolution (Koonin 2009; Kunin et al. 2005; Philippe et al. 2005). However, the results from these studies also challenge the present framework because the prokaryotic genomes are found to be dynamic and plastic, showing much more diversity than what we knew from their phenotypic or other genotypic characteristics such as the 16S rRNA gene sequences (Gogarten et al. 2002; Snel et al. 2005; Lawrence and Hendrickson 2005; Gogarten and Townsend 2005). Especially with the large number of cases of horizontal gene transfer (HGT) identified from genome sequences, a Darwinian tree-like representation of relationships between species has been questioned and a network of species has been proposed (Doolittle 1999; Kunin et al. 2005). However, since the detection of HGTs between species also depends upon the current phylogenetic/systematic framework, a sound understanding of the evolutionary relationships among different species is essential to accurately determine the incidence of HGTs. Before discussing the current state of microbial systematics and some of the important issues that needs to be understood in this regard, it would be helpful to revisit the concept of HGT in a little more detail and critical manner.

Influence of horizontal gene transfers on genome and species evolution

To determine whether the HGTs truly diminish the tree of life, two questions need be answered: (i) What is the extent that HGTs affect the prokaryotic genomes and (ii) whether a core genome still exists that is significantly not affected by HGTs (Snel et al. 2005). These two questions can be answered together. The extent of HGT is currently hotly debated, and due to different species sampling and detection methods/standards, a bacterial genome is suggested to have 0 to >20% genes obtained from outsources (Ochman et al. 2000; Nakamura et al. 2004; Ragan 2001). However, among these alien genes, some have detected homologues in other species so the transfer is evidenced, whereas many other genes are only found in particular genomes (Abby and Daubin 2007; Lerat et al. 2005). These later genes constitute a large fraction of the currently identified HGT products. However, due to their lack of homologues in other species, it is quite possible that such genes may have originated in those specific genomes. A number of comparative genomic studies have been carried out to carefully examine the influence of HGT events. For example, Novichkov et al. 2004 described a framework for identifying orthologous sets of genes that deviate from a clock-like model of evolution. For several hundred analyzed orthologous sets representing three well-defined bacterial lineages, they found that 70% of the genes were not affected by HGT, 15% of them showed anomalous behavior due to lineage-specific acceleration of evolution, while the remaining were probably caused by HGTs (Novichkov et al. 2004). Kunin et al. analyzed a 165-genome dataset and found 4.7–5.2% of events to be HGTs, 11.1–11.6% gene losses and 83.4–83.6% vertical transfers (Kunin et al. 2005; Kunin and Ouzounis 2003). Additionally, Beiko et al. 2005) have performed a rigorous phylogenetic analysis of >220,000 proteins from 144 prokaryotic genomes to determine the contribution of gene sharing to current prokaryotic diversity, and the inferred relationships suggest a pattern of inheritance that is largely vertical except among some closely related taxa and among some species that live in similar environments .

It is known that genes involved in translation and transcription show fewer indications of transfers (Koonin 2003). For example, the 31 orthologous genes employed by Ciccarelli et al. 2006 for construction of a universal phylogenetic tree are all involved in translation (Ciccarelli et al. 2006; Oren 2010). These proteins are highly connected in the cellular network, less exposed to immediate selective pressure and thus less susceptible to homologous replacement via HGT (Ragan and Beiko 2009; Aris-Brosou 2005). Although some studies have detected HGT events for some core genes, including the 16S rRNA, such cases are very few and the number of publications reporting them is countable (Gogarten and Townsend 2005; Gogarten et al. 2002; Oren 2010). Besides, single-gene based trees are already questioned and the current trend is to use a core set of non-transferred or rarely transferred genes to track the evolutionary history of prokaryotes (Ciccarelli et al. 2006; Williams et al. 2010; Gupta and Mathews 2010; Horiike et al. 2009).

These studies indicate that the concept or definition of HGT needs to be revised to take into consideration the evolutionary processes by which genomes and new species evolve. It is known that not all genes were inherited from the last universal common ancestor of life forms. Although the mechanisms by which new genes evolve are not known, it is believed that gene transfer is an important source of genome expansion throughout the evolutionary process (Daubin and Ochman 2004; Lerat et al. 2005). Gene transfer provides the bacterial genome with a new set of genes that helps it to explore and adapt to new ecological niches (Kuo and Ochman 2009; Stackebrandt et al. 2002; Oren 2010). If the gene transfers occurred at a deeper clade level and the new genes are retained by all the descendants from the progenitor, then the gene transfer events have likely contributed to the divergence of the clade and the new genes are already incorporated into the cellular protein interaction network (Lerat et al. 2005; Narra et al. 2008; Ragan and Beiko 2009). Besides, the genes acquired via lateral gene transfer over time get ameliorated and many of them exhibit little or no similarity to the original genes and they come to resemble the native genes with regard to characteristics such as their GC content or codon usage (Lawrence and Ochman 1997; Marri and Golding 2008; Koski et al. 2001). Thus, these "new genes", which could have been acquired by means of ancient gene transfers, actually record the divergence of the clades or lineages. Importantly, after introduction into the progenitor cell, these new genes follow a vertical inheritance pattern, which is different from the current concept of HGTs or LGTs. Hence, the genes which are restricted to specific lineages and passed on by vertical inheritance should not be regarded as "horizontal" or "lateral" gene transfers. Rather than randomly obscuring prokaryotic phylogeny, these genes actually promote and record the divergence of the species via the introduction of new genes at different evolutionary stages.

In summary, although the HGTs add an extra layer of complexity to the study of species evolution, they do not seriously affect reconstruction of the evolutionary history of life as most of the genes are still vertically inherited (Abby and Daubin 2007; Snel et al. 2005). Caution should also be exercised in extending the concept of HGT to the evolution of novel genes, as the mechanisms that give rise to them are not fully understood. Additionally, the gene transfer events that occur at different evolutionary stages tell us different stories about the evolution of species.

Current issues in microbial systematics

Apart from the noise and complexity that is introduced by HGTs in phylogenetic studies, there are many critical issues in microbial systematics that are currently not resolved. The most debated of these issues is how to define a species (Konstantinidis et al. 2006; Staley 2006; Fraser et al. 2009; Oren 2010; Stackebrandt et al. 2002). As a fundamental unit in the hierarchy of prokaryote classification, the development of a sound "species" concept is of particular importance. Since 1987, bacterial strains exhibiting >70% whole-cell DNA–DNA hybridization and sharing at least one distinctive phenotype are considered to be members of the same species (Wayne 1988). The above value for DNA–DNA hybridization roughly corresponds to ~97% rRNA sequence identity and this criterion is also commonly employed for species identification purposes (Brenner et al. 2005; Goris et al. 2007; Stackebrandt et al. 2002). However, in many cases, different species and even genera are found to exhibit >70% DNA–DNA similarity and yet for practical reasons they are regarded as different species or genera (Ludwig and Klenk 2005; Gevers et al. 2005; Oren 2010). Importantly, different strains of the same species are also found to differ greatly in terms of their genome sequences (Tettelin et al. 2005; Lukjancenko et al. 2010; Alcaraz et al. 2010). A recent detailed study on 44 *Streptococcus pneumoniae* genomes indicated that while about 74% of the genes were present in most strains, the remaining 21–32% genes (non-core) were restricted to different clusters (Donati et al. 2010). The sum of both core and non-core genes from different strains of a given species is now referred to as the "pan-genome" (Tettelin et al. 2008). The non-core genes are postulated to contribute to functions such as niche adaptation, antibiotic resistance and the ability to colonize new hosts (Kuo and Ochman 2009; Narra et al. 2008; Coleman and Chisholm 2010; Tettelin et al. 2008). Based upon their branching in phylogenetic trees, genomic arrangement and uniquely shared genes/proteins (Touchon et al. 2009; Liu et al. 1999; Gupta, unpublished results), different strains of some species can be divided into a number of distinct clades. This raises the questions what taxonomic rank should be assigned to strains from these clades and whether the current species definition is too broad and masks the diversity that exists within the prokaryotes. Although it has been suggested that new methods should be applied to define a prokaryotic species (Stackebrandt et al. 2002; Staley 2006; Fraser et al. 2009), due to lack of reliable means to define a species, no general agreement has been reached in this regard.

In contrast to the species level where a formal, although inadequate, definition exists, there are no agreement upon criteria for defining the higher taxonomic ranks within prokaryotes and all such rank assignments are based upon almost entirely arbitrary considerations (Stackebrandt 2006; Oren 2010; Stackebrandt et al. 2002; Ludwig and Klenk 2005). The arbitrariness of the present bacterial classification is well illustrated by the example of the phylum proteobacteria. The proteobacteria comprise the largest group within prokaryotes accounting for nearly 50% of all cultured bacteria (Ludwig and Klenk 2005; Maidak et al. 2001; Kersters et al. 2006; Gupta 2000b). Based upon their branching in the 16S rRNA trees, they are divided into five classes, named alpha-, beta-, gamma-, delta- and epsilon-proteobacteria (Ludwig and Klenk 2005; Maidak et al. 2001; Kersters et al. 2006; Garrity et al. 2005). Of these, alpha-, beta-, and gamma-proteobacteria harbor approximately 12, 8, and 26% of all cultured bacteria (Maidak et al. 2001). The species from these groups can also be clearly distinguished from each other and from all other bacteria based upon large numbers of molecular characteristics (Gupta 2000b, 2005, 2006; Gupta and Sneath 2007; Gupta and Mok 2007; Gao et al. 2009; Kersters et al. 2006; Ciccarelli et al. 2006). However, despite their phylogenetic and molecular distinctness, these large groups of bacteria are presently not recognized as distinct phyla, whereas numerous other poorly studied bacteria consisting of only few species are recognized as separate phyla of bacteria.

In the current taxonomic scheme, based upon branching pattern in the 16S rRNA trees, the relationships among various higher taxonomic clades are also generally not resolved. Thus, it is difficult to determine how different groups are related to each other or how they evolved from a common ancestor (Ludwig and Klenk 2005; Woese 2006; Gupta and Griffiths 2002; Gupta and Gao 2010). Additionally, for most of the prokaryotic groups of higher taxonomic ranks, except for their branching pattern in the phylogenetic trees, no molecular, biochemical or physiological characteristics are known that are specific for these groups and can be used to distinguish them from all others. Considering that systematics should ideally serve as the foundation and guide map for microbiological studies, in addition to indicating that a particular group of prokaryotes form a distinct clade in phylogenetic trees, it should be able to specify more of their commonly shared and unique characteristics. Hence, it is important to develop more reliable criteria to define and delimit the higher taxonomic ranks within the prokaryotes and also develop means to identify biochemical or physiological characteristics that are specific for different groups of prokaryotes. This should lead to the development of a more comprehensive and reliable systematics of prokaryotes that should be able to serve the guiding role in microbiology.

Conserved indels and lineage-specific proteins as novel tools for microbial systematics

The current unresolved issues regarding prokaryotic phylogeny and systematics make it necessary to search for novel characteristics that are unique to different prokaryotic lineages and also record their divergence from common ancestor. The characteristics that are ideally suited for such studies should meet the following requirements: "*These markers should be homologous apomorphic characters that evolved only once (synapomorphy) and not by convergence*" (Stackebrandt 2006; Gupta 1998; Gupta and Griffiths 2002). Such markers should also not be affected or minimally affected by factors such as multiple changes at a given site, long-branch attraction effects, differences in evolutionary rates between and among species, HGTs, etc., which confound the inferences from phylogenetic trees (Delsuc et al. 2005; Philippe et al. 2005; Gupta 1998).

Conserved inserts and deletions (indels) in gene/proteins sequences provide an important category of rare genetic changes (RGCs) for understanding bacterial phylogeny (Gupta 1998; Rokas and Holland 2000; Delsuc et al. 2005; Gupta and Griffiths 2002). Those indels which provide useful phylogenetic markers are generally of defined size and flanked on both sides by conserved regions to ensure their reliability (Gupta and Griffiths 2002; Gupta 1998). Because of the rarity and highly specific nature of such changes, it is less likely that they could arise independently by either convergent or parallel evolution (i.e., homoplasy) (Gupta 2000a; Rokas and Holland 2000). Other confounding factors such as differences in evolutionary rates at different sites or among different species should also not affect the interpretation of a conserved indel. Hence, when a conserved signature indel (CSI) of defined size is uniquely found in a phylogenetically defined group(s) of species, the simplest explanation for this observation is that the genetic change responsible for this CSI occurred once in a common ancestor of this group of species and then passed on vertically to the various descendents. Because the presence or absence of a given CSI in different species is not affected by factors such as differences in evolutionary rates, CSIs which are restricted to particular clade(s) have generally provided very good phylogenetic markers of common evolutionary descent (Gupta 1998; Gupta 2003; Lake et al. 2007). Also, since genetic changes leading to CSIs could be introduced at various stages during evolution, it is possible to identify CSIs in gene/protein sequences at different phylogenetic depths corresponding to various higher taxonomic groupings (e.g. phylum, order, family, genus and even single species and subspecies levels) (Gupta 2001; Gupta and Griffiths 2002; Gupta 1998; Gupta and Gao 2010; Gao and Gupta 2005; Ahmod et al. 2011). Such CSIs, in turn, can provide well-defined markers for identifying different taxonomic groups of bacteria in molecular terms. Recent work from our lab has identified a large number of CSIs that are restricted to many higher taxonomic groups within the prokaryotes, such as: alpha-proteobacteria, gamma-proteobacteria, epsilon-proteobacteria, Aquifiales, Chlamydia, Cyanobacteria, Deinococcus–Thermus, Bacteroidetes-Chlorobi, Actinobacteria, Thermotogae, Archaea, etc. (Gupta 2009; Gao et al. 2009; Griffiths and Gupta 2004a, 2004b, 2001, 2006; Griffiths et al. 2005; Gupta 1998, 2004, 2010; Gupta and Bhandari 2011; Gao and Gupta 2005; Gupta and Shami 2011; Naushad and Gupta 2011). These newly discovered CSIs provide useful markers for defining or circumscribing the above prokaryotic groups in clear molecular terms. Additionally, identified CSIs that are commonly shared by species from a number of different phyla provide valuable information regarding the branching order and interrelationships among different main groups of prokaryotes (Gupta 2001, 2003, 2000a, 2010, 2009; Gupta and Mok 2007). With the greatly expanded microbial genome database, the statistical study of large numbers of such RGCs certainly represents a promising avenue for unraveling the prokaryotic phylogeny.

Another type of RGC that can be useful for taxonomic classification as well as for understanding evolutionary relationships among different organisms are whole proteins that are uniquely present in particular groups or subgroups of prokaryotes but not found anywhere else (Kainth and Gupta 2005; Dutilh et al. 2008). Recent analyzes of genomic sequences have indicated that such conserved signature proteins (CSPs), which are also referred to as lineage-specific proteins, arise throughout the evolutionary process of a bacterial lineage (Gao and Gupta 2007; Lerat et al. 2005; Gupta and Mathews 2010). A vast number of lineage-specific proteins unique to certain species, strain or even genome, which are also called "ORFans", are introduced recently during speciation or strain divergence (Daubin and Ochman 2004). Studies have shown that these proteins present at the tips of the phylogeny evolve fast and are subject to loss if not conferring advantages to the host (Narra et al. 2008; Kuo and Ochman 2009). However, if the lineage-specific proteins originate deep within a clade and are retained by all the descendents from the progenitor, they are confined to the monophyletic group (Gao et al. 2006; Dutilh et al. 2008; Gupta and Gao 2010; Gupta and Mathews 2010). Thus, these proteins are no more solitary "orphans", but they are conserved signature proteins (CSPs), which are uniquely shared by every daughter lineage of that group and they provide useful molecular markers for defining or distinguishing that group from other bacteria (Gupta and Gao 2009; Gao et al. 2009; Gupta and Gao 2010). Furthermore, based on a number of CSPs that are specific to different lineages, it is possible to infer their branching order or interrelationship (Gupta and Mok 2007; Kainth and Gupta 2005; Gupta and Griffiths 2006; Gupta and Mathews 2010; Gupta and Gao 2010; Gupta 2010).

Similar to CSIs, comparative genomic studies have been carried out on several major prokaryotic phyla to identify CSPs that are unique to them, such as alphaproteobacteria, gamma-proteobacteria, epsilon-proteobacteria, Chlamydia, Cyanobacteria, Deinococcus–Thermus, Bacteroidetes-Chlorobi, Actinobacteria, Archaea, etc. (Kainth and Gupta 2005; Gao et al. 2009; Gupta 2006, 2009; Griffiths et al. 2006; Griffiths and Gupta 2007; Gupta and Lorenzini 2007; Gupta and Mok 2007; Gupta and Shami 2011). The identified CSPs unique to different prokaryotic groups have proved of great value in defining these major groups and have also provided useful information regarding the branching order of different lineages within them. Interestingly, a majority of identified CSPs are of hypothetical functions, which points to our lack of knowledge regarding many of the building blocks in the prokaryotic cell (Gupta and Gao 2010). Studies on these lineage-specific CSPs that originate at deeper clade levels are very meaningful for the following reasons: First, because of their retention in all daughter lineages, these proteins must perform important functions in species from these clades. For example, recent studies on the species distribution of key lipopolysaccharide (LPS) biosynthesis enzymes (Sutcliffe 2010) and a number of CSIs across different bacterial phyla have provided important insight concerning the evolution of the LPS-containing outer cell membrane, which is a defining characteristic of archetypical Gram-negative bacteria (Gupta 2011). Second, due to their uniqueness, their functions likely specify some distinctive characteristics that make the clade different from other bacteria. Third, a thorough understanding of their evolution as individual and components in the protein interaction network should provide insight into the mechanisms of genesis or speciation of new bacterial species and clades (Daubin and Ochman 2004; Kuo and Ochman 2009). Moreover, it is arguable that maintenance of particular CSP/CSI in a genome over countless generations is in itself a significant phenotype, in the sense that it is the expressed result of faithful replication under natural selection. Clearly maintenance of these sequences (either of CSI or of CSP) is likely to have phenotypic consequences, as has been demonstrated for CSI in the Hsp60 and Hsp70 proteins of *E. coli* (Singh and Gupta 2009).

Future directions

In order to better understand microbial systematics, it is important to map molecular characteristics such as CSIs and CSPs on to the phylogenetic tree. These markers not only provide additional evidence for the genetic or phylogenetic relatedness of different prokaryotic groups, but also provide new targets/tools to study the biology of these microbes. Although the GenBank currently has >1,700 complete genomes from different microbes, they are somewhat biased in terms of taxonomic sampling toward bacterial taxa that are either important pathogens or are important from biotechnology standpoints. However, the recent project of phylogeny-driven genomic encyclopedia of bacteria and archeae (GEBA) (Wu et al. 2009; Klenk and Goker 2010) should lead to sequencing of diverse prokaryotic genomes that should enable identification of more molecular markers for different groups and also provide the necessary means to rigorously test the specificity of these markers. The cultured bacteria or archaea represent only about 1% of the total microbial diversity (Amann et al. 1995; Delong and Pace 2001). Although we do not have reliable means to study the uncultured microbes, the metagenomics data from different environments (Tringe et al. 2005; Xu 2006; Turnbaugh and Gordon 2008; Gianoulis et al. 2009) have opened up new windows to explore microbial diversity in these environments. The CSIs and CSPs that are specific for different prokaryotic groups provide valuable tools for determining the presence or absence of species related to these groups in different metagenomic samples. The availability of increasing numbers of genomic sequences covering the depth and breadth of prokaryotic species, in conjunction with novel and more specific means to identify different prokaryotic groups at various taxonomic levels, such as CSIs and CSPs, bodes well for the future prospects of developing a stable and comprehensive foundation for microbial systematics.

Acknowledgments This work was supported by a research grant from the Natural Science and Engineering Research Council of Canada.

References

Abby S, Daubin V (2007) Comparative genomics and the evolution of prokaryotes. Trends Microbiol 15:135–141

Ahmod NZ, Gupta RS, Shah HN (2011) Identification of a *Bacillus anthracis* specific indel in the yeaC gene and development of a rapid pyrosequencing assay for distinguishing *B. anthracis* from the *B. cereus* group. J Microbiol Methods (in press)

Alcaraz LD, Moreno-Hagelsieb G, Eguiarte LE, Souza V, Herrera-Estrella L, Olmedo G (2010) Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. BMC Genomics 11:332

Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial-cells without cultivation. Microbiol Rev 59:143–169

Aris-Brosou S (2005) Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. Mol Biol Evol 22:200–209

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin JJ, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Dore J, Weissenbach J, Ehrlich SD, Bork P (2011) Enterotypes of the human gut microbiome. Nature 473:174–180

Beiko RG, Harlow TJ, Ragan MA (2005) Highways of gene sharing in prokaryotes. Proc Natl Acad Sci USA 102:14332–14337

Belda E, Moya A, Silva FJ (2005) Genome rearrangement distances and gene order phylogeny in gamma-proteobacteria. Mol Biol Evol 22:1456–1467

Brenner DJ, Staley JT, Krieg NR (2005) Classification of prokaryotic organisms and the concept of bacterial speciation. In: Brenner DJ, Krieg NR, Staley JT, Garrity GM (eds) Bergey's manual of systematic bacteriology, 2nd edn. Springer, Berlin, pp 27–32

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287

Coleman ML, Chisholm SW (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci USA 107:18634–18639

Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in *E.coli*. Genome Res 14:1036–1042

Delong EF, Pace NR (2001) Environmental diversity of bacteria and archaea. Syst Biol 50:470–478

Delong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311:496–503

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

Ding GH, Yu ZH, Zhao J, Wang Z, Li Y, Xing XB, Wang CA, Liu L, Li YX (2008) Tree of life based on genome context networks. PLoS One 3(10):e3357

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li LL, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan YJ, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F (2008) Functional metagenomic profiling of nine biomes. Nature 452:629–639

Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, Oggioni M, Hotopp JCD, Hu FZ, Riley DR, Covacci A, Mitchell TJ, Bentley SD, Kilian M, Ehrlich GD, Rappuoli R, Moxon ER, Masignani V (2010) Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol 11(10):R107

Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124–2128

Dutilh BE, Snel B, Ettema TJG, Huynen MA (2008) Signature genes as a phylogenomic tool. Mol Biol Evol 25:1659–1667

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, Mckenney K, Sutton G, Fitzhugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu LI, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, Mcdonald LA, Small KV, Fraser CM, Smith HO, Venter JC (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP (2009) The bacterial species challenge: making sense of genetic and ecological diversity. Science 323:741–746

Fraser-Liggett CM (2005) Insights on biology and evolution from microbial genome sequencing. Genome Res 15:1603–1610

Gao B, Gupta RS (2005) Conserved indels in protein sequences that are characteristic of the phylum actinobacteria. Int J Syst Evol Microbiol 55:2401–2412

Gao B, Gupta RS (2007) Phylogenomic analysis of proteins that are distinctive of archaea and its main subgroups and the origin of methanogenesis. BMC Genomics 8:86

Gao B, Paramanathan R, Gupta RS (2006) Signature proteins that are distinctive characteristics of actinobacteria and their subgroups. Antonie Van Leeuwenhoek 90:69–91

Gao B, Mohan R, Gupta RS (2009) Phylogenomics and protein signatures elucidating the evolutionary relationships among the gamma-proteobacteria. Int J Syst Evol Microbiol 59:234–247

Garrity GM, Bell JA, Lilburn T (2005) The revised road map to the manual. In: Brenner DJ, Krieg NR, Staley JT, Garrity GM (eds) Bergey's manual of systematic bacteriology, 2nd edn. Springer, Berlin, pp 159–187

Gershon D (1997) Bioinformatics in a post-genomics age. Nature 389:417–418

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL, Swings J (2005) Re-evaluating prokaryotic species. Nat Rev Microbiol 3:733–739

Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I, Yamada T, Paccanaro A, Jensen LJ, Snyder M, Bork P, Gerstein MB (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. Proc Natl Acad Sci USA 106:1374–1379

Gogarten JP, Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol 3:679–687

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

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

Griffiths E, Gupta RS (2001) The use of signature sequences in different proteins to determine the relative branching order of bacterial divisions: evidence that fibrobacter diverged at a similar time to Chlamydia and the Cytophaga-Flavobacterium-Bacteroides division. Microbiology-Sgm 147:2611–2622

Griffiths E, Gupta RS (2004a) Distinctive protein signatures provide molecular markers and evidence for the monophyletic nature of the Deinococcus-Thermus phylum. J Bacteriol 186:3097–3107

Griffiths E, Gupta RS (2004b) Signature sequences in diverse proteins provide evidence for the late divergence of the order Aquificales. Int Microbiol 7:41–52

Griffiths E, Gupta RS (2006) Molecular signatures in protein sequences that are characteristics of the phylum Aquificae. Int J Syst Evol Microbiol 56:99–107

Griffiths E, Gupta RS (2007) Identification of signature proteins that are distinctive of the Deinococcus-Thermus phylum. Int Microbiol 10:201–208

Griffiths E, Petrich AK, Gupta RS (2005) Conserved indels in essential proteins that are distinctive characteristics of Chlamydiales and provide novel means for their identification. Microbiology-Sgm 151:2647–2657

Griffiths E, Ventresca MS, Gupta RS (2006) BLAST screening of Chlamydial genomes to identify signature proteins that are unique for the Chlamydiales, Chlamydiaceae, Chlamydophila and Chlamydia groups of species. BMC Genomics 7:14

Gupta RS (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol Mol Biol Rev 62:1435–1491

Gupta RS (2000a) The natural evolutionary relationships among prokaryotes. Crit Rev Microbiol 26:111–131

Gupta RS (2000b) The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. FEMS Microbiol Rev 24:367–402

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 (2003) Evolutionary relationships among photosynthetic bacteria. Photosynth Res 76:173–183

Gupta RS (2004) The phylogeny and signature sequences characteristics of fibrobacteres, chlorobi, and bacteroidetes. Crit Rev Microbiol 30:123–143

Gupta RS (2005) Protein signatures distinctive of alpha-proteobacteria and its subgroups and a model for alpha-proteobacterial evolution. Crit Rev Microbiol 31:101–135

Gupta RS (2006) Molecular signatures (unique proteins and conserved indels) that are specific for the epsilon-proteobacteria (Campylobacterales). BMC Genomics 7:167

Gupta RS (2009) Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades. Int J Syst Evol Microbiol 59 1:2510–2526

Gupta RS (2010) Molecular signatures for the main phyla of photosynthetic bacteria and their subgroups. Photosynth Res 104:357–372

Gupta RS (2011) Origin of diderm (gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes. Antonie Van Leeuwenhoek 100:171–182

Gupta RS, Bhandari V (2011) Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups. Antonie Van Leeuwenhoek 100:1–34

Gupta RS, Gao B (2009) Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus *Clostridium* sensu stricto (cluster I). Int J Syst Evol Microbiol 59:285–294

Gupta RS, Gao B (2010) Recent advances in understanding microbial systematics. In: Xu JP (ed) Microbial population genetics. Caister Academic Press, Norfolk

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

Gupta RS, Griffiths E (2006) Chlamydiae-specific proteins and indels: novel tools for studies. Trends Microbiol 14:527–535

Gupta RS, Lorenzini E (2007) Phylogeny and molecular signatures (conserved proteins and indels) that are specific for the bacteroidetes and chlorobi species. BMC Evol Biol 7:71

Gupta RS, Mathews DW (2010) Signature proteins for the major clades of cyanobacteria. BMC Evol Biol 10:24

Gupta RS, Mok A (2007) Phylogenomics and signature proteins for the alpha-proteobacteria and its main groups. BMC Microbiol 7:106

Gupta RS, Shami A (2011) Molecular signatures for the crenarchaeota and the thaumarchaeota. Antonie Van Leeuwenhoek 99:133–157

Gupta RS, Sneath PHA (2007) Application of the character compatibility approach to generalized molecular sequence data: branching order of the proteobacterial subdivisions. J Mol Evol 64:90–100

Horiike T, Miyata D, Hamada K, Saruhashi S, Shinozawa T, Kumar S, Chakraborty R, Komiyama T, Tateno Y (2009) Phylogenetic construction of 17 bacterial phyla by new method and carefully selected orthologs. Gene 429:59–64

Kainth P, Gupta RS (2005) Signature proteins that are distinctive of alpha-proteobacteria. BMC Genomics 6:94

Kersters K, Devos P, Gillis M, Swings J, Vandamme P, Stackebrandt E (2006) Introduction to the proteobacteria. In: Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E (eds) The prokaryotes: a handbook on the biology of bacteria, 3rd edition, Release 3.12 edn. Springer, New York, pp 3–37

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

Konstantinidis KT, Ramette A, Tiedje JM (2006) The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci 361:1929–1940

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

Koonin EV (2009) Darwinian evolution in the light of genomics. Nucleic Acids Res 37:1011–1034

Koski LB, Morton RA, Golding GB (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol 18:404–412

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

Kunin V, Goldovsky L, Darzentas N, Ouzounis CA (2005) The net of life: reconstructing the microbial phylogenetic network. Genome Res 15:954–959

Kuo CH, Ochman H (2009) The fate of new bacterial genes. FEMS Microbiol Rev 33:38–43

Lake JA, Herbold CW, Rivera MC, Servin JA, Skophammer RG (2007) Rooting the tree of life using non-ubiquitous genes. Mol Biol Evol 24:130–136

Lathe WC, Snel B, Bork P (2000) Gene context conservation of a higher order than operons. Trends Biochem Sci 25:474–479

Lawrence JG, Hendrickson H (2005) Genome evolution in bacteria: order beneath chaos. Curr Opin Microbiol 8:572–578

Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44:383–397

Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. PLoS Biol 3:807–814

Liu SL, Schryvers AB, Sanderson KE, Johnston RN (1999) Bacterial phylogenetic clusters revealed by genome structure. J Bacteriol 181:6747–6755

Ludwig W, Klenk HP (2005) Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Brenner DJ, Krieg NR, Staley JT, Garrity GM (eds) Bergey's manual of systematic bacteriology. Springer, Berlin, pp 49–65

Lukjancenko O, Wassenaar TM, Ussery DW (2010) Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol 60:708–720

Maidak BL, Cole JR, Lilburn TG, Parker CT, Saxman PR, Farris RJ, Garrity GM, Olsen GJ, Schmidt TM, Tiedje JM (2001) The RDP-II (ribosomal database project). Nucleic Acids Res 29:173–174

Marri PR, Golding GB (2008) Gene amelioration demonstrated: the journey of nascent genes in bacteria. Genome 51:164–168

Nakamura Y, Itoh T, Matsuda H, Gojobori T (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet 36:760–766

Narra HP, Cordes MHJ, Ochman H (2008) Structural features and the persistence of acquired proteins. Proteomics 8:4772–4781

Naushad HS, Gupta RS (2011) Molecular signatures (conserved indels) in protein sequences that are specific for the order Pasteurellales and distinguish two of its main clades. Antonie Van Leeuwenhoek Epub ahead of print:

NCBI database (2011) NCBI completed microbial genomes. http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html. Ref Type: Generic

Novichkov PS, Omelchenko MV, Gelfand MS, Mironov AA, Wolf YI, Koonin EV (2004) Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J Bacteriol 186:6575–6585

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

Oren A (2010) Microbial systematics. In: Wang LK, Ivanov V, Jay JH (eds) Environmental biotechnology—handbook of environmental engineering. Springer, New York, pp 81–120

Pallen MJ, Wren BW (2007) Bacterial pathogenomics. Nature 449:835–842

Philippe H, Delsuc F, Brinkmann H, Lartillot N (2005) Phylogenomics. Annu Rev Ecol Syst 36:541–562

Ragan MA (2001) Detection of lateral gene transfer among microbial genomes. Curr Opin Genet Dev 11:620–626

Ragan MA, Beiko RG (2009) Lateral genetic transfer: open issues. Philos Trans R Soc Lond B Biol Sci 364:2241–2251

Rokas A, Holland PWH (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15:454–459

Singh B, Gupta RS (2009) Conserved inserts in Hsp60 (GroEL) and Hsp70 (DnaK) are essential for cellular growth. Mol Genet Genomics 281:361–373

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

Stackebrandt E (2006) Defining taxonomic ranks. In: Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E (eds) The prokaryotes. Springer, New York, pp 29–57

Stackebrandt E, Frederiksen W, Garrity GM, Grimont PAD, Kampfer P, Maiden MCJ, Nesme X, Rossello-Mora R, Swings J, Truper HG, Vauterin L, Ward AC, Whitman WB (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 52:1043–1047

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

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

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Ros IMY, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou LW, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA 102:13950–13955

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11:472–477

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, El Karoui M, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguenec C, Lescat M, Mangenot S, Martinez-Jehanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Saint Ruf C, Schneider D, Tourret J, Vacherie B, Vallenet D, Medigue C, Rocha EPC, Denamur E (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 5(1):e1000344

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005) Comparative metagenomics of microbial communities. Science 308: 554–557

Turnbaugh PJ, Gordon JI (2008) An invitation to the marriage of metagenomics and metabolomics. Cell 134:708–713

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu DY, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the sargasso sea. Science 304:66–74

Wayne LG (1988) International committee on systematic bacteriology: announcement of the report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Syst Appl Microbiol 10:99–100

Williams KP, Gillespie JJ, Sobral BWS, Nordberg EK, Snyder EE, Shallom JM, Dickerman AW (2010) Phylogeny of gammaproteobacteria. J Bacteriol 192:2305–2314

Woese CR (2006) How we do, don't and should look at bacteria and bacteriology. In: Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E (eds) The Prokaryotes. Springer, New York, pp 3–23

Wu DY, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng JF, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk HP, Eisen JA (2009) A phylogeny—driven genomic encyclopedia of bacteria and archaea. Nature 462:1056–1060

Xie W, Wang FP, Guo L, Chen ZL, Sievert SM, Meng J, Huang GR, Li YX, Yan QY, Wu S, Wang X, Chen SW, He GY, Xiao X, Xu AL (2011) Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. ISME J 5:414–426

Xu JP (2006) Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. Mol Ecol 15:1713–1731