Environmental Microbiology (2004) **6**(7), 754–759 doi:10.1111/j.1462-2920.2004.00617.x

Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins

Scott R. Santos and Howard Ochman*
Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, AZ 85721, USA.

Summary

Molecular characterizations of bacteria often employ ribosomal DNA (rDNA) to establish the identity and relationships among organisms, but the use of rRNA sequences can be problematic as the result of alignment ambiguities caused by indels, the lack of informative characters, and varying functional constraints over the molecule. Although protein-coding regions have been used as an alternative to rRNA, there is neither consensus among the genes examined nor ways to rapidly obtain sequence information for such genes from uncharacterized bacterial species. To standardize the set of protein-coding loci assayed in bacterial genomes, we examined over 100 widely distributed genes to identify sets of universal primers for use in the PCR amplification of protein coding regions that are common to virtually all bacteria. From this set, we developed primer sets that each target of 10 genes spanning an array of genomic locations and functional categories. Although many of the primers contain sequence degeneracies that aid in targeting genes across diverse taxa, most are adequate for direct sequencing of amplification products, thereby eliminating intermediate cloning before sequence determination. We foresee the analysis of these protein-coding regions as being complementary to ribosomal DNA for answering questions pertaining to bacterial identification, classification, phylogenetics and evolution.

Introduction

Communities of bacteria have been recovered and characterized from diverse environments, ranging from cultivated fields and arctic tundra (Torsvik and Ovreas, 2002; Buckley and Schmidt, 2003), to thermal springs and open ocean (Gordon and Giovannoni, 1996; DeLong, 2001; Blank *et al*., 2002), and including those within human and animal hosts (Kroes *et al*., 1999; Frank *et al*., 2003; Moran *et al*., 2003; Rolain *et al*., 2003; Schramm *et al*., 2003). Although bacteria are the most ancient and pervasive organisms on the planet (Whitman *et al*., 1998; Sheridan *et al*., 2003), identifying and distinguishing microbial species, and inferring their relationships, have been historically fraught with difficulties because of a lack of morphologically distinguishing features. Moreover, the initial dependence on biochemical and metabolic properties as key characteristics for the classification of bacteria has limited analysis to the estimated 1% of microbes that can be cultivated (Amann *et al*., 1995; Kaeberlein *et al*., 2002).

Since the 1980s, analysis of ribosomal RNA and DNA sequences has served as the standard to assess microbial diversity in nature and to classify bacterial species (Fox *et al*., 1980; Busse *et al*., 1996). The appeal of these molecules lies in their ubiquitous distribution and relatively slow rates of evolution, which enables comparisons among very divergent bacteria. In addition, the highly conserved domains of ribosomal DNA can serve as templates for designing amplification primers to generate the corresponding regions from all samples. By applying a universal genetic character by which all organisms, even those not amenable to cultivation, can be compared, these procedures have greatly enhanced our view of microbial diversity and relationships (DeLong and Pace, 2001; Smith *et al*., 2001; Hagstrom *et al*., 2002; Hayashi *et al*., 2002; Francino *et al*., 2003). To date, the nearly 75 000 bacterial isolates for which rDNA sequences are available have been classified into almost 40 phyla, over half of which have been recognized solely on the basis of their 16S rDNA sequences (Hugenholtz *et al*., 1998; Dojka *et al*., 2000; DeLong, 2002; Hugenholtz, 2002).

Despite these advances, several authors have noted shortcomings in using 16S rDNA sequences for assessing microbial diversity and for phylogenetic analysis. Quite apart from the fact the 16S rDNA spans a very small portion of the genome, the lack of informative characters and its slow evolutionary rate can complicate both the differentiation of closely related strains of bacteria as well as the resolution of evolutionary trees (e.g. Rogall *et al*., 1990; Bennasar *et al*., 1996). Furthermore, because rDNA does not encode a protein, the occurrence of addi-

Received 26 November, 2003; revised 29 January, 2004; accepted 29 January, 2004. *For correspondence. E-mail hochman@e-mail.arizona.edu; Tel. (+1) 520 626 8355; Fax (+1) 520 621 3709.

tions and deletions (indels) can introduce problems for sequence alignments, and selection on secondary structures can cause sequence convergence and saturation, thereby distorting the actual relationships among organisms (Hillis and Dixon, 1991; Dixon and Hillis, 1993; Kjer, 1995). Finally, there are questions pertaining to the degree to which rDNA sequences are resistant to lateral gene transfer (Yap *et al*., 1999) or whether any single molecule can accurately represent the true organismal phylogeny.

To circumvent many of these problems, we have focused on universally conserved protein-coding gene sequences. Unlike rDNA, such genes are typically present in single copies within genomes, are subject to very low rates of indels and can be readily partitioned into synonymous and non-synonymous sites, which undergo very different rates and patterns of evolution. Such features allow for both the accurate alignment of homologues from divergent organisms as well as the differentiation of very closely related lineages. Conserved protein-coding regions have been used as an alternative to 16S rDNA to address many questions pertaining to the relationships of microbes ranging from the deepest branching bacterial phylogenies (reviewed in Francino *et al*., 2003) to the assessment of genetic variation within individual species (Enright and Spratt, 1999; Feil *et al*., 2003). Unfortunately, there is no consensus among the genes examined across species, with the selection of loci often specific to a particular study. In order to: (i) standardize the set of loci assayed in bacterial genomes at all phylogenetic levels; (ii) complement and extend the utility of 16S rRNA in the identification and classification of bacteria, and (iii) explore differences in the substitution patterns across genes common to different bacterial groups, we have developed sets of conserved primers of limited degeneracy designed for use in recovering protein-coding genes present in virtually all bacteria (Eisen, 1995; 1998). By allowing access to more comprehensive portions of bacterial genomes, these primer sets should be of broad application in molecular, genetic and evolutionary analyses.

Results

A total of 143 genes was identified which fit the criteria of being both single copy and present in more than 95% of completely sequenced bacterial genomes. Based on analysis of the amino acid alignments of these broadly distributed sequences, candidate genes could be categorized into three classes: (i) those containing no regions that were sufficiently well-conserved to design low-degeneracy primers (35 genes); (ii) those with a single conserved region (69 genes), and (iii) those with at least two highly conserved regions separated by at least 100 amino acids (39 genes). Focusing first on this set containing 39 genes, we were able to design and synthesize primers for 10 functionally diverse genes (Table 1). Most of these genes are distributed around the bacterial chromosome; however, from a large group of conserved genes known to be clustered in most bacterial genomes (Hansmann and Martin, 2000; Lathe *et al*., 2000), we utilized three representatives (e.g. *fusA*, *rplB* and *rpoB*). For many of the selected loci, there were numerous potential primer binding sites conserved across bacterial taxa: at the extreme, this led to designing nine forward and seven reverse primers for *gyrB*.

Primer combinations resulting in high levels of background and non-specific amplification among the tested isolates, as well as those yielding low amplification efficiencies, were systematically removed. Ultimately, we identified at least one, but often several, combinations of primers for each of the 10 loci that produced either a single amplicon or one with a minimal background (Table 1) for more than two-thirds of the phylogentically diverse bacterial isolates screened. Primer combinations differ in their utility across taxonomic groups (Table 2); amplifications performed on representatives of the ε-Proteobacteria (*Helicobacter pylori*) and Spirochetes

Table 1. Primer sequences for universally conserved genes in bacteria.

Gene name	Primers[a]
fusA	fusAF: 5′-CATCGGCATCATGgcncayathga-3′
	fusAR: 5′-CAGCATCGGCTGCayncyttrtt-3′
gyrB	gyrBBAUP2: 5′-GCGGAAGCGGccngsnatgta-3′
	gyrBBNDN1: 5′-CCGTCCACGTCGgcrtcngycat-3′
ileS	ileSBCUP1: 5′-GCCCGGCTGGgaywsncaygg-3′
	ileSBKDN1: 5′-TGGAGCCGGAGTCGawccanmmntc-3′
lepA	lepABAUP1: 5′-CATCGCCCACATCgaycayggnaa-3′
	lepABAUP2: 5′-TGCATCATCGCCCACrtngaycaygg-3′
	lepABIDN1: 5′-CATGTGCAGCAGGccnarraancc-3′
leuS	leuSF: 5′-GAGACCGTGCTGGCCaygarsarrt-3′
	leuSBKDN1: 5′-GGGGCAGCcccarwanckyt-3′
pyrG	pyrGBAUP1: 5′-GGCGTGGTGTCCTCCntnggnaargg-3′
	pyrGBDDN2: 5′-GGAAGGGCAGGCACTcnatrtcnccna-3′
recA	recABDUP1: 5′-CCCGAGTCCTCCggnaaracnac-3′
	recABGDN2: 5′-CGTTGCCGCCGgkngtnryytc-3′
	recABHDN1: 5′-GAAGGGTGGGGCCanyttrttytt-3′
recG	recGBHUP2: 5′-GGGCGACGTGGGCdsnggnaarac-3′
	recGBMDN1: 5′-GGGTCCGGGGGatnggngtngc-3′
rplB	rplBBDUP1: 5′-CAAGGTGGAGCGCATCsantaygaycc-3′
	rplBBHDN1: 5′-GCCGCCGCCGwdnggrtgrtc-3′
	rplBR: 5′-CGCCGCCGCCGwrnggrtgrtc-3′
rpoB	rpoBBDUP1: 5′-GGGCACCTTCATCATCaayggndbnga-3′
	rpoBBDUP4: 5′-CATGGGCGACATCccnhwnatnac-3′
	rpoBBJDN2: 5′-CCGATGTTCGGGCcytcnggngtyt-3′
	rpoBBJDN3: 5′-GATGTTCGGGCCCtcnggngtytc-3′

a. Capital letters denote non-degenerate 5′ consensus clamp region, and lower case letters represent the degenerate 3′ core region of primers, as designed by CODEHOP (Rose *et al*., 1998). Abbreviations for degenerate nucleotides follow those of the IUPAC ambiguity codes.

Table 2. Testing amplification with universal primers on some representative bacterial DNAs. Genes which could be PCR amplified from an isolate with the universal primers are designated with a (+) whereas lack of PCR amplification is indicated by a (–).

	Gene									
Test species	*fusA*	*gyrB*	*ileS*	*lepA*	*leuS*	*pyrG*	*recA*	*recG*	*rplB*	*rpoB*
Agrobacterium tumefaciens (α-Proteobacteria)	+	+	+	+	+	+	+	+	+	+
Bacillus subtilis (Firmicutes)	+	+	+	+	–	+	+	+	–	–
Borrelia burgdorferi (Spirochetes)	+	+	–	–	–	–	+	–	–	–
Cellvibrio japonicus (γ-Proteobacteria)	–	+	–	+	–	–	+	–	+	+
Chlorobium tepidum (Chlorobi)	+	+	+	+	+	+	+	+	+	+
Clostridium vincentii (Firmicutes)	–	+	–	+	–	–	+	–	+	+
Escherichia coli (γ-Proteobacteria)	+	+	+	+	+	+	+	+	+	+
Flavobacterium hydatis (Bacteroidetes)	–	+	–	+	–	–	–	–	+	–
Helicobacter pylori (ε-Proteobacteria)	–	+	–	–	–	+	–	+	–	–
Neisseria gonorrhoeae (β-Proteobacteria)	–	+	+	+	+	+	+	+	+	+
Rhizobium leguminosarum (α-Proteobacteria)	+	–	–	+	+	–	+	–	+	+

(*Borrelia burgdorferi*) were particularly unsuccessful compared with other bacterial groups.

Although the diversity in target sequences required that some primers contain up to 512-fold degeneracy, these PCR primers were, by-and-large, adequate for directly sequencing amplification products, thereby eliminating the need for the intermediate cloning of amplicons prior to sequence determination. Sequence data generated directly from PCR amplicons were over 99% identical to sequences obtained from cloned inserts. As anticipated, amino acid sequences (translated from the nucleotide sequences) could be aligned with minimal effort.

Discussion

We have developed sets of universal primers that target protein-coding genes distributed in virtually all bacteria. With these primers, the DNA sequences of the corresponding regions of universally conserved genes can be generated via PCR, and can be used to: (i) recover sequences from small or heterogeneous environmental samples of bacteria; (ii) provide a common set of characters by which one can identify, classify and determine the relationships of bacteria at varied taxonomic levels, and (iii) examine the heterogeneity in rates and patterns of sequence evolution across taxa.

Like the conserved primers available for the amplification of ribosomal DNA, the primer pairs designed to amplify protein-coding regions will generate PCR products from most, but not all, bacterial genomes. For the panel of isolates that we tested, there was about a 60% success rate for obtaining the correct amplification product from each pair of primer (Table 2). However, the application of all possible primer combinations for a given gene greatly reduced the number of amplifications that yielded no, or multiple, PCR reaction products. Those few strains in which no reaction products could be recovered were apparently the result of the low G + C contents of some genomes coupled with the stringent primer annealing conditions that were applied. This problem is exacerbated by the fact that primers designed by CODEHOP possess a G + C rich 5′-consensus clamp that is integral to the amplification strategy (Rose *et al*., 1998). For such cases, we synthesized additional primers directed towards a more restricted taxonomic group (i.e. a particular phylum) based on known features (i.e. G + C content and codon usage biases) of constituent genomes. Although the design of the original primers ensures that the synthesis of additional primers will seldom be required, we adopted this alternative strategy to recover sequences from ε-Proteobacteria based on the available genome sequence of *Helicobacter pylori* and *Campylobacter jejuni* (data not shown).

The approach that we outline – i.e. using multiple conserved protein-coding regions for assessing the diversity and relationships of bacterial species – is similar, in concept, to the genotypic data obtained by multilocus sequence typing (MLST) which has been developed to assess genic diversity within bacterial species (Enright and Spratt, 1999; Spratt, 1999). However, MLST schemes

focus on the allelic variation at 6–8 enzyme loci within a particular species and makes little attempt to examine the same loci in different taxa because the genetic information is used principally to monitor the epidemiology and population structure within individual pathogenic species. The conserved primer sets designed in the present study can serve as a convenient starting point in the development of MLST schemes for a wide range of bacterial species, including those that have not been previously well characterized at the molecular level, and will further allow the comparative analysis of molecular and evolutionary processes across diverse taxa.

Many of the features that make the analysis of ribosomal DNA sequences so attractive for analysing bacterial diversity are also exhibited by conserved protein-coding regions. In addition, these genes are typically single copy within a genome, and their amino acid sequences promote alignments among very distantly related organisms. Finally, having access to additional regions of the genome can assist in screening clone libraries for conserved regions, phylogenetic reconstruction (by providing larger numbers of informative characters), and evolutionary studies (by permitting comparisons of the history and substitution patterns of different genes). Thus, information derived from universally conserved protein coding regions can complement and extend the utility of ribosomal DNA sequences for resolving questions pertaining to the molecular biology, genetics and evolution of bacterial genomes.

Experimental procedures

Choice of genes and primer design

A gene was considered a candidate for primer design if its orthologue was present in single copy in all ($n = 132$ as of August 2002), or nearly all, completely sequenced bacterial genomes. Amino acid sequences were retrieved from the NCBI Clusters of Orthologous Groups (COG) database (http://www.ncbi.nlm.nih.gov/COG/), supplemented with sequences from the Joint Genome Institute genome sequencing projects (http://www.jgi.doe.gov/JGI_microbial/html/index.html). Sequences were aligned in CLUSTAL X (Thompson *et al*., 1997).

To identify regions corresponding to the most highly conserved segments of each candidate gene, amino acid alignments were submitted to the BLOCKS Multiple Alignment Processor (http://blocks.fhcrc.org/blocks/process_blocks.html). Output from BLOCKS was used to design PCR primers with CODEHOP (Rose *et al*., 1998) employing the following parameters: annealing temperature of at least 55°C, equal codon usage, and a sequence degeneracy of ≤512. The remaining parameters in CODEHOP were maintained at their default settings. Primer combinations whose predicted products ranged from 400 to 2000 bp were selected for empirical testing. To improve the success of PCR amplifications, between two and six primers were chosen for synthesis when CODEHOP designed more than a single primer for a particular direction of a given gene. Resulting primers ranged from 20 to 30 nucleotides in length and were synthesized by Sigma-Genosys.

Bacterial isolates

To ascertain primer efficacy and amplification efficiency, a panel of taxonomically diverse bacteria was screened. The panel consisted of *Flavobacterium hydatis* (Bacteroidetes), *Chlorobium tepidum* (Chlorobi), *Bacillus subtilis* and *Clostridium vincentii* (Firmicutes), *Agrobacterium tumefaciens* and *Rhizobium leguminosarum* (α-Proteobacteria), *Neisseria gonorrhoeae* (β-Proteobacteria), *Helicobacter pylori* (ε-Proteobacteria), *Cellvibrio japonicus* and *Escherichia coli* (γ-Proteobacteria) and *Borrelia burgdorferi* (Spirochetes). DNA from *F. hydatis*, *C. vincentii*, *R. leguminosarum* and *C. japonicus* was kindly provided by D. R. Humphry (University of York). The remaining bacterial isolates were purchased from the American Type Culture Collection (ATCC). Cultures were grown overnight in the recommended medium and DNA extracted using the Qiagen DNeasy Tissue Kit.

Amplification, cloning and sequencing methods

Polymerase chain reactions were conducted in 25 μl volumes containing 50 mM KCl, 10 mM Tris-HCl (pH 8.3), 1.5 mM $MgCl_2$, 200 μM of each dNTP, 15 pmol of each primer, 1 U Eppendorf *Taq* polymerase, and 10 ng of template DNA. For genes in which more than a single forward or reverse primer were designed, all possible primer combinations were tested. Reactions incorporated a 'touchdown' PCR procedure (Don *et al*., 1991) using the following conditions: initial denaturation at 94°C for 2 min, 10 cycles of 94°C for 1 min, 60°C for 1 min (−1°C/cycle), and 72°C for 1 min followed by 21 cycles of 94°C for 1 min, 50°C for 1 min, and 72°C for 1 min, with a final extension of 72°C for 5 min.

Given the size distribution of anticipated products (400–2000 bp), PCR amplifications were assessed on 2% agarose gels in 0.5× Tris-borate (TBE). Amplifications that resulted in a single product were purified with the QIAquick PCR purification kit (Qiagen). In cases where reactions generated multiple products, the resulting DNA fragments were gel purified in a 2% 1× Tris-acetate (TAE) agarose gel. Products of the anticipated size were excised from the gel and recovered using the QIAquick gel extraction kit (Qiagen). In situations where extraneous products could not be fractionated from the desired amplicon by gel purification, amplicons were cloned using the TOPO TA Cloning′ kit (Invitrogen) before sequencing. In all other cases, amplicons were sequenced directly without intermediate cloning. Sequence reactions utilized the ABI PRISM Big Dye terminator cycle sequencing ready reaction kit (Perkin Elmer). Cloned products were sequenced from multiple independent inserts to ensure adequate representation. Conceptual translations of nucleotide sequences were identified and verified via BLAST (Altschul *et al*., 1997) searches to the GenBank database.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped

BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev* **59:** 143–169.

Bennasar, A., Rossello Mora, R., Lalucat, J., and Moore, E.R.B. (1996) 16S rRNA gene sequence analysis relative to genomovars of *Pseudomonas stutzeri* and proposal of *Pseudomonas balearica* sp nov. *Int J Syst Bacteriol* **46:** 200–205.

Blank, C.E., Cady, S.L., and Pace, N.R. (2002) Microbial composition of near-boiling silica-depositing thermal springs throughout Yellowstone National Park. *Appl Environ Microbiol* **68:** 5123–5135.

Buckley, D.H., and Schmidt, T.M. (2003) Diversity and dynamics of microbial communities in soils from agroecosystems. *Environ Microbiol* **5:** 441–452.

Busse, H.J., Denner, E.B.M., and Lubitz, W. (1996) Classification and identification of bacteria: Current approaches to an old problem. Overview of methods used in bacterial systematics. *J Biotech* **47:** 3–38.

DeLong, E.F. (2001) Microbial seascapes revisited. *Curr Opin Microbiol* **4:** 290–295.

DeLong, E.F. (2002) Microbial population genomics and ecology. *Curr Opin Microbiol* **5:** 520–524.

DeLong, E.F., and Pace, N.R. (2001) Environmental diversity of bacteria and archaea. *Syst Biol* **50:** 470–478.

Dixon, M.T., and Hillis, D.M. (1993) Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. *Mol Biol Evol* **10:** 256–267.

Dojka, M.A., Harris, J.K., and Pace, N.R. (2000) Expanding the known diversity and environmental distribution of an uncultured phylogenetic division of bacteria. *Appl Environ Microbiol* **66:** 1617–1621.

Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K., and Mattick, J.S. (1991) 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res* **19:** 4008.

Eisen, J.A. (1995) The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *J Mol Evol* **41:** 1105–1123.

Eisen, J.A. (1998) A phylogenomic study of the MutS family of proteins. *Nucleic Acids Res* **26:** 4291–4300.

Enright, M.C., and Spratt, B.G. (1999) Multilocus sequence typing. *Trends Microbiol* **7:** 482–487.

Feil, E.J., Cooper, J.E., Grundmann, H., Robinson, D.A., Enright, M.C., Berendt, T., *et al.* (2003) How clonal is *Staphylococcus aureus*? *J Bacteriol* **85:** 3307–3316.

Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., *et al.* (1980) The phylogeny of prokaryotes. *Science* **209:** 457–463.

Francino, M.P., Santos, S.R., and Ochman, H. (2003) Phylogenetic relationships of bacteria with special reference to endosymbionts and enteric species. In *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community*, 3rd edn. release 3.13, May 12, 2003. Dworkin, M. (ed.-in-Chief). New York, USA: Springer-Verlag. [WWW document] URL http://link.springer-ny.com/link/service/books/10125/

Frank, D.N., Spiegelman, G.B., Davis, W., Wagner, E., Lyons, E., and Pace, N.R. (2003) Culture-independent molecular analysis of microbial constituents of the healthy human outer ear. *J Clin Microbiol* **41:** 295–303.

Gordon, D.A., and Giovannoni, S.J. (1996) Detection of stratified microbial populations related to *Chlorobium* and *Fibrobacter* species in the Atlantic and Pacific oceans. *Appl Environ Microbiol* **62:** 1171–1177.

Hagstrom, A., Pommier, T., Rohwer, F., Simu, K., Stolte, W., Svensson, D., and Zweifel, U.L. (2002) Use of 16S ribosomal DNA for delineation of marine bacterioplankton species. *Appl Environ Microbiol* **68:** 3628–3633.

Hansmann, S., and Martin, W. (2000) Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *Int J Syst Evol Microbiol* **50:** 1655–1663.

Hayashi, H., Sakamoto, M., and Benno, Y. (2002) Phylogenetic analysis of the human gut microbiota using 16S rDNA clone libraries and strictly anaerobic culture-based methods. *Microbiol Immun* **46:** 535–548.

Hillis, D.M., and Dixon, M.T. (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. *Quart Rev Biol* **66:** 411–453.

Hugenholtz, P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3:** REVIEWS0003.

Hugenholtz, P., Goebel, B.M., and Pace, N.R. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* **180:** 4765–4774.

Kaeberlein, T., Lewis, K., and Epstein, S.S. (2002) Isolating 'uncultivable' microorganisms in pure culture in a simulated natural environment. *Science* **296:** 1127–1129.

Kjer, K.M. (1995) Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. *Mol Phylogenet Evol* **4:** 314–330.

Kroes, I., Lepp, P.W., and Relman, D.A. (1999) Bacterial diversity within the human subgingival crevice. *Proc Natl Acad Sci USA* **96:** 14547–14552.

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

Moran, N.A., Dale, C., Dunbar, H., Smith, W.A., and Ochman, H. (2003) Intracellular symbionts of sharpshooters (Insecta: Hemiptera: Cicadellinae) form a distinct clade with a small genome. *Environ Microbiol* **5:** 116–126.

Rogall, T., Wolters, J., Flohr, T., and Bottger, E.C. (1990) Towards a phylogeny and definition of species at the molecular level within the genus *Mycobacterium*. *Int J Syst Bacteriol* **40:** 323–330.

Rolain, J.M., Franc, M., Davoust, B., and Raoult, D. (2003) Molecular detection of *Bartonella quintana*, *B. koehlerae*, *B. henselae*, *B. clarridgeiae*, *Rickettsia felis*, and *Wolbachia pipientis* in cat fleas, France. *Emerg Infect Dis* **9:** 338–342.

Rose, T.M., Schultz, E.R., Henikoff, J.G., Pietrokovski, S., McCallum, C.M., and Henikoff, S. (1998) Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res* **26:** 1628–1635.

Schramm, A., Davidson, S.K., Dodsworth, J.A., Drake, H.L., Stahl, D.A., and Dubilier, N. (2003) *Acidovorax*-like symbionts in the nephridia of earthworms. *Environ Microbiol* **5:** 804–809.

Sheridan, P.P., Freeman, K.H., and Brenchley, J.E. (2003) Estimated minimal divergence times of the major bacterial and archaeal phyla. *Geomicrobiol J* **20:** 1–14.

Smith, Z., McCaig, A.E., Stephen, J.R., Embley, T.M., and Prosser, J.I. (2001) Species diversity of uncultured and cultured populations of soil and marine ammonia oxidizing bacteria. *Microbial Ecol* **42:** 228–237.

Spratt, B.G. (1999) Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet. *Curr Opin Microbiol* **2:** 312–316.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997) The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25:** 4876–4882.

Torsvik, V., and Ovreas, L. (2002) Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol* **5:** 240–245.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* **95:** 6578–6583.

Yap, W.H., Zhang, Z., and Wang, Y. (1999) Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J Bacteriol* **181:** 5201–5209.