JOURNAL OF CLINICAL MICROBIOLOGY, Sept. 2007, p. 2761–2764 Vol. 45, No. 9
0095-1137/07/$08.00+0 doi:10.1128/JCM.01228-07

MINIREVIEW

16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls[▽]

J. Michael Janda* and Sharon L. Abbott

Microbial Diseases Laboratory, Division of Communicable Disease Control, California Department of Public Health, Richmond, California 94804

The use of 16S rRNA gene sequences to study bacterial phylogeny and taxonomy has been by far the most common housekeeping genetic marker used for a number of reasons. These reasons include (i) its presence in almost all bacteria, often existing as a multigene family, or operons; (ii) the function of the 16S rRNA gene over time has not changed, suggesting that random sequence changes are a more accurate measure of time (evolution); and (iii) the 16S rRNA gene (1,500 bp) is large enough for informatics purposes (12). In 1980 in the *Approved Lists*, 1,791 valid names were recognized at the rank of species. Today, this number has ballooned to 8,168 species, a 456% increase (http://www.bacterio.cict.fr/number.html#total). The explosion in the number of recognized taxa is directly attributable to the ease in performance of 16S rRNA gene sequencing studies as opposed to the more cumbersome manipulations involving DNA-DNA hybridization investigations. DNA-DNA hybridization is unequivocally the "gold standard" for proposed new species and for the definitive assignment of a strain with ambiguous properties to the correct taxonomic unit. Based upon DNA-DNA reassociation kinetics, the genetic definition of a species is quantifiable, i.e., (i) ca. ≥70% DNA-DNA relatedness and (ii) 5°C or less ΔT_m for the stability of heteroduplex molecules. DNA hybridization assays are not without their shortcomings, however, being time-consuming, labor-intensive, and expensive to perform. Today, fewer and fewer laboratories worldwide perform such assays, and many studies describing new species are solely based upon small subunit (SSU) sequences or other polyphasic data.

In the early 1990s the availability DNA sequencers in terms of cost, methodologies, and technology improved dramatically, such that many centers can now afford such instrumentation. In 1994, Stackebrandt and Goebel (15) summarized the emergence of SSU sequence technology and its potential usefulness in the definition of a species. Although it has been demonstrated that 16S rRNA gene sequence data on an individual strain with a nearest neighbor exhibiting a similarity score of <97% represents a new species, the meaning of similarity scores of >97% is not as clear (13). This latter value can represent a new species or, alternatively, indicate clustering within a previously defined taxon. DNA-DNA hybridization studies have traditionally been required to provide definitive answers for such questions. Whereas 16S rRNA gene sequence data can be used for a multiplicity of purposes, unlike DNA hybridization (>70% reassociation) there are no defined "threshold values" (e.g., 98.5% similarity) above which there is universal agreement of what constitutes definitive and conclusive identification to the rank of species.

BACTERIAL IDENTIFICATION USING 16S RRNA SEQUENCING

Unidentified bacteria or isolates with ambiguous profiles. One of the most attractive potential uses of 16S rRNA gene sequence informatics is to provide genus and species identification for isolates that do not fit any recognized biochemical profiles, for strains generating only a "low likelihood" or "acceptable" identification according to commercial systems, or for taxa that are rarely associated with human infectious diseases. The cumulative results from a limited number of studies to date suggest that 16S rRNA gene sequencing provides genus identification in most cases (>90%) but less so with regard to species (65 to 83%), with from 1 to 14% of the isolates remaining unidentified after testing (5, 11, 17). Difficulties encountered in obtaining a genus and species identification include the recognition of novel taxa, too few sequences deposited in nucleotide databases, species sharing similar and/or identical 16S rRNA sequences, or nomenclature problems arising from multiple genomovars assigned to single species or complexes.

Routine isolates. Surveys have looked at the feasibility of identifying routine clinical isolates or specific groups of medically important bacteria using SSU gene sequence data. In each of these studies, SSU sequence data has been compared to identification results obtained either in conventional or commercial test formats (Table 1). A couple of general observations can be made from these investigations, namely, (i) a higher percentage of species identifications were obtained using SSU sequence results than with either conventional or commercial methods and (ii) most studies, with the exception of one study by Fontana et al. (6), have found that 16S yielded species identification rates of 62 to 91%. In the study by Fontana et al. (6) the closest match in the MicroSeq 500 database was considered the identification no matter what the distance

* Corresponding author. Mailing address: Microbial Diseases Laboratory, 850 Marina Bay Parkway, Rm. E164, Richmond, CA 94804. Phone: (510) 412-3700. Fax: (510) 412-3722. E-mail: JohnMichael.Janda@cdph.ca.gov.

[▽] Published ahead of print on 11 July 2007.

TABLE 1. 16S species identification for routine isolates

No. of strains	Group studied[a]	16S			Commercial system(s)	Species identification (%)[c]			Reference
		Size(s) (bp)	Database[b]	Criteria (%)[c]		Conv	Comm	16S	
72	GNB	1,189, 527, 418	MicroSeq	CM	Conv, MIDI, Biolog	90	67.7–84.6	89.2	16
328	Mycobacteria	500	MicroSeq	≥99	Conv.	42		62.5	8
83	GNB, GPB	527	MicroSeq	CM	Vitek 2, Phoenix		77.1	100	6
231	*Bacteroides*	899, 711	GenBank	≥99	Conv	74.5		83.1	14
47	CNS	1,500	GenBank	>97	API StaphID, Phoenix		63.8–85.1	87.2	9
20	GPA	1,500	GenBank MicroSeq	≥98	Vitek ANA, RapID ANA II, API 20A		20–45	65	10
107	GNNFB	796	GenBank, EMBL, DDBJ	≥99	API 20NE, Vitek 2		53.2–54.2	91.6	2

[a] CNS, coagulase-negative staphylococci; GNB, gram-negative bacteria; GNNFB, gram-negative nonfermentative bacteria; GPA, gram-positive anaerobes; GPB, gram-positive bacteria.

[b] DDBJ, DNA Data Bank Japan; EMBL, European Molecular Biology Laboratory.

[c] CM, closest match; Comm, commercial system; Conv, conventional phenotypic tests.

score was. For bacteria that are difficult to grow or identify the identification rates were lower with 16S rRNA sequencing (62 to 83%) than the values traditionally acceptable in the clinical laboratory (i.e., ≥90%) (12). Problems again revolved around complete and accurate databases and groups that are not easily distinguishable by 16S rRNA gene sequencing (2, 8).

ISSUES

It is clear from the information listed in Table 1 that 16S rRNA gene sequence information has an expanding role in the identification of bacteria in clinical or public health settings. However, the data also clearly show that it is not foolproof and applicable in each and every situation.

Bacterial nomenclature in relation to 16S rRNA gene sequencing. There were more than 1,700 species on the 1980 *Approved Lists*, but this list does not imply that all of these taxa are valid. Many names included predate modern DNA-DNA hybridization studies and most certainly phylogenetic investigations. Thus, the type strains for many species may not accurately reflect the entire genomic composition of the nomenspecies, and such situations have a direct bearing on SSU studies with reference to microbial identification. Some bacterial species exist as "phenospecies" or "complexes," that is, more than one genomovar (DNA group) exists within that species and cannot be separated phenotypically. Examples of these kinds of situations include *Enterobacter cloacae* (at least 7 genomovars originally), *Pseudomonas stutzeri* (18 genomovars originally), and the genus *Acinetobacter* (22 genomovars originally).

Resolution of 16S rRNA gene sequencing. Although 16S rRNA gene sequencing is highly useful in regards to bacterial classification, it has low phylogenetic power at the species level and poor discriminatory power for some genera (2, 11), and DNA relatedness studies are necessary to provide absolute resolution to these taxonomic problems. The genus *Bacillus* is a good example of this. The type strains of *B. globisporus* and *B. psychrophilus* share >99.5% sequence similarity with regard to their 16S rRNA genes, and yet at the DNA level exhibit only 23 to 50% relatedness in reciprocal hybridization reactions (7). In our laboratory we have found that the type strains of *Edwardsiella* species exhibit 99.35 to 99.81% similarity to each other, and yet these three species are clearly distinguishable biochemically and by DNA homology (28 to 50% relatedness). Such examples indicate that SSU sequence similarity even to a very high level does not in each case imply identity or accuracy in microbial identifications. Many investigators have found resolution problems at the genus and/or species level with 16S rRNA gene sequencing data (Table 2). These groups include (not exclusively), the family *Enterobacteriaceae* (in particular, *Enterobacter* and *Pantoea*), rapid-growing mycobacteria, the *Acinetobacter baumannii-A. calcoaceticus* complex, *Achromobacter*, *Stenotrophomonas*, and *Actinomyces*. Some of these problems are related to bacterial nomenclature and taxonomy while others are related to different issues cited below.

A further problem regarding the resolution of 16S rRNA gene sequencing concerns sequence identity or very high similarity scores. Reports have documented 16S rRNA gene sequence similarities or identity for the *Streptococcus mitis* group and other nonfermenters (Table 2). In such instances 16S rRNA gene sequence data cannot provide a definitive answer since it cannot distinguish between recently diverged species (13, 16). In other instances, the difference between the closest and next closest match to the unknown strain is <0.5% divergence (>99.5% similarity). In these circumstances, such small differences cannot justify choosing the closest match as a definitive identification, although in some studies this is exactly what was done (6).

TABLE 2. Selected examples of bacterial genera and species with identification problems using 16S rRNA gene sequencing

Genus	Species
Aeromonas	*A. veronii*
Bacillus	*B. anthracis*, *B. cereus*, *B. globisporus*, *B. psychrophilus*
Bordetella	*B. bronchiseptica*, *B. parapertussis*, *B. pertussis*
Burkholderia	*B. cocovenenans*, *B. gladioli*, *B. pseudomallei*, *B. thailandensis*
Campylobacter	Non-*jejuni-coli* group
Edwardsiella	*E. tarda*, *E. hoshinae*, *E. ictaluri*
Enterobacter	*E. cloacae*
Neisseria	*N. cinerea*, *N. meningitidis*
Pseudomonas	*P. fluorescens*, *P. jessenii*
Streptococcus	*S. mitis*, *S. oralis*, *S. pneumoniae*

TABLE 3. Recommended guidelines for use of 16S rRNA gene sequencing for microbial identification

Category	Guidelines
Strain to be sequenced	Phenetic profile of strain is not known by general grouping to present difficulties for identification by 16S rRNA gene analysis (Table 2) For strains such as those in Table 2 requiring molecular identification, another housekeeping gene is required (e.g., *rpoB*)
16S rRNA gene sequencing	Minimum: 500 to 525 bp sequenced; ideal: 1,300 to 1,500 bp sequenced <1% position ambiguities
Criteria for species identification	Minimum: >99% sequence similarity; ideal: >99.5% sequence similarity Sequence match is to type strain or reference strain of species that has undergone DNA-relatedness studies For matches with distance scores <0.5% to the next closest species, other properties, including phenotype, should be considered in final species identification

Public and private nucleotide databases. The usefulness of 16S rRNA gene sequencing as a tool in microbial identification is dependent upon two key elements, deposition of complete unambiguous nucleotide sequences into public or private databases and applying the correct "label" to each sequence. Years ago the overall quality of nucleotide sequences deposited in public databases was questionable, since many depositions were of poor quality (9, 13). Much of this misinformation that was originally present in such databases was thought to have been corrected; however, a recent multicenter study from the United Kingdom (1) conservatively estimates that at least 5% of the 1,399 sequences searched had substantial errors associated with them ranging from chimeras (64%) to sequencing errors or anomalies (35%). A 1995 study by Clayton et al. (4) also revealed that at least 26% of 16S rRNA gene sequence pairs (two sequences deposited for the same species) in GenBank had >1% random sequencing errors and, of these, almost half had >2% random sequencing errors.

Species identification definition using 16S rRNA gene sequence data. Unfortunately, no universal definition for species identification via 16S rRNA gene sequencing exists, and authors vary widely in their use of acceptable criteria for establishing a "species" match (Table 1). In none of these studies does the definition of a species "match" ever exceed 99% similarity (<1% divergence). Based on the data listed above, even this threshold value may not be sufficient in all instances to guarantee an accurate identification. In the case of *Aeromonas veronii* the genome can contain up to six copies of the 16S rRNA gene that differ by up to 1.5% among themselves. This implies intragenomic heterogeneity of the 16S rRNA gene among aeromonads and would preclude the use of this technology alone for species identification. The collective data described above strongly suggest that any microbial identifications using 16S rRNA distance scores of >1% are unsatisfactory for a diagnostic or public health reference laboratory.

Miscellaneous issues. A number of other issues related to SSU gene sequencing merit brief mention. These include the number of position ambiguities, sequence gaps, and use of gap and/or nongapped programs with regard to sequence evaluation and analysis. Other concerns involve isolate purity, DNA extraction methods, and possible chimeric molecule formation (9, 16, 17). All of these problems to some extent affect final identifications.

POSSIBLE SOLUTIONS

The use of 16S rRNA gene sequencing in the clinical laboratory is becoming commonplace for identifying biochemically unidentified bacteria or for providing reference identifications for unusual strains. Although some researchers would never question using a molecular identification over a conventional one, 16S rRNA gene sequencing is not infallible, and examples of such misidentifications have been published (3). Although it is clear that SSU sequencing plays an important role in the identification of unknown isolates or those with ambiguous biochemical profiles, it is less clear what that role is in other situations. An intriguing question concerns how accurate is our routine identification of very common species using conventional methodologies or commercial systems. Although it is generally regarded that these identifications are highly accurate, we now have a more convenient and precise mechanism for checking these identifications on a molecular basis. Such studies need to be performed and published.

The use of 16S rRNA gene sequencing for definitive microbial identifications and for publication requires a harmonious set of guidelines for interpretation of sequence data that needs to be implemented so that results from one study can be accurately compared to another. In 2000, Drancourt et al. (5) made several recommendations concerning proposed criteria for 16S rRNA gene sequencing as a reference method for bacterial identification. We support Drancourt's guidelines for including full 16S rRNA gene sequences whenever possible, and in particular, for groups such as *Campylobacter* species that absolutely require it for accurate species identifications. Table 3 expands on these recommendations for use in the diagnostic setting. It is clear that the appropriate use of such technology requires the adoption of standards similar to those previously defined for DNA-DNA hybridization. Because the adaptation of 16S rRNA gene sequencing as a tool in species identification is still a relatively new phenomenon in most clinical laboratories, such standards will most likely continue to evolve over time. Furthermore, use of microarray-based technologies with 16S or other housekeeping gene targets in the future may provide a much more sensitive and definitive platform for molecular species identification in the future.

REFERENCES

1. **Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman.** 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl. Environ. Microbiol. **71:**7724–7736.
2. **Bosshard, P. P., R. Zbinden, S. Abels, B. Böddinghaus, M. Altwegg, and E. C. Böttger.** 2006. 16S rRNA gene sequencing versus the API 20 NE system and the Vitek 2 ID-GNB card for identification of nonfermenting gram-negative bacteria in the clinical laboratory. J. Clin. Microbiol. **44:**1359–1366.
3. **Boudewijns, M., J. M. Bakkers, P. D. J. Sturm, and W. J. G. Melchers.** 2006. 16S rRNA gene sequencing and the routine clinical microbiology laboratory: a perfect marriage? J. Clin. Microbiol. **44:**3469–3470.
4. **Clayton, R. A., G. Sutton, P. S. Hinkle, Jr., C. Bult, and C. Fields.** 1995. Intraspecific variation in small-subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa. Int. J. Syst. Bacteriol. **45:**595–599.
5. **Drancourt, M., C. Bollet, A. Carlioz, R. Martelin, J.-P. Gayral, and D. Raoult.** 2000. 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. J. Clin. Microbiol. **38:**3623–3630.
6. **Fontana, C., M. Favaro, M. Pelliccioni, E. S. Pistoia, and C. Favalli.** 2005. Use of the MicroSeq 16S rRNA gene-based sequencing for identification of bacterial isolates that commercial automated systems failed to identify correctly. J. Clin. Microbiol. **43:**615–619.
7. **Fox, G. E., J. D. Wisotzkey, and P. Jurtshuk, Jr.** 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int. J. Syst. Bacteriol. **42:**166–170.
8. **Hall, L., K. A. Doerr, S. L. Wohlfiel, and G. D. Roberts.** 2003. Evaluation of the MicroSeq system for identification of mycobacteria by 16S ribosomal DNA sequencing and its integration into a routine clinical mycobacteriology laboratory. J. Clin. Microbiol. **41:**1447–1453.
9. **Heikens, E., A. Fleer, A. Paauw, A. Florijn, and A. C. Fluitt.** 2005. Comparison of genotypic and phenotypic methods for species-level identification of clinical isolates of coagulase-negative staphylococci. J. Clin. Microbiol. **43:** 2286–2290.
10. **Lau, S. K. P., K. H. L. Ng, P. C. Y. Woo, K.-T. Yip, A. M. Y. Fung, G. K. S. Woo, K.-M. Chan, T. I. Que, and K.-Y. Yuen.** 2006. Usefulness of the MicroSeq 500 16S rDNA bacterial identification of anaerobic gram-positive bacilli isolated from blood cultures. J. Clin. Pathol. **59:**219–222.
11. **Mignard, S., and J. P. Flandrois.** 2006. 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. J. Microbiol. Methods **67:** 574–581.
12. **Patel, J. B.** 2001. 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. Mol. Diagn. **6:**313–321.
13. **Petti, C. A.** 2007. Detection and identification of microorganisms by gene amplification and sequencing. Clin. Infect. Dis. **44:**1108–1114.
14. **Song, Y., C. Liu, M. Bolaňos, J. Lee, M. McTeague, and S. M. Finegold.** 2005. Evaluation of 16S rRNA sequencing and reevaluation of a short biochemical scheme for identification of clinically significant *Bacteroides* species. J. Clin. Microbiol. **43:**1531–1537.
15. **Stackebrandt, E., and B. M. Goebel.** 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.
16. **Tang, Y.-W., N. M. Ellis, M. K. Hopkins, D. H. Smith, D. E. Dodge, and D. H. Persing.** 1998. Comparison of phenotypic and genotypic techniques for identification of unusual aerobic pathogenic gram-negative bacilli. J. Clin. Microbiol. **36:**3674–3679.
17. **Woo, P. C. Y., K. H. I. Ng, S. K. P. Lau, K.-T. Yip, A. M. Y. Fung, K.-W. Leung, D. M. W. Tam, T.-L. Que, and K.-Y. Yuen.** 2003. Usefulness of the MicroSeq 500 16S ribosomal DNA-based identification system for identification of clinically significant bacterial isolates with ambiguous biochemical profiles. J. Clin. Microbiol. **41:**1996–2001.