Application of SmartGene IDNS Software to Partial 16S rRNA Gene Sequences for a Diverse Group of Bacteria in a Clinical Laboratory[▽]

Keith E. Simmon,[1] Ann C. Croft,[1] and Cathy A. Petti[1,2]*

Associated Regional and University Pathologists Laboratories, Salt Lake City, Utah,[1] and Departments of Pathology and Medicine, University of Utah School of Medicine, Salt Lake City, Utah[2]

Laboratories often receive clinical isolates for bacterial identification that have ambiguous biochemical profiles by conventional testing. With the emergence of 16S rRNA gene sequencing as an identification tool, we evaluated the usefulness of SmartGene IDNS, a 16S rRNA sequence database and software program for microbial identification. Identification by conventional methods of a diverse group of bacterial clinical isolates was compared with gene sequences interrogated by the SmartGene and MicroSeq databases. Of 300 isolates, SmartGene identified 295 (98%) to the genus level and 262 (87%) to the species level, with 5 (2%) being inconclusive. MicroSeq identified 271 (90%) to the genus level and 223 (74%) to the species level, with 29 (10%) being inconclusive. SmartGene and MicroSeq agreed on the genus for 233 (78%) isolates and the species for 212 (71%) isolates. Conventional methods identified 291 (97%) isolates to the genus level and 208 (69%) to the species level, with 9 (3%) being inconclusive. SmartGene, MicroSeq, and conventional identifications agreed for 193 (64%) of the results. Twenty-seven microorganisms were not represented in MicroSeq, compared to only 2 not represented in SmartGene. Overall, SmartGene IDNS provides comprehensive and accurate identification of a diverse group of bacteria and has the added benefit of being a user-friendly program that can be modified to meet the unique needs of clinical laboratories.

The identification of microorganisms historically has relied on phenotypic methods that are often time-consuming, potentially inaccurate because of the inherent mutability of biochemical characteristics, and subject to interpretive bias (5, 24). Because of the growing microbial diversity with emergence of common pathogens having rare or unique phenotypic characteristics and new pathogenic microorganisms with poorly defined phenotypes, conventional methods often cannot fully characterize bacterial isolates, and laboratories are now relying on partial 16S rRNA gene sequencing for bacterial identification (5–7, 9, 14–16, 21, 22). Previous studies have illustrated improved accuracy with 16S rRNA gene sequencing using the MicroSeq (11, 22, 25, 27), GenBank (5–7, 14, 15) or Ribosomal Differentiation of Microorganisms (RIDOM) (2) databases, but each of these databases has limitations, and with the exception of MicroSeq, they are not specifically designed for routine use by clinical laboratories. MicroSeq (Applied Biosystems, Foster City, Calif.) is an established, commercially available reference sequence library based on type strains of microorganisms with a sequence repository housing approximately 2,000 16S rRNA sequences. GenBank is a large public database of nucleotide sequences with over 200,000 named 16S rRNA gene sequences maintained by the National Center for Biotechnology Information (3) but may lack proper oversight, resulting in reference sequences of poor quality or with anomalies (1, 9). SmartGene IDNS (SmartGene, Inc., Raleigh, NC) is advertised as a new, user-friendly database and software

program that currently contains 112,000 rRNA gene sequences and is quality controlled because it is based on the most current deposited GenBank reference sequences which have been screened for quality (e.g., numerous ambiguous bases). SmartGene also offers the creation of an internal reference database, the ability to add alternative gene target databases (e.g., *rpoB*), the capacity to store and then compare previous clinical sequences, and the ability to flag poor sequences.

While public databases have the added advantages of open access, we realized the importance of using databases that can be easily modified to meet the needs of a clinical laboratory and, more importantly, have a greater level of quality assurance. Our aim was to evaluate the use of SmartGene as a tool to identify partial 16S rRNA gene sequences from clinical isolates of a diverse group of bacteria with comparisons to conventional methods and to gene sequences interrogated by the MicroSeq database.

MATERIALS AND METHODS

From February to May 2005, 302 clinical isolates of a diverse group of bacteria that were referred to ARUP Laboratories for identification were evaluated prospectively (Table 1).

Conventional methods. Phenotypic identification comprised a combination of manual biochemical testing (19) and the use of API systems (bio-Merieux, Marcy I'Etoile, France) and/or Microlog software v4.2 (gram-negative database v6.01; gram-positive database v6.11) (Biolog, Hayward, Calif.), which were performed by standard laboratory protocols.

Sequencing of 16S rRNA. Bacterial DNA was extracted from suspensions of colonies in molecular-grade water using QIAmp DNA mini kit (QIAGEN, Hilden, Germany) according to the manufacturer's instructions. PCRs were performed in 20-μl volume containing 1× *Taq* buffer; 0.25 U of TaKaRa *Taq*; 3.0 mM MgCl$_2$ (Takara Bio, Inc., Shiga, Japan); 200 μM each dATP, dGTP, and dCTP; 600 μM dUTP (Roche Diagnostics Corporation, Alameda, Calif.); 0.2 μM each primer; and 2 μl of template. Before template was added, master mix

* Corresponding author. Mailing address: ARUP Laboratories, 500 Chipeta Way, Salt Lake City, UT 84108. Phone: (801) 583-2787. Fax: (801) 584-5207. E-mail: cathy.petti@aruplab.com.

TABLE 1. Genera represented based on SmartGene identifications

Genus (n)	No. of isolates	Method of conventional identification[a]
Gram-negative bacilli		
Nonfermenters (70)		
Achromobacter	8	Biolog, CM
Acinetobacter	11	Biolog, CM
Brevundimonas	1	Biolog, CM
Burkholderia	5	Biolog, CM
Chryseobacterium	1	Biolog, CM
Chryseomonas	2	Biolog, CM
Comamonas	3	Biolog, CM
Moraxella	8	Biolog, CM
Ochrobactrum	1	Biolog, CM
Oligella	1	Biolog, CM
Pseudomonas	17	Biolog, CM
Ralstonia	3	Biolog, CM
Sphingomonas	1	Biolog, CM
Stenotrophomonas	8	Biolog, CM
Enterobacteriaceae (24)		
Citrobacter	1	Biolog, CM
Enterobacter	5	Biolog, CM
Escherichia	14	Biolog, CM
Klebsiella	2	Biolog, CM
Shigella	1	Biolog, CM
Yersinia	1	Biolog, CM
Other gram-negative bacilli (48)		
Aeromonas	3	Biolog, CM
Bordetella	2	Biolog, CM
Campylobacter	1	CM
Capnocytophaga	3	CM
Eikenella	1	Biolog, CM
Haemophilus	10	Biolog, CM
Kingella	1	Biolog, CM
Legionella	1	DFA
Methylobacterium	1	Biolog, CM
Neisseria elongata	13	Biolog, CM
Pasteurella	5	Biolog, CM
Roseomonas	4	CM
Inconclusive	3	
Streptococcus sp. and *Streptococcus*-like organisms (55)		
Abiotrophia	2	Biolog, CM
Aerococcus	4	Biolog, CM
Enterococcus	8	Biolog, CM
Gemella	4	Biolog, CM
Streptococcus	36	Biolog, CM
Weissella	1	Biolog, CM
Staphylococcus sp. and other catalase-positive cocci (11)		
Staphylococcus	9	CM
Micrococcus	2	CM
Gram-positive bacilli (92)		
Actinobaculum	1	Biolog, CM
Actinomyces	5	CM
Arcanobacterium	2	Biolog, CM
Arthrobacter	2	Biolog, CM
Bacillus	9	CM
Brachybacterium	1	Biolog, CM
Brevibacterium	2	Biolog, CM
Cellulomonas	1	Biolog, CM
Cellulosimicrobium	1	Biolog, CM
Corynebacterium	31	Biolog, CM
Lactobacillus	6	CM
Listeria	2	CM
Microbacterium	4	Biolog, CM
Nocardia	1	CM
Paenibacillus	5	CM
Propionibacterium	4	CM
Rothia	8	Biolog, CM
Streptomyces	5	CM
Inconclusive	2	

[a] CM denotes conventional biochemical testing and, when indicated, API methods.

was filtered through a 100-kDa filter (Millipore, Billerica, Mass.) to remove contaminating DNA. The primers used for amplification were 5F (5′-TTGGA GAGTTTGATCCTGGCTC-3′) and 1194R (5′-ACGTCATCCCCACCTTCCT C-3′), which bind to conserved regions near bp 5 and 1194 of the 16S rRNA gene. The PCR mixtures were amplified by initial hold at 94°C for 5 min and then 30 cycles of denaturing at 94°C for 30s, annealing for 30s at 60°C, and extension at 72°C for 2 min. The reaction ended with a final extension at 72°C for 2 min and a hold at 4°C. Positive PCRs and amplicon size were confirmed by gel electrophoresis. PCR products were purified using ExoSAP-IT reagent (USB Corporation, Cleveland, Ohio) per the manufacturer's instructions. PCR products were bidirectionally sequenced with the original amplification primer, 5F, and a reverse primer, 810R (5′GGCGTGGACTTCCAGGGTATCT-3′). Sequencing reactions were performed with Big Dye terminator reagents on an ABI prism 310 or 3730xl (Applied Biosystems, Foster City, Calf.) by the standard automated sequencer protocol.

Sequence analysis. The parameters for SeqScape (Applied Biosystems) analysis was set as follows: KB.bcp (basecaller), KB_3730_POP7_BDTv1.mob (primer/dye set), with 5 N's in 15 bases (ending base). Mixed-base calls were made if the secondary peak height was 75% of the primary peak. The Phred quality values were <30, and reference trimming was used. A reference *Shigella* sp. (GenBank accession no. NC_004741) trimmed after the sequence GTGCC AGCAGCCGCGGTA was used as a reference strain. ABI files were imported into SeqScape software, and consensus sequences were generated. Consensus sequences were compared to related sequences in the SmartGene (version 3.2.3r8) and MicroSeq (version 1.4.3) databases, where all consensus sequences were trimmed to correspond to the primers in the MicroSeq 500-bp kit. For each clinical isolate, the top 20 matches were evaluated. Only consensus sequences with a minimum length of 300 bp were analyzed. Genus- and species-level identifications were assigned using the following criteria: ≥99% identity to a reference entry identified a microorganism to the species level, and 97.0 to 98.9% identity identified a microorganism to the genus level, whereas microorganisms with <97% identity to any sequence were considered unable to be identified definitively. When isolates matched 100% to a reference sequence, matches of ≤99.9% were ignored. Multiple species were assigned to isolates when the top matches were between 99.0 and 99.9%. Variants in 16S rRNA gene copies were identified during manual editing of electropherograms and were defined as abrupt changes in sequence quality. Truncated sequences caused by copy variants in an isolate were repeated by reextracting and resequencing to rule out amplification-induced mutation.

Phylogenetic analysis. Alignments and phylogenetic trees were constructed by the neighbor-joining method using Kimura's two-parameter distance correction model and 1,000 bootstrap replications in the MEGA version 3.1 software package (18). Reference sequences from the MicroSeq and SmartGene databases were used for comparison.

Comparative analysis. The results of partial gene sequencing and phenotypic tests were compared. For isolates with discordant phenotypic and genotypic results, we applied the following criteria. (i) If an isolate shared <97% identity with the phenotypic identification of the genus, phenotypic identification was considered incorrect. (ii) If an isolate had 97 to 98.9% identity with the phenotypic identification of the genus, both phenotypic and genotypic identifications to the genus level were considered correct to the genus level. (iii) If an isolate had ≥99% identity to a species that differed from the phenotypic species identification, the genotypic species identification was considered correct. (iv) If an isolate had ≥99% identity for multiple species that included the phenotypic species identification, the phenotypic species identification was considered correct with the genotypic identification unresolved at the multiple-species level. All microorganisms flagged as having indistinguishable 16S rRNA gene sequences underwent further biochemical testing for identification by standard laboratory methods (see Table 6) (19).

RESULTS

Of 302 clinical isolates collected prospectively, 300 isolates with consensus sequences greater than 300 bp were included in the study. The mean and median fragment sizes of the generated consensus sequences were 489 and 498 bp, respectively. A total of 158 gram-positive and 142 gram-negative organisms were identified (Table 1).

Identification by conventional methods. For 300 isolates, conventional methods identified 291 (97%) to the genus level

TABLE 2. Level of identification by each method according to microorganism group

Organism group (*n*)	No. of isolates identified by method								
	Conventional			SmartGene			MicroSeq		
	Genus	Species	Inconclusive	Genus	Species	Inconclusive	Genus	Species	Inconclusive
Nonfermenters, gram-negative bacilli (70)	69	46	1	70	59	0	65	47	5
Enterobacteriaceae (24)	23	23	1	24	22	0	24	24	0
Other gram-negative bacilli (48)	46	36	2	45	43	3	41	37	7
Enterococcus (8)	8	8	0	8	7	0	8	7	0
Streptococcus sp. and *Streptococcus*-like organisms (47)	47	38	0	47	47	0	44	40	3
Staphylococcus sp. and other catalase-positive cocci (11)	11	7	0	11	11	0	11	11	0
Gram-positive bacilli (92)	87	50	5	90	73	2	78	57	14
Total (300)	291	208	9	295	262	5	271	223	29

and 208 (69%) to the species level, with 9 (3%) being inconclusive (Table 2). Of those identified to the species level, 36 (17%) were assigned to multiple species or to a species group (e.g., *Streptococcus mitis* group).

Identification by sequencing using SmartGene and MicroSeq databases. Of 300 isolates, SmartGene identified 295 (98%) to the genus level and 262 (87%) to the species level, with 5 (2%) being inconclusive (Table 2). Of those identified to the species level, 81 (31%) were assigned to multiple species. MicroSeq identified 271 (90%) to the genus level and 223 (74%) to the species level, with 29 (10%) being inconclusive. Of those identified to the species level, 37 (17%) were assigned to multiple species.

Comparison of SmartGene and MicroSeq databases. Of 300 isolates, SmartGene and MicroSeq agreed on the genus for 233 (78%) and on the species for 212 (71%), with 4 (1%) inconclusive. For 63 isolates with discrepant identifications, SmartGene identified 11 isolates to the genus level that remained unidentified using MicroSeq. Conversely, MicroSeq identified two isolates to the genus level and species that SmartGene identified to the genus level or could not identify. Forty discrepant results occurred from species classification by SmartGene, with MicroSeq assigning 26 to the genus only and 14 as inconclusive.

For the remaining 10 discrepant isolates, SmartGene and MicroSeq disagreed with species identification. All 10 isolates were identified as the same genus using the SmartGene and MicroSeq databases. Three isolates were identified as *Achromobacter xylosoxidans* or *Alcaligenes faecalis* by SmartGene and as *A. piechaudii* by MicroSeq. Both SmartGene (GenBank AB010841) and MicroSeq reference sequences for *Achromobacter piechaudii* were derived from ATCC 43552; however, an alignment of these two reference sequences showed that they differed by 5 bp. The phenotypic identification for all three isolates was *A. xylosoxidans/A. faecalis* (*A. piechaudii* was not in the Biolog database). Further phylogenetic analysis could not definitively identify this isolate to a single species. We considered *A. piechaudii* to be a correct identification given that the sequence was derived from a type strain; therefore, the isolate was unresolved and classified as *A. faecalis* or *A. piechaudii*.

Another example of discordance in species identification involved two isolates identified as *Corynebacterium amycola-*

tum by SmartGene, but as *Corynebacterium xerosis* (ATCC 373) by MicroSeq. Three of six SmartGene reference sequences for *C. xerosis* (X81914, AF145257, and AF024653) shared 100% homology to the MicroSeq reference sequence. Sequence X81914 was obtained from ATCC 373. Two of the SmartGene references for *C. xerosis* (M59058 and X84446) shared 98.7 to 98.9% homology with the MicroSeq reference. (Note that M59058 was derived from ATCC 373.). The remaining SmartGene reference sequence for *C. xerosis* (X81906) was derived from ATCC 7711 and shared 93.9% homology to the MicroSeq reference sequence. For *C. amycolatum*, MicroSeq contained reference DSM 455 and SmartGene contained four reference sequences (X84244, X82057, AY831726, and X82050). Two of the SmartGene reference sequences (X84244 and X82057) shared 100% homology with the MicroSeq reference strain. The remaining two SmartGene references (AY831726 and X82050), which were identical to each other, matched most closely to our two clinical isolates, but shared only 98.3% homology with the MicroSeq reference sequence. The phenotypic identification for both clinical isolates was *C. amycolatum*, which was our final identification.

Disagreement in species identification was common among *Pseudomonas* spp. For example, SmartGene identified one isolate as *Pseudomonas psychrotolerans* with a 100% identity score (type strain DSM 15758), whereas MicroSeq identified the isolate as *P. oryzihabitans* with a 100% identity score (*P. psychrotolerans* was not in MicroSeq). The isolate was identified phenotypically as *P. oryzihabitans*. Of the six *P. oryzihabitans* reference sequences in SmartGene, none shared more than 95% homology with our isolate. We resolved this isolate to be *P. psychrotolerans*. Additionally, SmartGene and MicroSeq identified an isolate with 100% identity scores as a member of the *Pseudomonas fluorescens* group, with SmartGene assigning the isolate to *Pseudomonas gessardii* or *Pseudomonas libaniensis* and MicroSeq assigning the isolate to *Pseudomonas synxantha* or *Pseudomonas mucidolens*. SmartGene contained references for *P. synxantha* and *P. mucidolens* sharing 99.4% homology with the query isolate. MicroSeq did not have entries for *P. gessardii* or *P. libaniensis*. Phenotypically, the isolate was identified as *P. fluorescens*. We assigned the organism to the *P. fluorescens* group because partial 16S rRNA gene sequencing was unable to discriminate among species in this group.

Overall, for the 63 discrepant isolates, the lack of represen-

TABLE 3. Representative examples of intraspecies variation
in partial 16S rRNA gene sequences

Species (no. of representative sequences in SmartGene)	No. of representative sequences in SmartGene matching single MicroSeq reference (% homology)	No. of isolates identified by SmartGene	Conventional identification
Aerococcus urinae (13)	0 (100)		
	3 (99.0–99.9)	1	*Aerococcus* sp.
	10 (98.0–98.9)	2	*A. urinae*
Stenotrophomonas maltophilia (58)	3 (100)	3	*S. maltophilia*
	15 (99.0–99.9)	2	*S. maltophilia*
	27 (98.0–98.9)	2	*S. maltophilia*
	11 (95.0–97.9)		
	2 (<95)		
Corynebacterium jeikeium (5)	1 (100)	3	*C. jeikeium*
	1 (99.0–99.9)		
	1 (98.0–98.9)	1	*C. jeikeium*
	1 (95.0–97.9)		
	1 (<95)		
Neisseria canis (5)	0 (100)		
	1 (99.0–99.9)		
	0 (98.0–98.9)		
	3 (95.0–97.9)	1	CDC group EF-4
	1 (<95)		

TABLE 4. Isolates identified to species level by conventional
methods and to genus level or unidentified by SmartGene

Conventional identification	SmartGene reference	% Identity score[a]
Brevibacterium otitidis	*Brevibacterium paucivorans*	97.8
Comamonas testosteroni	*Acidovorax delafieldii*	98.8
Corynebacterium jeikeium	*Cornynebacterium jeikeium*	98.8
Corynebacterium variabile	*Corynebacterium variabile*	98.3
Enterobacter cloacae	*Enterobacter* sp.[b]	99.0
Enterobacter cloacae	*Enterobacter* sp. *or Pantoea* sp.[b]	99.4
Enterococcus gallinarum	*Enterococcus gallinarum*	98.6
Escherichia coli	*Citrobacter braakii*	96.4
Haemophilus aphrophilus	*Haemophilus aphrophilus*	98.8
Ralstonia paucula	*Ralstonia paucula*	98.7
Rothia dentocariosa	*Rothia dentocariosa*	98.6
Stenotrophomonas maltophilia	*Stenotrophomonas maltophilia*	98.0
Weeksella virosa	*Naxibacter alkalitolerans*	96.7

[a] Criterion for species identification is ≥99% identity.
[b] SmartGene references at >99% identity contained a genus-level identification only.

tative species in the MicroSeq database accounted for 29 of these discrepancies. Additionally, inadequate sequence diversity in the MicroSeq database was observed and less sequence variation between reference sequences was present (Table 3).

Comparison of SmartGene and conventional identifications. By comparing the two methods, we found that 193 (64%) of the results were concordant, with 176 (59%) showing species concordance. Of the 107 isolates that disagreed, SmartGene identified three isolates (*Brachybacterium* sp., *Acinetobacter* sp., and *Actinobaculum* sp.) to the genus level that were unidentifiable by phenotypic methods; their phenotypic characteristics were consistent with the SmartGene identification. Additionally, SmartGene identified 66 to the species level, whereas phenotypic methods identified 61 to their corresponding genera, with only the remaining 5 unclassified.

Conventional methods identified 13 isolates to the species level that SmartGene identified to their corresponding genera (*n* = 11) or could not classify (*n* = 2). Table 4 compares conventional with their corresponding SmartGene identifications. All 13 discrepant isolates identified phenotypically were found in the SmartGene database. Seven of SmartGene's top identity scores corresponded to the phenotypic identification but did not meet our case definition for species identification of ≥99%. For nine isolates, SmartGene and MicroSeq identifications to the genus level only were in agreement.

Two isolates were identified as *Microbacterium* sp. phenotypically but were classified as unidentifiable using our criteria for sequence-based identification. For both isolates, SmartGene's top identity scores were *Microbacterium* sp.—namely, *Microbacterium testaceum* (96.5%) and *Microbacterium phyllosphaerae* (96.8%)—but did not meet our case definition for genus identification of ≥97%.

The results for seven isolates disagreed on genus. These isolates were identified phenotypically as *Actinomyces* sp., *Leuconostoc* sp., *Escherichia coli*, *Klebsiella pneumoniae*, *Comamo-*

nas testosteroni, *Chryseomonas* sp., and *Streptomyces* sp., and the corresponding SmartGene identifications were *Streptococcus mutans*, *Weissella confusa*, *Shigella sonnei*, *Raoultella ornithinolytica* or *Raoultella planticola*, *Ralstonia* sp., *Sphingomonas* sp., and *Microbacterium* sp., respectively. By constructing phylogenetic trees, we determined that SmartGene identifications were more accurate, with the exception of the *E. coli*/*S. sonnei* isolate, which was resolved as *E. coli* based on motility.

The results for 16 isolates agreed on the genus, but disagreed on the species between conventional methods and SmartGene (Table 5). Nine of these isolates were not included in the Biolog database. Two isolates did not have a representative strain in SmartGene. Eighty-one isolates were indistinguishable by 16S rRNA gene sequencing to more than one species and were assigned to a group (e.g., salivarius group) or, when clinically indicated, required further biochemical testing for definitive identification. Table 6 summarizes a few clinically relevant microorganisms with indistinguishable 16S rRNA sequences encountered in this study. Overall SmartGene, MicroSeq, and conventional identifications agreed for 193 (64%) isolates.

DISCUSSION

A reliable reference database for 16S rRNA sequences that is user friendly for bench technologists is important for the routine use of partial 16S rRNA gene sequencing in a clinical microbiology laboratory. In this study, partial 16S rRNA gene sequencing using SmartGene software was confirmed to be a more accurate method for microbial identification than conventional methods. To our knowledge, this is the first comprehensive study evaluating the ability of the SmartGene IDNS system to identify a diverse group of bacteria in a clinical laboratory and, additionally, the only direct comparison with the MicroSeq reference database. Our analysis of 300 clinical isolates illustrates that microbial identification improved with SmartGene compared to MicroSeq interrogations. Since SmartGene draws and compiles reference sequences from GenBank, it contains a large number of comparative sequences

TABLE 5. Discrepant species identification between SmartGene and conventional methods

| SmartGene identification | Partial 16S rRNA gene sequencing | | |
	No. of reference sequences matching ≥99%	Conventional identification	Highest homology (%) of SmartGene sequence corresponding to conventional identification
Aeromonas hydrophila or *A. media*	10 (8 *A. hydrophila*, 2 *A. media*)	*Aeromonas cavaie*	No reference in SmartGene
Arcanobacterium hemolyticum	1	*Arcanobacterium pyogenes*	90.4
Brevibacterium paucivorans	1	*Brevibacterium otitidis*	94.1
Campylobacter lari	7	*Campylobacter jejuni*	98.3
Comamonas kersterii	2	*Comamonas testosteroni*	94.4
Corynebacterium aurimucosum	4	*Corynebacterium minutissiumum*	98.8
Corynebacterium simulans	5	*Corynebacterium striatum*	99.2
Enterobacter aerogenes	1	*Enterobacter cloacae*	99.0
Pseudomonas migulae	1	*Pseudomonas mendocina*	95.9
Pseudomonas psychrotolerans	1	*Pseudomonas oryzihabitans*	94.6
Pseudomonas migulae	1	*Pseudomonas stutzeri*	96.6
Rothia dentocariosa	8	*Rothia mucilaginosa*	96.7
Staphylococcus caprae or *Staphylococcus capitis*	3 (2 *S. caprae*, 1 *S. capitis*)	*Staphylococcus saccharolyticus*	99.6
Streptococcus cristatus	3	*Streptococcus acidominimus*	No reference in SmartGene
Streptococcus gordonii	9	*Streptococcus anginosus* group	93
Streptococcus mitis or *S. parasanguis*	2	*Streptococcus infantarius*	94.1

for the same and different species, providing microbiologists with greater microbial diversity and increasing the ability to accurately identify microorganisms. In our study, 27 microorganisms were not represented in the MicroSeq database compared to only 2 not represented in SmartGene that impacted final identification. While MicroSeq potentially has higher-quality sequence data because it relies on type strains, we have demonstrated that use of a single type strain to represent entire taxa is often inadequate. The sequence heterogeneity of specific microorganisms within SmartGene is particularly valuable for identifying species, since many studies have demonstrated the existence of distinct interspecies and intraspecies variability (4, 8, 10, 12, 26). The software also allows selection of ambiguous base codes as a feature for BLAST searches which can be particularly useful for those microorganisms with high numbers of ambiguous bases (e.g., *Nocardia* sp. and *Moraxella* sp.) resulting from multiple copies of the 16S rRNA gene (13) Table 7 summarizes the features of several gene sequencing database systems.

Any reference database can be challenging because of the potential for erroneous reference sequences resulting from anomalous entries (1). To overcome this problem, we established a stringent protocol to identify reference sequences that are incorrectly named or with poor sequence quality. Our laboratory algorithm requires that the technologist review at least the top 20 reference sequences to determine the presence of "outliers." Once outliers are identified, they are compared to other references of the same and closely related organisms by constructing phylogenetic trees in SmartGene. Aberrant sequences are subsequently reported to SmartGene to modify their database. Clearly SmartGene lacks complete quality control for deposited sequences, and reference laboratories should be aware that identifications may require algorithms such as those described above for accurate identifications.

Currently, there is no standardized or consensus guideline to establish the case definition for species or genus classification by partial 16S rRNA sequencing (20, 23). Our case definition for species sharing ≥99% identity with a reference was based on previously published data (5, 6), and we developed a more rigorous criterion for genus identification which often limited our ability to identify several isolates (e.g., *Microbacterium* sp.). Those clinical isolates that shared less than 97% identity with available SmartGene reference sequences were reported as most closely related to the nearest relative from the reference database. In time, we believe that as more laboratories employ

TABLE 6. Representative common clinical isolates with indistinguishable partial 16S rRNA gene sequences

SmartGene identification	No. of isolates	Biochemical testing
Bacillus cereus group, including *B. anthracis*	4	Beta-hemolysis, motility, colony morphology, and penicillin disk
Escherichia sp. and *Shigella* sp.	10	Motility, indole, or lactose fermentation
Neisseria meningitidis and *Neisseria cinerea*	1	Colistin disk or conventional kits
Bordetella bronchiseptica, *Bordetella parapertussis*, or *Bordetella pertussis*	1	PCR or DFA for *B. pertussis*
Listeria monocytogenes and *Listeria innocua*	2	Beta-hemolysis

TABLE 7. Summary of advantages and disadvantages of database systems

Type of database	Cost	Database size (no. of sequences)	Comprehensive phylogenetic analysis	Creation of private sequence database	Automatic creation of searchable clinical isolate database	Flagging of questionable reference sequences	Quality control	Updates
SmartGene	$	~112,000	Yes	Yes	Yes	Yes	Partial	Weekly
MicroSeq	$	~2,000	Yes	Yes	No	No	All type strains	Periodically
Public domain	Free	>200,000	No	No	No	No	No	Daily

16S rRNA sequencing, the criteria for species and genus classifications will evolve and prove to be microorganism specific. For example, 16S rRNA gene sequences sharing ≥99.4% identity have already been proposed to identify *Mycobacterium* species (11).

Another important consideration for the routine use of partial 16S rRNA gene sequencing is the need for technical expertise and its cost. The SmartGene software and database have the advantage over software available in the public domain because they enable the less experienced non-molecular-bench technologist to perform the sequencing and interrogate and analyze gene sequences. To reduce the labor-intensive process of sequence editing and subsequent costs, we used the automated sequence analysis software program SeqScape, which has been proven to be an accurate and reliable method for sequence analysis which assigns quality values to each base call, trims sequences, and assembles forward and reverse sequences (7, 17). SmartGene plans to introduce their automated sequencing analysis software in the near future. While previous authors have supported the use of only the forward or reverse sequence (5) to reduce costs, we continue to use both forward and reverse sequences to resolve sequence discrepancies and evaluate the impact of copy variants within a single 16S rRNA genome (13). The high rate of intracellular polymorphisms within the coryneform group (~10% in this study) makes reliance on single-strand sequencing problematic and inadequate for prompt identification. Finally, to ensure judicious use of gene sequencing, we developed algorithms to screen for those isolates that can be adequately identified by conventional methods, with only a subset of isolates undergoing 16S rRNA gene sequencing routinely. For example we continue to use conventional methods to identify *Staphylococcus aureus*, *Listeria monocytogenes*, *Lactobacillus* sp., *Moraxella* sp., and *Eikenella* sp. In the 9 months this laboratory has employed partial 16S rRNA gene sequencing for bacteria, we experienced a labor savings of least one full-time equivalent certified medical technologist.

In summary, SmartGene IDNS has several advantages over MicroSeq in terms of the ability to identify microorganisms with its unique software interface. We found that SmartGene offers a database system with software enhancements, user-friendly programs, and the potential for labor savings for certain clinical laboratories. Overall, for those laboratories performing high-test-volume sequencing, the SmartGene database and software programs are excellent tools for providing comprehensive and accurate identification of a diverse group of bacteria.

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. **Becker, K., D. Harmsen, A. Mellmann, C. Meier, P. Schumann, G. Peters, and C. von Eiff.** 2004. Development and evaluation of a quality-controlled ribosomal sequence database for 16S ribosomal DNA-based identification of *Staphylococcus* species. J. Clin. Microbiol. **42:**4988–4995.
3. **Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler.** 2006. GenBank. Nucleic Acids Res. **34:**D16–D20.
4. **Beumer, A., and J. B. Robinson.** 2005. A broad-host-range, generalized transducing phage (SN-T) acquires 16S rRNA genes from different genera of bacteria. Appl. Environ. Microbiol. **71:**8301–8304.
5. **Bosshard, P. P., S. Abels, R. Zbinden, E. C. Böttger, and M. Altwegg.** 2003. Ribosomal DNA sequencing for identification of aerobic gram-positive rods in the clinical laboratory (an 18-month evaluation). J. Clin. Microbiol. **41:**4134–4140.
6. **Bosshard, P. P., S. Abels, M. Altwegg, E. C. Böttger, and R. Zbinden.** 2004. Comparison of conventional and molecular methods for identification of aerobic catalase-negative gram-positive cocci in the clinical laboratory. J. Clin. Microbiol. **42:**2065–2073.
7. **Christensen, J. J., K. Andresen, T. Justensen, and M. Kemp.** 2005. Ribosomal DNA sequencing: experiences from use in the Danish National Reference Laboratory for identification of bacteria. APMIS **113:**621–628.
8. **Christensen, J. J., M. Kilian, V. Fussing, K. Andresen, J. Blom, B. Korner, and A. G. Steigerwalt.** 2005. *Aerococcus urinae*: polyphasic characterization of the species. APMIS **113:**517–525.
9. **Clarridge, J. E., III.** 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin. Microbiol. Rev. **17:**840–862.
10. **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.
11. **Cloud, J. L., H. Neal, R. Rosenberry, C. Y. Turenne, M. Jama, D. R. Hillyard, and K. C. Carroll.** 2002. Identification of *Mycobacterium* spp. by using a commercial 16S ribosomal DNA sequencing kit and additional sequencing libraries. J. Clin. Microbiol. **40:**400–406.
12. **Coenye, T., and P. Vandamme.** 2003. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. FEMS Microbiol. Lett. **228:**45–49.
13. **Conville, P. S., and F. G. Witebsky.** 2005. Multiple copies of the 16S rRNA gene in *Nocardia nova* isolates and implications for sequence-based identification procedures. J. Clin. Microbiol. **43:**2881–2885.
14. **Daley, P., D. L. Church, D. B. Gregson, and S. Elsayed.** 2005. Species-level molecular identification of invasive "*Streptococcus milleri*" group clinical isolates by nucleic acid sequencing in a centralized regional microbiology laboratory. J. Clin. Microbiol. **43:**2987–2988.
15. **Drancourt, M., P. Berger, and D. Raoult.** 2004. Systematic 16S rRNA gene sequencing of atypical clinical isolates identified 27 new bacterial species associated with humans. J. Clin. Microbiol. **42:**2197–2202.
16. **Ferroni, A., I. Sermet-Gaudelus, E. Abachin, G. Quesne, G. Lenoir, P. Berche, and J.-L. Gaillard.** 2002. Use of 16S rRNA gene sequencing for identification of nonfermenting gram-negative bacilli recovered from patients attending a single cystic fibrosis center. J. Clin. Microbiol. **40:**3793–3797.
17. **Gautier, A.-L., D. Dubois, F. Escande, J.-L. Avril, P. Trieu-Cuot, and O. Gaillot.**

2005. Accurate identification of human isolates of *Pasteurella* and related species by sequencing the *sodA* gene. J. Clin. Microbiol. **43:**2307–2314.

18. **Kumar, S., K. Tamura, and N. Nei.** 2004. MEGA 3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. **5:**150–163.

19. **Murray, P. R., E. J. Baron, J. H. Jorgensen, M. A. Pfaller, and R. H. Yolken (ed.).** 2003. Manual of clinical microbiology, 8th ed. American Society for Microbiology, Washington, D.C.

20. **Ochman, H., E. Lerat, and V. Daubin.** 2005. Examining bacterial species under the specter of gene transfer and exchange. Proc. Natl. Acad. Sci. USA **102:**6595–6599.

21. **Pace, N. R.** 1997. A molecular view of microbial diversity and the biosphere. Science **276:**734–740.

22. **Patel, J. B., R. J. Wallace, Jr., B. A. Brown-Elliott, T. Taylor, C. Imperatrice, D. G. B. Leonard, R. W. Wilson, L. Mann, K. C. Jost, and I. Nachamkin.** 2004. Sequence-based identification of aerobic actinomycetes. J. Clin. Microbiol. **42:**2530–2540.

23. **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.

24. **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.

25. **Tang, Y.-W., A. von Graevenitz, M. G. Waddington, M. K. Hopkins, D. H. Smith, H. Li, C. P. Kolbert, S. O. Montgomery, and D. H. Persing.** 2000. Identification of coryneform bacterial isolates by ribosomal DNA sequence analysis. J. Clin. Microbiol. **38:**1676–1678.

26. **Vasquez, A., G. Molin, B. Pettersson, M. Antonsson, and S. Ahrne.** 2005. DNA-based classification and sequence heterogeneities in the 16S rRNA genes of *Lactobacillus casei/paracasei* and related species. Syst. Appl. Microbiol. **28:**430–431.

27. **Woo, P. C. Y., K. H. L. 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 bacterial identification system for identification of clinically significant bacterial isolates with ambiguous biochemical profiles. J. Clin. Microbiol. **41:**1996–2001.