Simultaneous Sequence Analysis of the 16S rRNA and $rpoB$ Genes by Use of RipSeq Software To Identify Mycobacterium Species

Keith E. Simmon,1* Øyvind Kommedal,2,3 Øystein Saebo,3 Bjarte Karlsen,3 and Cathy A. Petti1,4

**JAU** MJol and **UT** Motal Pathology, 500 Chipeta Way, Salt Lake City, UT 84108. Phone: (801) 583-2787, ext. 2035. Fax: (801) 584-5109. E-mail: keith.simmon@aruplab.com.

Received 22 February 2010/Returned for modification 9 June 2010/Accepted 24 June 2010

The 16S rRNA gene is commonly used to identify *Mycobacterium* spp., but alternative DNA targets can provide better resolution to the species level. We evaluated a novel system that enables simultaneous amplification, sequencing, and analysis of two different DNA targets in a single tube to identify clinical isolates of *Mycobacterium* spp. For 139 clinical isolates, we found that the 16S rRNA gene alone identified 67 (48%) isolates as single species, 68 (49%) isolates to the complex or group level, and 4 (3%) isolates to the genus level only. The $rpoB$ gene alone identified 117 (84%) isolates as single species, 10 (7%) isolates to the complex or group level, and 12 (8%) isolates to the genus level only. Combining the separate analyses for sequencing of two gene targets, 119 (86%) isolates were identified as single species and 10 (7%) isolates were identified to the complex or group level. Seven (5%) isolates were identified as novel species within established groups, and 3 (2%) were identified to the genus level only. Dual-locus identification identified 110 (79%) isolates as single species and 22 (16%) isolates to the complex or group level. Six (4%) were identified as novel species within established groups, and 1 (1%) was identified to the genus level only. Identifications were more accurate when both the 16S rRNA and $rpoB$ genes were screened, and reliance on a single gene target was suboptimal. RipSeq dual-locus software provides an accurate alternative method for laboratories using two different gene targets for microorganism identification.

Accurate identification of *Mycobacterium* species is important because the clinical relevance and susceptibility patterns differ among species and impact clinical decision-making. Routine molecular identification of *Mycobacterium* species is often hindered by the lack of genetic heterogeneity within the 16S rRNA gene. Other DNA targets have been used to resolve taxonomic uncertainties, and studies have demonstrated their use for identification in clinical laboratories (1–3, 5, 7, 11, 16). Recently, more scrutiny has been applied to these targets, particularly on their relative values for describing new species, and data have suggested that reliance on a single DNA target for identification may have limitations (14, 15, 20). Routine implementation of alternative targets for identification is limited by cost, longer turnaround times, a lack of references in databases, a lack of consensus on the optimal target or region within a DNA target (1, 10), and uncertainty about expected intraspecies variability (14, 18).

We hypothesized that concurrent sequencing of the 16S rRNA gene and an alternative DNA target could provide better discrimination for identification of *Mycobacterium* spp., and we sought a method that would not significantly impact workflow, cost, or analysis time. RipSeq software (iSentio, Bergen, Norway) offers a solution for laboratories that use two genetic loci for identification by enabling simultaneous amplification, sequencing, and analysis of two different DNA targets in a single tube. The RipSeq dual-locus identification system has an algorithm that analyzes mixed electropherograms obtained by sequencing of two different gene targets, searches two different databases, each representing one of the amplified targets, and provides a ranked list of best-match results for each target gene. Using clinical isolates, we evaluated the added value of using two DNA targets for identification of *Mycobacterium* sp. by standard laboratory protocols performed according to CLSI guidelines (6). We also assessed the accuracy of the RipSeq dual-locus identification system for identification of *Mycobacterium* spp. by comparing the identity scores of consensus sequences obtained by sequencing of two gene targets individually to those obtained by the simultaneous dual-locus method.

**MATERIALS AND METHODS**

In April 2008 and August 2008, we prospectively collected consecutive clinical isolates that were identified at ARUP Laboratories as *Mycobacterium* spp. by partial 16S rRNA gene sequencing. Template DNA was prepared by a 15-min boil of isolate suspensions in PrepMan Ultra sample preparation reagent (Applied Biosystems, Foster City, CA). Amplification was performed in a 25-µl reaction mixture containing 1× FastStart PCR master mix (Roche Diagnostics Corporation, Alameda, CA), 0.2 µM concentration of each primer, and 2 µl of template. Primers for amplification of the 16S rRNA and $rpoB$ (2) genes were 16S-5F (5’-TTGGAGAGTTTGATCCTGGCTC-3’), 16S-3R (5’-TACCCGGGCTGTCTGGCAC-3’), $rpoB$-mycoF (5’-GGCAAGGTCACCCGGAAAGG-3’), and $rpoB$-mycoR (5’-AGCCGGTGCTGGTGATCATC-3’). PCR mixtures were amplified by an initial hold at 95°C for 10 min and then 35 cycles of denaturing at 95°C for 30 s, annealing at 55°C for 30 s, and extension at 72°C for 1 min 30 s. The reaction ended with a final extension at 72°C for 2 min and a hold at 4°C. A single amplification reaction mix was prepared for each isolate and contained primers for both the 16S rRNA and $rpoB$ genes. Gel electrophoresis was used to confirm the amplification of both targets (16S rRNA gene product, ~500 bp; and $rpoB$ product, ~750 bp). PCR products were purified using Ampure beads (Agencourt Bioscience Corp., Beverly, MA) per the manufacturer’s
instructions and were sequenced with the following primers or primer combinations: 16S-5F, 16S-534R, rpoB-mycoF, rpoB-mycoR, 16S-5F plus rpoB-mycoF, and 16S-534R plus rpoB-mycoR. Sequencing was done using Big Dye Terminator reagents on a model 3730xl sequencer (Applied Biosystems) by a standard automated sequencer protocol.

Editing of sequences generated with a single primer was done by manual inspection, using Contig Express in the Vector NTI suite (Invitrogen, Carlsbad, CA). Edited sequences were aligned in MEGA 4.0 (19). All sequences were trimmed to the same size for comparison. Bootstrapped neighbor-joining trees were constructed using the Jukes-Cantor parameter. Dual-locus electropherograms were uploaded and bulk analyzed directly in the RipSeq software.

Identification of isolates by use of individual gene targets. Isolate identifications based on the 16S rRNA gene were based on CLSI guidelines (6), which state that 100% identity to a reference sequence is required for species identification. Identification with the rpoB gene was performed using criteria outlined by Adekambi et al., who recommended an identity of 98.0 to 100% for species identification (1, 2). GenBank reference sequences were used for identification for both the 16S rRNA and rpoB genes.

Analysis of sequences with RipSeq software. All samples were initially run with standard settings. The basis for the dual-locus module is the RipSeq algorithm, which has been described previously (12, 13). The dual-locus module includes the RipSeq base-calling procedure and a modified alignment strategy. The basic principles of the search are to divide the mixed electropherograms into smaller pieces and to construct every possible combination of bases (words) within each piece. Each single word is subsequently compared to the corresponding area of the reference sequence in the database. Before a RipSeq search can be performed, three parameters are set. The first two are the 5′-end trimming and 3′-end trimming of the electropherograms. The third is the y axis cutoff. A y axis cutoff is used to avoid base calling of low signals from the baseline of the electropherograms, as these would significantly reduce the specificity of the analysis.

Because the RipSeq dual-locus system is intended for use on cultured isolates and the ratio between the different targets is constant, the signal peak height and technical quality of the electropherograms should be similar every time. This observation enables establishment of standard parameter settings and bulk analysis of multiple samples. For each sample, the program will generate a consensus answer based on the forward and reverse sequences for both gene targets. The consensus answer will be generated only when there is agreement within all results. If one of the results is not in agreement or if one of the targets is not detected, a warning will be displayed, and the available results for that sample must be evaluated manually.

Interpretation by the RipSeq algorithm. A RipSeq interpretation scheme set identification to the species level by use of the 16S rRNA gene to 99.5 to 100% and identification to the species level by use of the rpoB gene to 97.8 to 100%. When an rpoB sequence matched a reference sequence with an identity of >99.0%, the algorithm weighted the rpoB result with a higher priority. When an identity score obtained using the rpoB gene reference did not meet species guidelines, the database was interrogated. When the database lacked an rpoB gene reference for the species, 16S rRNA gene identification was used. When a corresponding rpoB reference was in the database but below the cutoff value, we classified the isolate as a Mycobacterium species closely related to its corresponding 16S rRNA gene identity.

The quality and accuracy of base calling for the 16S rRNA and rpoB genes were determined by inferring a consensus sequence based on the variance in identity scores toward the top reference for the forward and reverse sequencing reactions. The identity scores toward the reference were compared for the dual-locus method and the consensus sequences of individual gene targets.

RESULTS

16S rRNA gene identification. Alignment of 462 bp of the 16S rRNA genes of all 139 isolates gave 28 (20%) unique DNA sequences, with 72 (16%) variable positions over the aligned area (Fig. 1). The largest cluster of isolates sharing an identical sequence belonged to the Mycobacterium chelonae/M. abscessus complex (n = 38), and this cluster represented the majority of clinical isolates. M. gordonae was the second largest group and comprised three sequence variants, with 17, 10, and 1 isolate of the variants. The difference between these three variants was 2 to 4 bp, and each had 100% identity to a GenBank reference annotated as M. gordonae. Overall, 135 (97%) isolates shared 100% identity, 2 shared 99.8% identity, and 1 isolate each shared 99.6% and 99.4% identities to a GenBank reference sequence. Using CLSI guidelines, 67 (48%) isolates were identified as single species, 68 (49%) isolates were identified to the complex or group level, and 4 (3%) were identified to the genus level only (Table 1).

rpoB gene identification. Alignment of 684 bp of the rpoB gene gave 70 (50%) unique DNA sequences with 260 (38%) variable positions. Among the 38 M. chelonae/M. abscessus complex isolates with identical 16S rRNA gene sequences, 14 different rpoB gene sequences were obtained. The largest cluster of isolates sharing identical rpoB sequences belonged to M. abscessus (n = 25) (Fig. 2). Twenty-eight isolates identified as M. gordonae by 16S rRNA gene sequencing contained the largest variability in the rpoB gene sequence. Overall, 65 (47%) isolates shared 100% identity, 63 (45%) shared 98.0 to 99.9% identity, and 11 (8%) shared <98.0% identity to a GenBank reference sequence. Using guidelines set forth by Adekambi et al. (1, 2), 117 (84%) isolates were identified as single species, 10 (7%) isolates were identified to the complex or group level, and 12 (8%) were identified to the genus level only (Table 1).

Comparison of 16S rRNA and rpoB genes. With integration of the rpoB gene, 50 additional organisms were identified to the species level compared with identification by use of the 16S rRNA gene alone (117 [84%] versus 67 [48%] isolates). Most notably, the rpoB gene distinguished between members of the M. chelonae/M. abscessus complex, M. kansasi from M. gastri, M. mucogenum from M. phocaicum, and M. conceptionens from M. senegalense. Four isolates that were identified to the genus level only by both targets probably represented undescribed species in the Mycobacterium genus. For an additional 6 isolates, 16S rRNA gene results indicated species-level descriptions, although the rpoB sequence data may suggest undescribed species. Five of these isolates were related to M. gordonae, and one was a member of the M. chelonae/M. abscessus complex. A lack of agreement was also observed between 16S rRNA and rpoB gene identifications for organisms within the M. fortuitum group and the M. avium complex. Genetic heterogeneity was observed among 28 clinical isolates in the M. gordonae group, with 3 unique 16S rRNA gene sequences and 14 unique rpoB sequences. Compared to all other groups, this group exhibited the most rpoB gene heterogeneity, and 5 isolates shared <98% identity to a GenBank reference sequence. Combining the separate analyses of the sequences of two gene targets, 119 (86%) isolates were identified as single species and 10 (7%) isolates were identified to the complex or group level. Seven (5%) were identified as novel species within established groups, and 3 (2%) were identified to the genus level only (Table 1).

Dual-locus identification with 16S rRNA and rpoB genes by use of RipSeq software. The dual-locus algorithm was able to identify closely related 16S rRNA gene reference sequences for 138 of 139 samples without manual inspection of the mixed electropherograms or database scrutiny. For one sample, the reverse sequence was not successfully matched to the 16S rRNA gene reference, and manual inspection of the electropherograms was required. Comparison of the percent identities between the forward and reverse sequences showed an
average percent identity variance of 0.2%; the lowest variance was 0.0%, and the highest was 0.5%, with a median of 0.2%.

The dual-locus algorithm was able to identify closely related rpoB gene reference sequences for 134 of 139 samples without manual inspection of the electropherograms or database scrutiny. Of the 5 failures, 4 were due to a lack of reference sequences in GenBank to place in the rpoB gene database. All 4 samples were identified using 16S rRNA gene identities. The remaining failure was due to lower peaks in the electropherogram for the rpoB gene than in that for the 16S rRNA gene.

Comparison of the percent identities between the forward and reverse sequences of successful samples showed an average percent identity variance of 0.3%; the lowest variance was 0.0%, and the highest variance was 3.0%, with a median variance of 0.2%. When we reviewed the isolate with the highest variance, we found that this was due to very poor technical quality in the nonmixed 3’/H11032 end of the forward electropherogram. After reanalysis, the variance for this particular isolate was reduced to 0.1%, and the maximum variance for any isolate was 1.7%.

With the dual-locus algorithm, 110 (79%) isolates were identified as single species, and 22 (16%) isolates were identified to the complex or group level. Six (4%) were identified as novel species within established groups, and 1 (1%) was identified to the genus level only (Table 1). Comparison of the percent identities for the averaged forward and reverse sequence reactions from the RipSeq algorithm to those for the consensus sequences of individual reactions showed average percent identity variances of 0.1% (range, 0.0 to 0.6%; median, 0.1%) and 0.3% (range, 0.0 to 1.7%; median, 0.2%) for the 16S rRNA and rpoB genes, respectively.

**DISCUSSION**

In this study, we examined identifications obtained by use of individual gene targets and then compared them to those obtained by a novel method for sequencing two loci in a single reaction tube. Regardless of method, identifications were more accurate when both the 16S rRNA and rpoB genes were
screened. Although the rpoB gene target alone provided a larger number of species-level identifications than the 16S rRNA gene, the 16S rRNA gene allowed us to evaluate isolates nearing rpoB cutoff guidelines. For example, using the rpoB gene, 12 isolates could not be identified further than the genus level, compared with only 4 isolates identified by use of the 16S rRNA gene. Results from 16S rRNA gene sequencing also improved our ability to interpret rpoB gene sequence data. Additionally, the 16S rRNA gene allowed us to classify isolates in the correct phylogenetic context in the absence of species descriptions or rpoB gene references. Based on these observations, we believe that reliance on either the 16S rRNA or rpoB gene alone is not the optimal approach for definitive identification of Mycobacterium species.

In addition to demonstrating the importance of two DNA targets for species-level identification, we evaluated an alternative to multitarget DNA analysis by using the RipSeq dual-locus module. Previously, the RipSeq algorithm was shown to resolve mixed electropherograms for a single gene (12, 13). In this study, we demonstrated that this software provides a reliable algorithm to accurately identify Mycobacterium species by discriminating two gene targets in a single sample, without a significant loss of fidelity. Performance of dual-locus sequencing in a single reaction tube, as opposed to individual reaction tubes for the two gene targets, had the added advantages of limiting overall reagent costs and labor and improving turnaround times for reporting results to the species level. The editing of individual sequences and their final analyses were the most time-consuming steps, as it took about 10 min to assemble, edit, and export a consensus sequence and to perform a BLAST search for each sequence. We calculated that these steps would involve approximately 2 to 3 days of work for 139 isolates. The automation in the RipSeq application allowed for the identification of the 139 isolates in roughly 4 h, including evaluation of a few suboptimal electropherograms. Even with some manual adjustment using the bulk run feature, analysis of samples was significantly faster than that by the conventional individual sequencing method.

One of the limitations of RipSeq software is an inability to construct consensus sequences. Mixed samples require separate analyses of the forward and reverse electropherograms. Also, the base-calling step cannot differentiate which base comes from a particular template; instead, base-calling results in a string of DNA letters containing a large number of IUPAC codes (e.g., Y = T and C) that are used directly in the subsequent matching with the reference databases. As a consequence, the actual template sequences cannot be derived directly from the dual-locus base-calling method. In cases of new variants or novel species, the RipSeq identity scores will indicate them, but confirmation and further investigations will require individual sequencing of the respective targets. Additionally, for dual-locus sequences, sequence quality currently

FIG. 2. Neighbor-joining tree of 684-bp region of rpoB genes of 139 clinical isolates and reference strains. Branch support is recorded at nodes, as a percentage of 100 bootstrap iterations. Footnote symbol a indicates clinical isolates 1, 8, 13, 21, 25, 27, 28, 40, 51, 52, 53, 54, 61, 69, 73, 85, 95, 100, 118, 119, 128, 130, 135, 150, and 155.
cannot be assessed using traditional phred scores (8). As an alternative, we assessed the quality of sequences based on the identity scores obtained from each forward and reverse sequence and measured the overall identity variance. When the analysis algorithm was followed, the average variances were 0.2% and 0.3% for the 16S rRNA and rpoB genes, respectively. More variance was observed with the rpoB gene, and this observation can be explained partly by the larger proportion of nonperfect matches in the database for the rpoB gene.

We selected the 16S rRNA and rpoB genes because they have been well-recognized and accepted targets for identification of Mycobacterium spp. (1–3, 9). Prior investigations have utilized type or well-characterized strains to evaluate the performances of alternative targets for microorganism identification and, similarly, used well-characterized strains and/or type strains as references. We believe that the strengths of our study are that we (i) used only clinical isolates and (ii) did not limit the reference set to only type strains for identification, instead using all available reference sequences in GenBank. For the 16S rRNA gene, a type strain sequence was available for each species identified in this study. Not surprisingly, type strain sequences were much more limited for the region of the rpoB gene and were not available for M. arupense, M. kansasi, M. gastri, M. gordonea, and M. szulzig. Although the 16S rRNA and rpoB genes were robust targets, we found a lack of interspecies variability in the rpoB gene for both the M. fortuitum and M. avium complex isolates, particularly for M. fortuitum and M. houstonense or M. intracellulare and M. chimaera. Prior studies revealed that there is a lack of consensus about the phylogeny of these complexes, and without the use of multiple DNA targets, definitive species identifications cannot be determined reliably (4, 17).

Identities were more accurate when both the 16S rRNA and rpoB genes were screened, and reliance on a single gene target was suboptimal. The RipSeq dual-locus software provides an accurate, timesaving, and cost-effective alternative method for laboratories that use two different gene targets for microorganism identification. Although more isolates were identified to the group rather than species level with the dual-locus module, this observation is a reflection of known and still unresolved phylogenetic ambiguities for particular Mycobacterium sp., such as the M. fortuitum and M. avium complexes. In this study, we illustrated the potential of same-tube sequencing of two well-characterized gene regions for the identification of mycobacteria. This tool also has the potential to couple gene regions important in antibiotic resistance, virulence, and/or identification.

REFERENCES