A method for the elimination of misidentified organisms in rRNA sequence alignments for improved primer design and diagnostic assay performance

Friday, June 2, 2017
Poster 761

R. Lum[1], J. Hogan[2], and J. Liu[1]
[1]Qvella Corporation, Richmond Hill, ON, Canada [2]JJ Hogan & Associates, San Diego, CA, USA

Qvella Corporation
ronniel@qvella.com
289-317-0414

Introduction

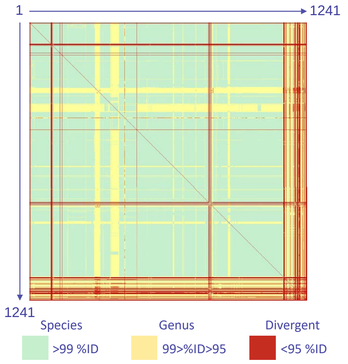
With the introduction of high-throughput sequencing methods, collections of rRNA sequences, such as the Ribosomal Database Project (RDP), SILVA, and Greengenes, have grown exponentially. The accuracy of these databases is crucial for developing primers/probes for molecular diagnostic tests for pathogenic organisms with sufficient specificity and target coverage. Major sources of error are poor quality sequencing, errors during assembly, sequence chimeras, and the improper identification of organisms by non-molecular-based methods. To address this problem, we have developed a Composite Variable Region Analyzer (COVA™), a tool to detect potentially misidentified sequences and to elucidate their true identities. Furthermore, we used COVA™ to assess the extent to which database entries have been misidentified using *Klebsiella pneumoniae* as a model. Finally, we highlight a method to assess mismatches in base-pairing regions of rRNA (stems) for validity. Since the conservation of rRNA secondary structures exceeds that of its nucleotides[1], a mismatch where no compensatory substitution is found to preserve the structure is likely a sequencing error.

Methodology

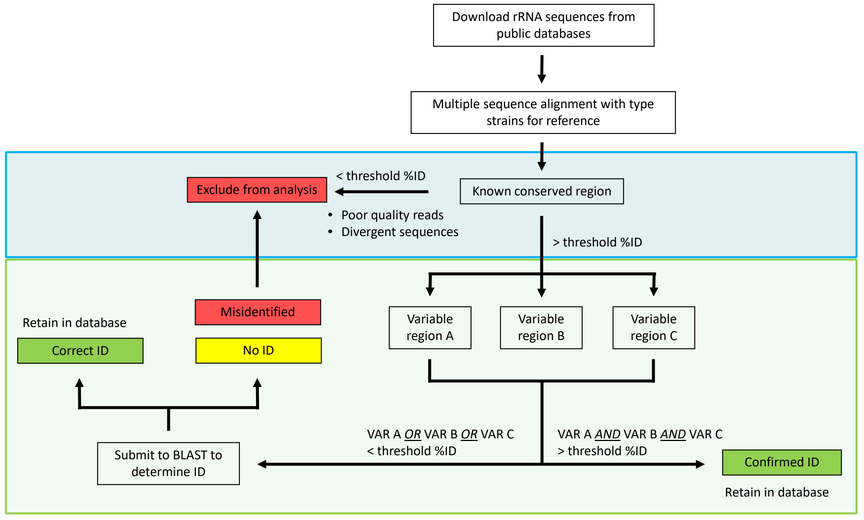
K. pneumoniae rRNA sequences (16s) were obtained from the RDP and Silva databases. Sequences were aligned and analyzed using CLUSTALW2 and GENEIOUS. Sequences with Percent Identity below a threshold compared to type strains in a known conserved 100 bp segment were classified as ambiguous and discarded. Alignments of three variable regions (between 50-100 bp in length) were constructed and distance matrices of each variable region were generated by pair-wise comparison to reference sequences of type strains. Sequences that met a minimum Percent Identity threshold for all of the variable regions were defined as putative *K. pneumonia* sequences. Sequences that did not meet the minimum threshold criteria were submitted to BLAST for identification.

Misidentification among *K. pneumoniae* sequences

- Percent Identity was calculated for every pairwise alignment between all 1241 16s rRNA sequences (>1.5 million alignments)
- Percent Identity values were colour-coded to generate a heat map; colours approximate taxonomic distinction
- Bands of yellow and red (12.8% of the entire database) indicate divergent sequences from the majority of *K. pneumoniae* sequences



COVA™ Algorithm



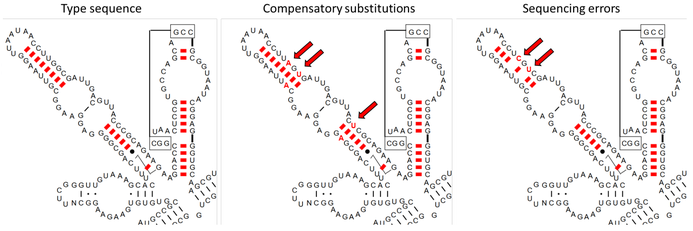
COVA™ analysis of *K. pneumoniae* sequences

- *K. pneumoniae* sequences (16s) were analyzed by COVA™
- Misidentified sequences were identified as *Enterobacter aerogenes*, *Raoultella ornithinolytica*, or other *Enterobacteriaceae* members
- It is difficult to biochemically distinguish between *K. pneumoniae* and some strains of *E. aerogenes*, which have different antibiotic sensitivity profiles[2]

COVA analysis - *K. pneumoniae*	
Sequences analyzed	1241
Confirmed ID	1181
Excluded due to low %ID at CON fragment	35
Misidentified	25
% Misidentified	2%

Mismatch confirmation by preservation of secondary structure

- The majority of ribonucleotides in rRNA are base paired:
  - Canonical pairing (G-C, A-U)
  - Allowable non-canonical pairing (G-U, G-A, U-U)



- A mismatch of a base that is paired in the rRNA structure must have a compensatory substitution in the complementary base to preserve the rRNA secondary structure
- If no compensatory substitution is found, the mismatch is likely a sequencing error

Conclusions

- Misidentification of sequences is a real problem
- Composite Variable Region Analyzer (COVA™) can detect potentially misidentified sequences and elucidate their true identities to:
  - Distinguish between intra-species sequence variation and misidentification
  - Eliminate misidentified sequences from analysis to facilitate assay design
- A mismatched base that is paired in the secondary structure must have a compensatory substitution in the complementary base to preserve the rRNA structure

[1]Kjer K. Use of ribosomal-RNA secondary structure in phylogenetic studies to identify homologous positions—an example of alignment and data presentation from the frogs. Mol Phylogenet Evol. 1995; 4(3):314–330.

[2]Claeys *et al.*, Extended-Spectrum β-lactamase (ESBL) producing *Enterobacter aerogenes* phenotypically misidentified as *Klebsiella pneumoniae* or *K. terrigena*. BMC Microbiol. 2004; 4: 49.