Diagnostic Microbiology and Infectious Disease 78 (2014) 105–115

Contents lists available at ScienceDirect

Diagnostic Microbiology and Infectious Disease

journal homepage: www.elsevier.com/locate/diagmicrobio

Evaluation of 16SpathDB 2.0, an automated 16S rRNA gene sequence database, using 689 complete bacterial genomes

Jade L.L. Teng [a,b], Tom C.C. Ho [a], Ronald S.Y. Yeung [a,c], Annette Y.P. Wong [a], Haiyin Wang [d], Chen Chen [d], Kitty S.C. Fung [c], Susanna K.P. Lau [a,b,e,f,*], Patrick C.Y. Woo [a,b,e,f,*]

[a] Department of Microbiology, The University of Hong Kong, Hong Kong, China
[b] Research Centre of Infection and Immunology, The University of Hong Kong, Hong Kong, China
[c] Department of Pathology, United Christian Hospital, Hong Kong, China
[d] National Institute for Communicable Disease Control and Prevention, Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control, Beijing, China
[e] State Key Laboratory of Emerging Infectious Diseases, Hong Kong, China
[f] Carol Yu Centre for Infection, The University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 20 July 2013
Received in revised form 21 October 2013
Accepted 27 October 2013
Available online 4 November 2013

Keywords:
16SpathDB
16S rRNA
Automated
Database
Genome
Evaluation
Bacteria

ABSTRACT

Interpretation of 16S rRNA sequences is a difficult problem faced by clinical microbiologists and technicians. In this study, we evaluated the updated 16SpathDB 2.0 database, using 689 16S rRNA sequences from 689 complete genomes of medically important bacteria. Among these 689 16S rRNA sequences, none was wrongly identified, with 35.8% reported as a single bacterial species having >98% identity with the query sequence (category 1), 63.9% reported as more than 1 bacterial species having >98% identity with the query sequence (category 2), 0.3% reported to the genus level (category 3), and none reported as no match (category 4). For the 16S rRNA sequences of non-duplicated bacterial species reported as category 1 or 2, the percentage of bacterial species reported as category 1 was significantly higher for anaerobic Gram-positive/Gram-negative bacteria than aerobic/facultative anaerobic Gram-positive/Gram-negative bacteria. 16SpathDB 2.0 is a user-friendly and accurate database for 16S rRNA sequence interpretation in clinical laboratories.

1. Introduction

In the 1970s, Carl Woese and others started to analyze and sequence the 16S rRNA genes of various bacteria, using DNA sequencing, a state-of-the-art technology at that time, and used the sequences for phylogenetic studies (Fox et al., 1977; Gupta et al., 1983). The invention of polymerase chain reaction and automated DNA sequencing and their widespread use in the last 20 years have enabled 16S rRNA sequencing to play a pivotal role in accurate identification of medically important bacteria in both clinical microbiology and research laboratories (Lau et al., 2013; Woo et al., 2008). On a patient-to-patient basis, accurate identification is crucial in determining the choice and duration of antibiotics and the appropriate infection control measure (Woo et al., 2002a, 2002b). On a population scale, accurate identification is important in analyzing the epidemiology, antibiotic resistance patterns, treatment plans, and outcomes of infections associated with a particular bacterium (Lee et al., 2003; Woo et al., 2001, 2004).

Although 16S rRNA sequencing has had great impact on clinical microbiology and infectious diseases, interpretation of 16S rRNA sequence results is one of the most difficult problems faced by inexperienced clinical microbiologists and technical staff. There is a lack of a universal threshold value or cutoff for species assignment, as different levels sequence diversities are observed among different bacterial taxa, which evolve at different rates. While a >97% similarity level has been previously proposed for bacterial speciation, a >0.5% difference may be indicative of a new species (Stackebrandt and Goebel, 1994; Palys et al., 1997). Recently, a more relaxed value of 98.5% has been proposed by Stackebrandt and Ebers (2006) after inspecting a large dataset about the relationship between rRNA gene sequence identities and DNA-DNA hybridization (DDH) values. Perhaps the above controversy can be partly solved when the present emphasis on DDH values for species delineation shifts to genome-based approach in the near future. Moreover, the available software and databases for analysis of 16S rRNA sequences are associated with problems that make them not ideal for clinical microbiology laboratory users (Woo et al., 2008). For example, as a result of the large number of unvalidated 16S rRNA sequences in GenBank, it is often not easy for inexperienced users to decide whether the "first hit" is the real identity of a

* Corresponding authors. Tel.: +852-22554892; fax: +852-28551241.
E-mail addresses: skplau@hkucc.hku.hk (S.K.P. Lau), pcywoo@hku.hk (P.C.Y. Woo).

0732-8893/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.diagmicrobio.2013.10.019

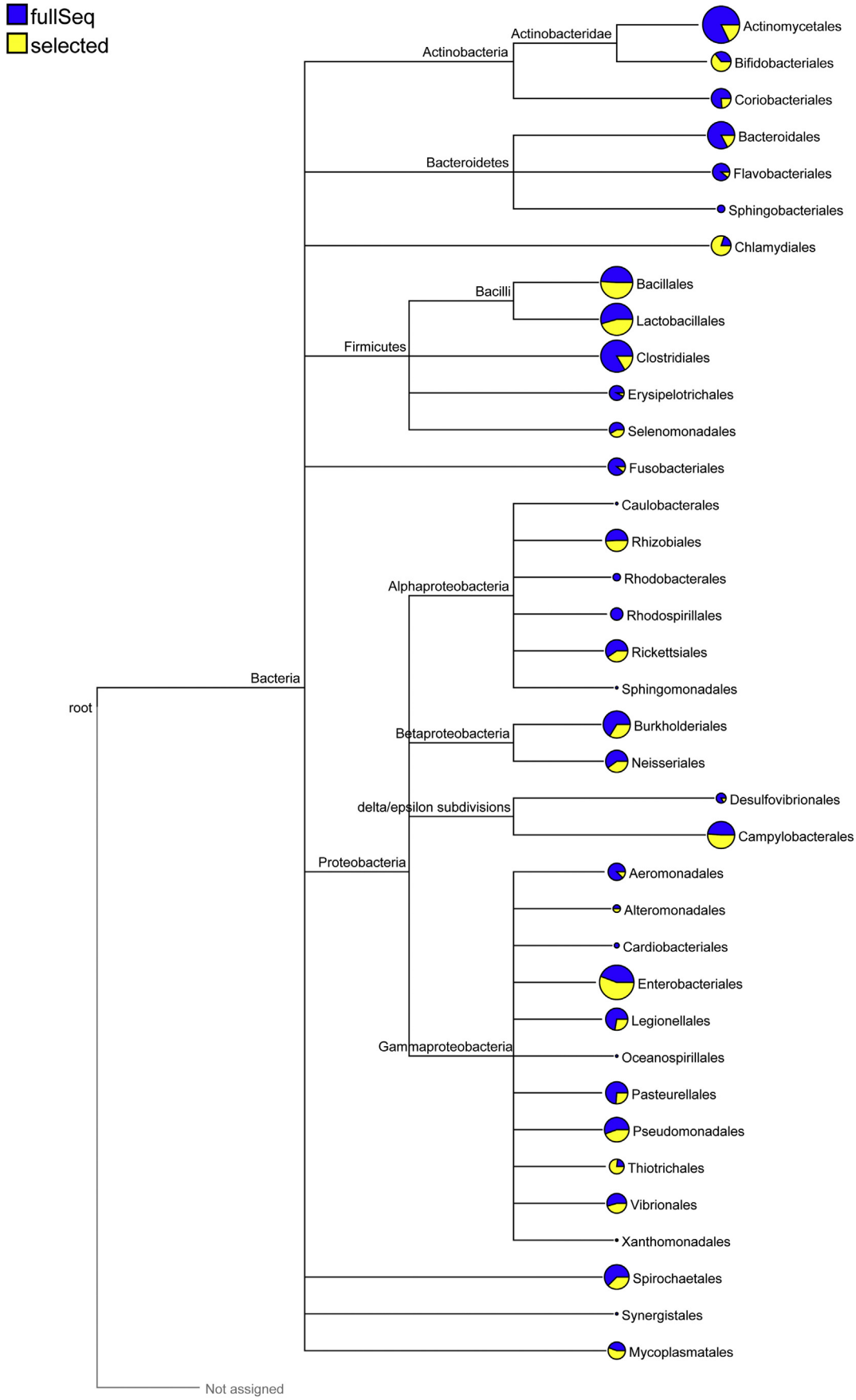

Fig. 1. Classification of 16S rRNA sequences by taxonomy i) 16S rRNA sequences in the 16SpathDB 2.0 database, ii) 16S rRNA sequences extracted from complete genomes of medically important bacteria for the validation of 16SpathDB 2.0. The circles representing sequences classified are scaled according to the number of sequences in each bacterial class. The relative numbers of sequences are represented by the i) blue and ii) yellow sectors. The figure was generated using MEGAN version 4.70.4.