Title: The salivary microbiome for differentiating individuals: proof of principle.

Authors: Leake SL, Pagni M, Falquet L, Taroni F, Greub G

Journal: Microbes and infection / Institut Pasteur

Year: 2016 Jun

Volume: 18

Issue: 6

Pages: 399-405

DOI: 10.1016/j.micinf.2016.03.011

The salivary microbiome for differentiating individuals: proof of principle

Sarah L. Leake[a], Marco Pagni[b], Laurent Falquet[b,c], Franco Taroni[aΔ], Gilbert Greub[dΔ*]

a: School of Criminal Justice, University of Lausanne, Lausanne, Switzerland

b: Swiss Institute of Bioinformatics, Vital-IT group, Lausanne, Switzerland

c: Department of Biology, University of Fribourg, Fribourg, Switzerland

d: Institute of Microbiology, Lausanne, Switzerland

* Corresponding author:

Prof Gillbert Greub
Institute of Microbiology
University of Lausanne
Bugnon 48
1011 Lausanne
041 21 314 49 79
gilbert.greub@chuv.ch

Δ: Both authors contributed equally to this work

Keywords: saliva microbiota; forensic science; Illumina sequencing; human identification

Abstract

Human identification has played a prominent role in forensic science for the past two decades. Identification based on unique genetic traits is driving the field. However, this may have limitations, for instance, for twins. Moreover, high-throughput sequencing techniques are now available and may provide a high amount of data likely useful in forensic science.

This study investigates the potential for bacteria found in the salivary microbiome to be used to differentiate individuals. Two different targets (16S rRNA and *rpoB*) were chosen to maximise coverage of the salivary microbiome and when combined, they increase the power of differentiation (identification). Paired-end Illumina high-throughput sequencing was used to analyse the bacterial composition of saliva from two different people at four different time points (t=0 and t=28 days and then one year later at t=0 and t=28 days). Five major phyla dominate the samples: Firmicutes, Proteobacteria, Actinobacteria, Bacteroidetes and Fusobacteria. *Streptococcus*, a Firmicutes, is one of the most abundant aerobic genera found in saliva and targeting *Streptococcus rpoB* has enabled a deeper characterisation of the different streptococci species, which cannot be differentiated using 16S rRNA alone. We have observed that samples from the same person group together regardless of time of sampling. The results indicate that it is possible to distinguish two people using the bacterial microbiota present in their saliva.

1. Introduction

Current methods of human identification in forensic science rely heavily upon the analysis of human DNA. However, there are limitations to the use of human DNA namely its degradation and low quantity. For example, in sexual assault cases, the DNA from the perpetrator is often masked by the DNA of the victim making identification difficult. In such cases saliva is commonly found due to it being transferred through, amongst others, biting, kissing and licking. To overcome the current unsatisfactory situation, the potential of other targets, for example bacteria, needs to be investigated. Why is bacterial DNA interesting in this context? Firstly, bacterial DNA is better protected than human DNA and more resistant to degradation. Therefore, bacterial DNA will persist better once deposited on a surface. Secondly, it may be possible to distinguish twins using bacterial DNA [1], a feat impossible with current human DNA based methods.

It has been estimated that 99% of bacteria found in the environment cannot be cultured [2]. However, with the arrival of next generation sequencing (NGS) the analysis of bacterial community composition has reached depths previously unachievable. There is now potential to exploit bacteria for forensic purposes. Fierer et al. demonstrated that the analysis of the skin microbiome could be used to link an individual to an object they touched and that the bacterial community found on the object was more similar to the community on the owners hand than to 270 other hands, indicating the potential of this technique for forensic identification [3]. This study extends the idea presented by Fierer et al. by demonstrating the potential of NGS analysis of the salivary microbiota for forensic identification.

A number of studies showing saliva bacterial community composition using NGS have been published [1,4-9]. To date the main gene targeted is 16S rRNA because it is ubiquitous and essential for bacterial life [10,11]. However, there are limitations to targeting 16S rRNA namely, intra-genomic heterogeneity, mosaicism and the lack of a universal threshold sequence identity value [12]. Therefore, in order to have a more complete picture of a microbiome, analysing a second (single-copy) target is essential. In this study the second gene targeted was *rpoB* which, encodes the beta-subunit of RNA polymerase, a very important enzyme that is highly conserved throughout bacteria. It has been shown that like the 16S rRNA gene the *rpoB* gene contains alternating variable and conserved regions [13]. The hypervariable regions of rpoB have shown promise for bacterial identification down to the species and subspecies levels [14-16]. Specifically studies have shown that humans have many different strains of the same \textit{Streptococcus} species, the most prevalent genus in saliva, with many strains being unique to individuals [17,18]. Using 16S rRNA alone these strains would not be detected and therefore an important part of the salivary microbiome would be missed out. By combining *rpoB* with 16S rRNA a deeper level of identification is possible.

Saliva unlike sperm and blood, the other main biological fluids found in criminal cases, is not sterile. Indeed, saliva contains, as many as 500 million bacterial cells per millilitre (ml) and at least 700 different bacterial species [19]. The average composition of the salivary microbiome being known [1,8], we wondered whether there is enough variation to differentiate salivary microbiomes of two different people. To date, studies have shown that differences in salivary microbial communities between individuals are present [5,20], however whether these differences are great enough to differentiate individuals has yet to be explored. Additionally, the salivary microbiome has been shown to be stable over a couple of months [5,8] but no longer, however studies on gut microbiota show stability over a few years [21,22], further work is required to see if this pattern is observed in saliva microbiota. Thus, this study investigates the intra and inter-individual variation of the salivary microbiome of two healthy subjects to investigate the potential of saliva microbiota in forensic science.

2. Materials and Methods

2.1. Sampling and DNA extraction

This study was approved by the Ethics Committee of the Canton of Vaud, Switzerland (protocol 357/11). Saliva samples were obtained from two healthy adult individuals at four time points; t=0 and t=30 days and one year later at t=0 and t=30, with informed consent. Volunteers were asked to brush their teeth in the morning and not eat or drink one hour before sampling. The saliva was collected by spitting into a sterile tube and then stored at -20°C until processing. DNA extraction was performed using the automated MagNA Pure 96 DNA and Viral Nucleic Acid small volume kit (Roche) following the Pathogen Universal 200 v2.0 protocol [23]. Samples were then stored at -20°C.

2.2. PCR and sequencing

In order to maximise coverage of the salivary microbiome, two different targets were chosen; 16S rRNA and *rpoB*. Practically two different pairs of primers targeting *rpoB* were used to investigate the biodiversity of streptococci (*rpoB*1) and other bacteria (*rpoB*2*)*. For 16S rRNA, primers were designed to amplify the V5 region and for *rpoB*, two sets of primers covered the V1 region. Primers were designed using general target species then checked against species known to be found in saliva (see table Table 1 for final primer sequences). Each target was amplified separately in a reaction containing 5 µl of DNA extract, 0.5 µM of both forward and reverse primer, 1x Phusion® HF buffer, 200 µM each dNTP, 0.02U/µl Phusion® Hot Start II DNA polymerase, 3% DMSO and 1mM $MgCl_2$ in a total volume of 50 µl. The following thermal cycling parameters were used: initial denaturation at 98°C for 30 seconds, 35 cycles of denaturation at 98°C for 5 seconds, primer dependant annealing temperature (see Table 1 for annealing temperatures) for 15 seconds and extension at 72°C for 10 seconds with a final extension of 5 minutes at 72°C.

Table 1

All amplified targets from the same sample were pooled together and the pooled sample barcoded. To pool samples equal molar amounts of each sample are necessary, in this case approximately ten picomoles of each were used. The samples were then purified using Agencourt AMPure XP PCR purification (Beckman Coulter). The purified products were then separated on an agarose gel and the band corresponding to the target size (120bp) excised. Finally, the sequencing libraries were prepared using the TruSeq DNA sample preparation kit (Illumina) [24]. Then, 100 cycles of paired-end sequencing were performed on a HiSeq 2000 (Illumina).

2.3. Sequence analysis

Base-calling was performed by HCS 2.0.12/RTA 1.17.21.3 and quality control by the CASAVA 1.8.2 pipeline using standard parameters. Specifically FastQC was used for quality control, by running FastQC in Casava mode the sequences which did not pass the quality threshold were removed [25]. FLASH was used to overlap the paired reads [26]. As each sample contained the sequences for three targets, each target was separated out using barcode splitter (from the FASTX-tool kit [27]) with exact matching for the primer sequence (sequences available in the European Nucleotide Archive under accession number PRJEB6052). This step also removes chimeric sequences.

Sequences were clustered into operational taxonomic units (OTUs) using CD-HIT-EST 4.5.4 [28]. For 16S rRNA 97% identity was used and for rpoB 95%. Any clusters containing less than twenty sequences were removed helping to reduce the number of OTUs resulting from sequencing errors and contamination. Then a representative sequence for each cluster was inputted into BLAST and compared against the entire nucleotide database using the best-hit algorithm to give the 'top' hit. The same process was carried out for both targets to enable direct comparison of results.

In order to compare the taxa abundances between the two experiments the data was normalised using DESeq [29], despite it being designed for RNAseq data, it can also be applied to microbiome data [30]. To minimise the effect of highly abundant taxa the data was then transformed by taking the $\log_{10}(x+1)$ of each count (x). To compare the taxa abundances, the samples from each individual were combined and the mean calculated, producing a mean abundance for each individual per taxon, per target gene. Two statistical inferential approaches have been performed. On one side, from a frequentist perspective, a 2-tailed unpaired t-test was used to compare the means ($\boldsymbol{\theta}$_1 for individual 1 and $\boldsymbol{\theta}$_2 for individual 2, respectively) and then the taxa were ranked by p-values. On the other hand, a Bayesian perspective was adopted by calculating Bayes factors (BF) to test the hypothesis H_0: $\boldsymbol{\theta}$_1 – $\boldsymbol{\theta}$_2 = 0 versus H_1: $\boldsymbol{\theta}$_1 – $\boldsymbol{\theta}$_2 ≠ 0. Due to the small sample size hierarchical clustering using the Ward method was used to group the data and a dendrogram used to visualise the grouping. The R packages hclust and as.dendrogram were used to carry out the clustering analyses. To combine data from different targets taxa considered as significant from each target were inputted into a table and hierarchical analysis performed.

3. Results

3.1. Illumina sequencing results

The saliva microbiome composition of 2 individuals was explored at 4 different time points. The samples were split into two sequencing runs with samples taken one month apart being sequenced together. Therefore, each run contained two samples per individual making 4 samples in total, per run. Run one was performed one year before run two. In total, run one produced 193,221,302 reads. After quality control, pairing and filtering 59,971,947 reads were used for analysis with the following target breakdown: 16S rRNA - 21,534,203, *rpoB*1 - 29,693,058 and *rpoB*2 - 8,744,686. In total, run two produced 201,692,619 reads. After quality filtering and pairing 56,762,234 reads were used for analysis with the following target breakdown: 16S rRNA - 30,604,336, *rpoB*1 - 17,007,924 and *rpoB*2 - 9,149,974. A breakdown of the number of different OTUs found per sample, per target can be found in Table 2.

3.2. Microbiome composition

The use of three targets enables the microbiome composition to be analysed to a greater depth. Fig.1 shows the proportion of the top five phyla per individual and per target. For both *rpoB*1 and 16S rRNA, Firmicutes is the most common phyla constituting over 90% and 70% of the population respectively. For rpoB2 the population is composed of over 90% Actinobacteria. The large difference in taxa found by each *rpoB* primer pair is expected as they were designed to amplify different taxa, demonstrating the benefit of targeting more than one region of the same target gene.

Fig.1

The addition of rpoB enables certain genera to be analysed down to the species and even strain level. Specifically, with 16S rRNA *Streptococcus* can be detected at the genus level and occasionally the species level (9 different OTUs); however, with rpoB it can be detected to the species/strain level (53 different OTUs) enabling a deeper characterisation of this part of the saliva microbiome. This is important as *Streptococcus* makes up about 80% of Firmicutes, the most abundant phylum.

3.3. Minimum sequences required

This study used the HiSeq2000 to analyse the samples, a machine which can produce over one billion reads, as at the outset of this study the number of sequences required to separate two individuals was unknown. To calculate the minimum number of sequences necessary the data were randomly sub-sampled at different levels: 1000, 10000, 50000, 100000, 500000 and 1000000 sequences. The analysis was performed to the end and the relative distances calculated between the samples at all levels are shown in Fig.2 For *rpoB*2 that provides the smallest separation, at least 50000 sequences were required to adequately discriminate the two investigated individuals. 16S rRNA provides the best separation when looking at the targets individually. However, when 16S rRNA and *rpoB*1 are combined the separation is improved. Combining all three targets produces the best separation, however the addition of *rpoB*2 does not greatly improve the separation except at 50000 sequences where the separation is significantly improved.

Fig.2

3.4. Clustering threshold

Unlike previous studies the main aim of this study was to investigate whether the bacteria found in saliva could be used to separate samples from different individuals and not just characterise the microbiome. Different clustering thresholds were tested to see which one gave the best separation taking into account analysis time i.e. the total time required to analyse the data after sequencing. Fig.3 shows that as the percent identity, generally, increases so does the relative distance between the two individuals. The results for both *rpoB* targets are shown in Fig.3A where the dashed line indicates the chosen threshold of 95%. In Fig.3B the dashed line highlights the chosen threshold for 16S rRNA of 97%. These percentages correspond to previously published studies for species level characterisation for *rpoB* and 16S rRNA, respectively [10,31]. For both targets 100% identity provides the best separation however the analysis time, for 16S rRNA especially, is very long and therefore it is not the most efficient solution.

Fig.3

3.5. Hierarchical clustering

Firstly the normalised logged data was filtered by performing a 2-tailed unpaired t-test and ranking the taxa by p-value and only the taxa with a p-value < 0.1 (and a BF <1) were kept for analysis. The data was further filtered by removing any taxa that did not appear in both experiments. Hierarchical clustering was performed by first calculating the Euclidean distance and then using the Ward method to produce relative distances between each sample. Fig.4

shows the dendrograms representing the relative distances between the samples, for each target, (A-C) and then for all targets combined (D). For all targets, samples from different individuals are separated, due to a significant inter-individual variation. Concerning the intra-individual variation samples sequenced in the same run are expected to be more similar and therefore logically grouped together as seen in Fig.4B and D. Conversely, the intra-individual separation for *rpoB*1 (Fig.4A) and 16S rRNA (Fig.4C) is not ideal. However, when all three targets are combined good inter and intra-individual separation could also be achieved, demonstrating the benefit of analysing more than one target gene.

Fig.4

4. Discussion

This paper presented the first study into the use of the salivary microbiome for human identification. It has shown that the salivary microbiome exhibits a significant biodiversity and by using a PCR-based metagenomic approach the discrimination of two unrelated individuals was possible. The biodiversity revealed in all samples was similar to that found by previous studies, showing that the designed primers are robust. However, the abundances do differ but this has been observed previously [1].

Previous studies [1,6,8] have shown that the most common phlya found in saliva are: Firmicutes, Proteobacteria, Actinobacteria, Bacteroidetes and Fusobacteria and this study concurs with these findings; however the abundances differ slightly. Stahringer et al. analysed 264 saliva samples and showed that bacterial abundances varied greatly, this study falls within the observed variation. In the same study they defined a genus-level core microbiome containing eight genera [1]. By combining three targets in this study a genus-level core microbiome of 58 genera was observed. This high number of genera covers about 95% of the population of each individual implying that most differences come from the species/strain level. However, this study is limited by a small sample size and more samples may reveal the core microbiome to be similar to previous studies. Such a small sample size was chosen, as the depth of sequencing required to differentiate two individuals was unknown. Therefore, this was one of the major goals of the research. Had too many samples been analysed in one run, the minimum number of sequences required may not have been achieved, so we remained conservative with regards to sample size.

The results showed that the minimum number of sequences for this type of analysis is 100,000 as this provided a good separation between individuals with all targets. However, the addition of *rpoB*2 did not significantly increase the discrimination. One of the main advantages of *rpoB* is that it identifies a fair number of species/strains and both primer pairs identify different species. However, *rpoB*2 identifies much less than *rpoB*1. Even though the best separation is achieved with sequences of all three target genes, very good separation is still achieved when combining only 16S rRNA and *rpoB*1. Therefore, the choice of target combination would depend on how many samples were to be sequenced in one run. By only using two target genes, more samples could be sequenced, making the technique more economical whilst achieving rather similar results. By choosing a clustering threshold, which enables identification down to the species/strain level whilst remaining time efficient, the whole analysis could be carried out in about one week, depending on which high-throughput sequencer is used.

To perform the hierarchical clustering the data was filtered to only use the taxa found to be significant by a 2-tailed unpaired t-test (and a BF < 1 meaning a support, generally with values that very strongly support the hypothesis H_1). Due to the number of OTUs found obviously not all of them are useful for separating samples from different individuals. To reduce analysis complexity, only OTUs found in both sequencing runs were kept as they could be more accurately attributed to an individual and techniques used in forensic science are required to be as robust as possible. Inevitably there is some natural variation in saliva microbiota due to it being a dynamic fluid and certain bacteria will not always be detected, being either absent or in too few numbers. To ensure that no sequencing errors were included, any clusters containing less than twenty sequences were removed prior to analysis. Even with this highly conservative algorithm, samples from one individual can be successfully separated from those of a second individual (Fig.4) whilst minimising the intra-individual variation. Altogether, our technique proved to be highly robust and is innovative not only for its putative application in forensic science, but also by using a combination of a highly discriminative gene (*rpoB*) with the 16S rRNA target generally used for PCR-based metagenomics. However, the present work only represents a first proof of principle and we need to study twins in order to confirm that saliva microbiota may indeed differentiate twins.
A recent study by Stahringer et al. showed that for twins aged between 12-24 years their salivary microbiome was not statistically more similar than for any other pair [1]. This indicates that overall there is very little or no genetic influence on salivary microbiome composition and that the differences observed between twins mainly come from environmental factors. Indeed a number of environmental factors such as diet, oral hygiene, smoking, alcohol and drug consumption may influence the salivary microbiome [1]. Therefore a person's microbiome could be used as intelligence to inform about their lifestyle.

One major environmental factor is antibiotics. Lazarevic et al. described the effects of amoxicillin treatment on the salivary microbiota in children with acute otitis media. They showed that directly after treatment there was a change in the microbiota in terms of both species richness and diversity [32]. However, three weeks after the end of treatment the microbiota had mainly recovered back to pre-antibiotic diversity. This, would only impact cases where the saliva was deposited on a crime scene whilst the perpetrator was taking antibiotics. In such cases, presence of antibiotics in the sample might be determined and an additional sample might then be obtained upon treatment with the same antimicrobial substance. In the case where the perpetrator is taking antibiotics when apprehended a reference sample could be taken at a later date once the salivary microbiome had recovered.

Another important point to consider with regards to forensic traces is how resistant the traces (i.e. here the bacterial DNA) are to external factors. Indeed, UV light, heat and humidity can degrade human DNA, environmental conditions which are often found at crime scenes. One advantage of mircobiota based forensic investigation is that bacterial DNA is better protected from degradation than human DNA as bacterial DNA is circular often highly condensed as "nucleoid" and therefore harder to be degraded by enzymes. Moreover, prokaryotic cells have a cell wall, which is chemically complex with a peptidoglycan matrix that better protects the contents of the cell compared to the cell membrane of eukaryotic cells. Therefore bacterial DNA should be more resistant than eukaryotic DNA to external factors taking longer to be degraded.

The goal of this technique is not to replace current methods used for human identification but to be complementary. When these methods do not produce satisfactory results there is no other option from a biological identification standpoint. By analysing the salivary microbiome, new options become available that previously were not possible. There are two main applications of this technique in forensic science: human identification and intelligence. The first will only be possible if a reference sample is available. The second application uses the same data but looks at the presence of specific bacteria, which could indicate a certain lifestyle. This information might be used to help guide an investigation. If an identification is not possible then the data acquired could still provide valuable information to a case. However, much more work is needed to relate given species to given lifestyle habits.

In conclusion, Illumina high-throughput sequencing of the salivary microbiome can be used to identify saliva samples from two different individuals. This technique shows promise for human identification, specifically for twins and other cases where standard DNA typing does not provide satisfactory results due to degradation of human DNA. The results could also be used for intelligence purposes by providing information concerning a person's lifestyle. Further work is required to investigate the benefit and limitations of this technique.

Acknowledgments

The authors would like to thank Sebastien Aeby (Institute of Microbiology) for his technical support as well as the Lausanne Genomic Technologies Facility for carrying out the sequencing and Vital-IT (Swiss Institute for Bioinformatics) for computing power.

References

[1] Stahringer SS, Clemente JC, Corley RP, Hewitt J, Knights D, Walters WA, et al. Nurture trumps nature in a longitudinal survey of salivary bacterial communities in twins from early adolescence to early adulthood. Genome Research 2012;22:2146–52.

[2] Handelsman J. Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol Mol Biol Rev 2004;68:669–85.

[3] Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Forensic identification using skin bacterial communities. P Natl Acad Sci Usa 2010;107:6477–81.

[4] Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. Defining the Normal Bacterial Flora of the Oral Cavity. Journal of Clinical Microbiology 2005;43:5721–32.

[5] Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial Community Variation in Human Body Habitats Across Space and Time. Science 2009;326:1694–7.

[6] Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Østerås M, et al. Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. Journal of Microbiological Methods 2009;79:266–71.

[7] Zaura E, Keijser BJF, Huse SM, Crielaard W. Defining the healthy core microbiome of oral microbial communities. BMC Microbiology 2009;9.

[8] Lazarevic V, Whiteson K, Hernandez D, Francois P, Schrenzel J. Study of inter- and intra-individual variations in the salivary microbiota. BMC Genomics 2010;11:523.

[9] Caporaso JG, Lauber C, Costello E, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. Genome Biology 2011;12:R50.

[10] Case RJ, Boucher Y, Dahllof I, Holmstrom C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. Applied and Environmental Microbiology 2007;73:278–88.

[11] Weisburg WG, Barns SM, Pelletier DA, Lane DJ. 16S ribosomal DNA amplification for phylogenetic study. Journal of Bacteriology 1991;173:697–703.

[12] Rajendhran J, Gunasekaran P. Microbial phylogeny and diversity: Small subunit ribosomal RNA sequence analysis and beyond. Microbiological Research 2011;166:99–110.

[13] Boor KJ, Duncan ML, Price CW. Genetic and Transcriptional Organization of the Region Encoding the Subunit of Bacillus subtilis RNA Polymerase. Journal of Biological Chemistry 1995;270:20329–36.

[14] Mollet C, Drancourt M, Raoult D. rpoB sequence analysis as a novel basis for bacterial identification. Molecular Microbiology 1997;26:1005–11.

[15] Adekambi T, Colson P, Drancourt M. rpoB-Based Identification of Nonpigmented and Late-Pigmenting Rapidly Growing Mycobacteria. Journal of Clinical Microbiology 2003;41:5699–708.

[16] Scola BL, Bui LTM, Baranton G, Khamis A, Raoult D. Partial rpoB gene sequencing for identification of Leptospira species. FEMS Microbiology Letters 2006;263:142–7.

[17] Rudney JD, Larson CJ. Use of restriction fragment polymorphism analysis of rRNA genes to assign species to unknown clinical isolates of oral viridans streptococci. Journal of Clinical Microbiology 1994;32:437–43.

[18] Wisplinghoff H, Reinert RR, Cornely O, Seifert H. Molecular Relationships and Antimicrobial Susceptibilities of Viridans Group Streptococci Isolated from Blood of Neutropenic Cancer Patients. Journal of Clinical Microbiology 1999;37:1876–80.

[19] Paster BJ, Olsen I, Aas JA, Dewhirst FE. The breadth of bacterial diversity in the human periodontal pocket and other oral sites. Periodontology 2000 2006;42:80–7.

[20] Consortium THMP. Structure, function and diversity of the healthy human microbiome. Nature 2012;486:207–14.

[21] Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The Long-Term Stability of the Human Gut Microbiota. Science 2013;341.

[22] Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature 2013;493:45–50.

[23] Roche Diagnostics GmbH. MagNA Pure 96 DNA and Viral NA Small Volume Kit. Version 08 2012.

[24] Illumina, Inc. TruSeq DNA Sample Prep Kits. Illumina; 2012.

[25] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. BioinformaticsBabrahamAcUk n.d.

[26] Magofç T, Salzberg SL. FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies. Bioinformatics 2011.

[27] http://hannonlab.cshl.edu/fastx_toolkit/index.html. HannonlabCshlEdu n.d.

[28] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–9.

[29] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology 2010;11:R106.

[30] McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol 2014;10:e1003531.

[31] Drancourt M, Raoult D. rpoB Gene Sequence-Based Identification of Staphylococcus Species. Journal of Clinical Microbiology 2002;40:1333–8.

[32] Lazarevic V, Manzano S, Gaïa N, Girard M, Whiteson K, Hibbs J, et al. Effects of amoxicillin treatment on the salivary microbiota in children with acute otitis media. Clinical Microbiology and Infection 2013;19:E335–42.

Figure legends

Fig.1. Relative abundance of the top five phyla, per individual, per target gene. A and B are different individuals and the target genes are shown in brackets.

Fig.2. Number of sequences required for sample separation. The relative distance corresponds to the distance between two individuals calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species with a p-value < 0.1 from a t-test between the samples from each individual or a BF < 1 were used.

Fig.3. Comparison of clustering thresholds for the separation of individuals. The percent identity is that used for clustering the sequences into OTUs with CD-HIT. The relative distance corresponds to the distance between two individuals calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species with a p-value < 0.1 from a t-test between the samples from each individual or a BF < 1 were used. A = both rpoB targets and B = 16S rRNA. The dashed line highlights the chosen threshold.

Fig.4. Hierarchical clustering of all eight samples for each target. The relative distance corresponds to the distance between two individuals (A and B) calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species occurring in both experiments and with a p-value < 0.1 from a t-test between the samples from each individual or a BF < 1 were used.

Table 1. Primers designed for each gene target. Primer name for 16S rRNA and *rpoB*2 corresponds to the *Escherichia coli* positions and for *rpoB*1 to the *Streptococcus bovis* positions.

Gene	Primer name	Primer sequence (5'-3')	Tm (℃)
16S rRNA	792 F	AGGATTAGATACCCTGGTAG	56
	891R	CGTACTCCCCAGGCGG	
*rpoB*1	130F	GGACCTGGTGGTTTGAC	64
	220R	CGATGTTAGGTCCTTCAGG	
rpoB2	340F	GGACCAGAACAACCCG	60
	434R	GGGTGTCCGTCTCGAAC	

Table 2. Species-level OTUs for all samples, per target.

Sample	No. OTUs 16S rRNA	No. OTUs *rpoB1*	No. OTUs *rpoB2*
Experiment 1			
A1	810	145	20
A2	793	147	23
B1	839	149	25
B2	828	144	29
Experiment 2			
A3	1273	182	46
A4	1267	185	44
B3	1291	169	44
B4	1283	171	48

Figure 1

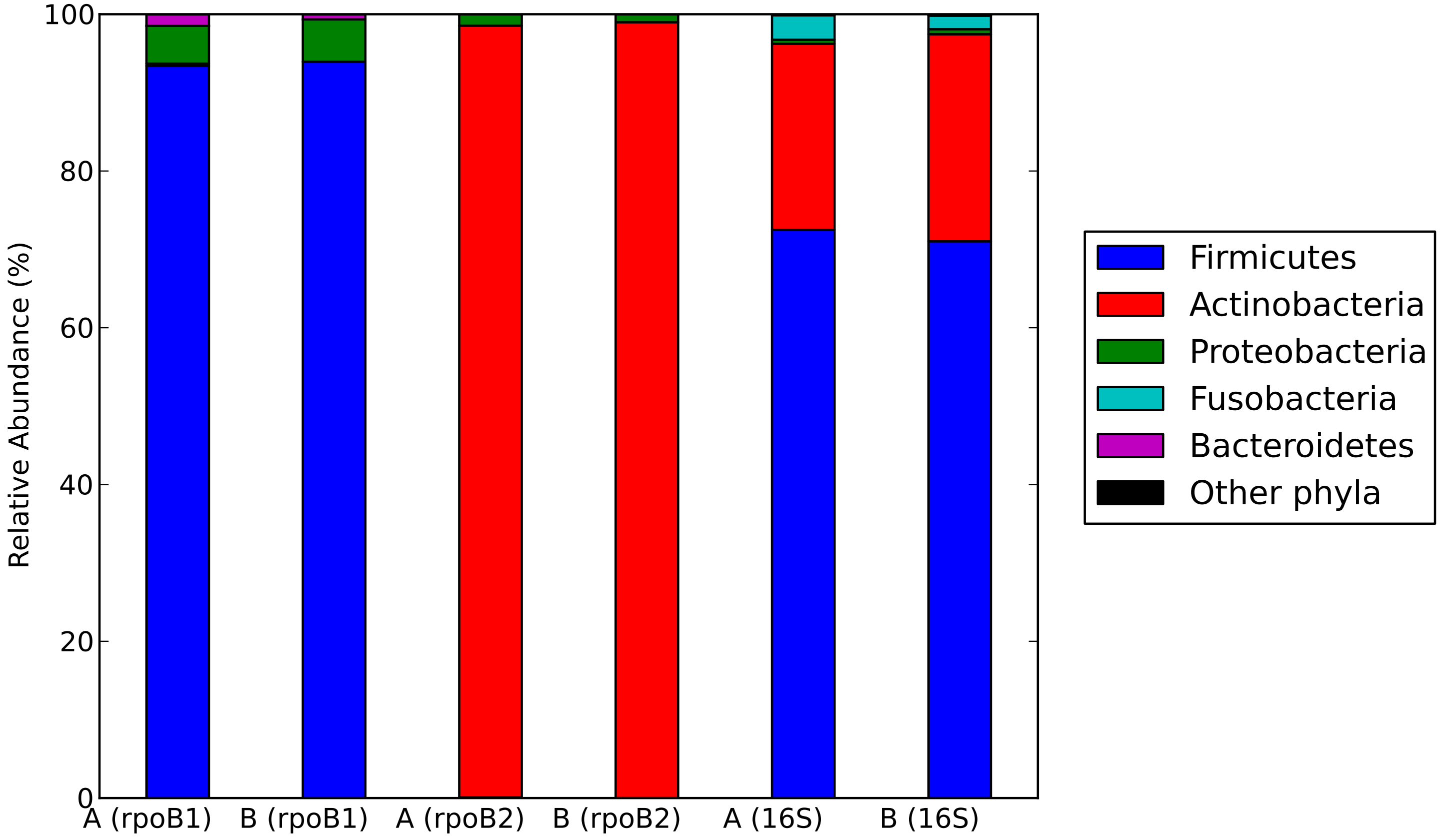

100
80
60
40
20
0
Relative Abundance (%)
A (rpoB1)
B (rpoB1)
A (rpoB2)
B (rpoB2)
A (16S)
B (16S)
Firmicutes
Actinobacteria
Proteobacteria
Fusobacteria
Bacteroidetes
Other phyla

Figure 2

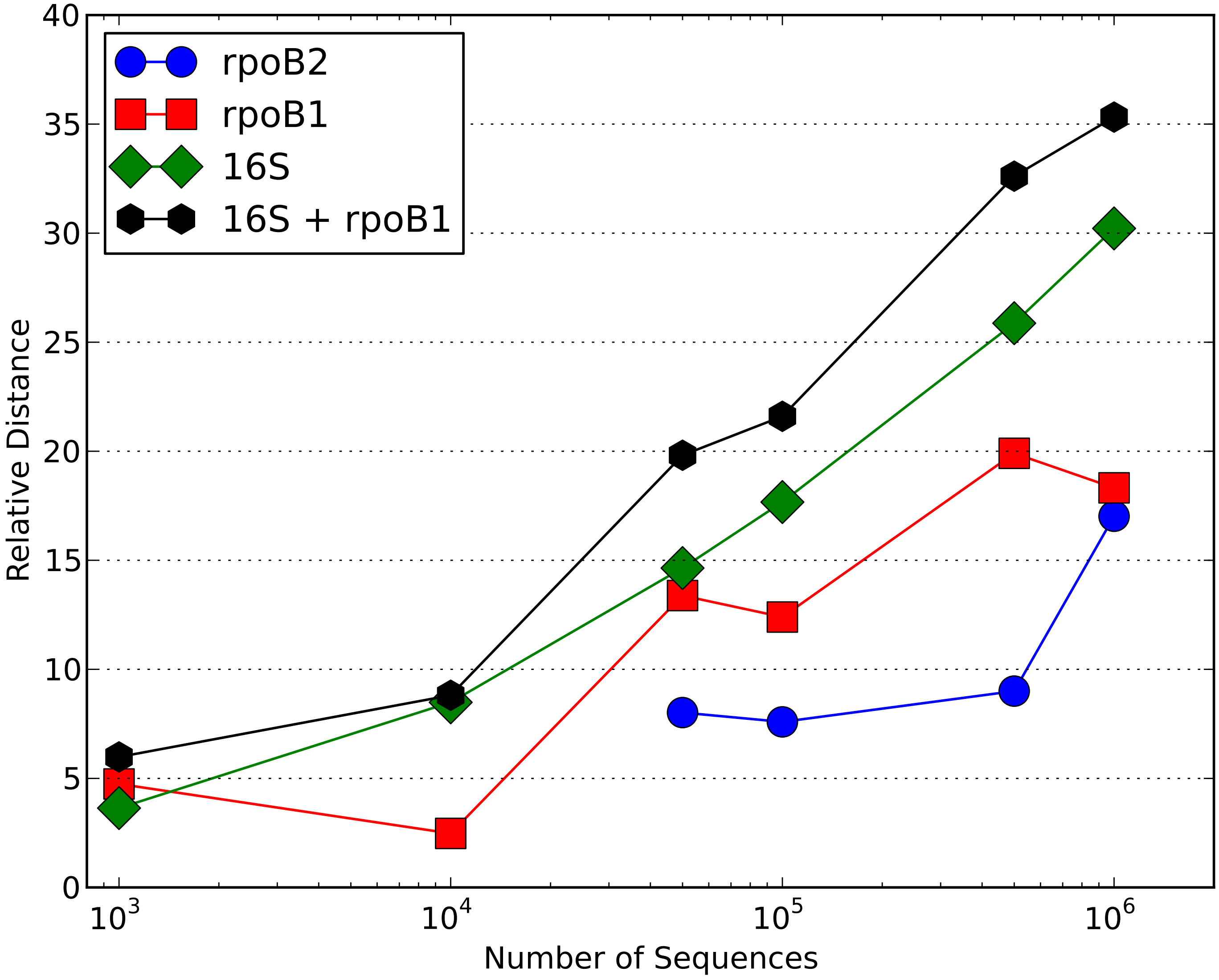

rpoB2
rpoB1
16S
16S + rpoB1
Relative Distance
Number of Sequences
0
5
10
15
20
25
30
35
40
10^3
10^4
10^5
10^6

Figure 3

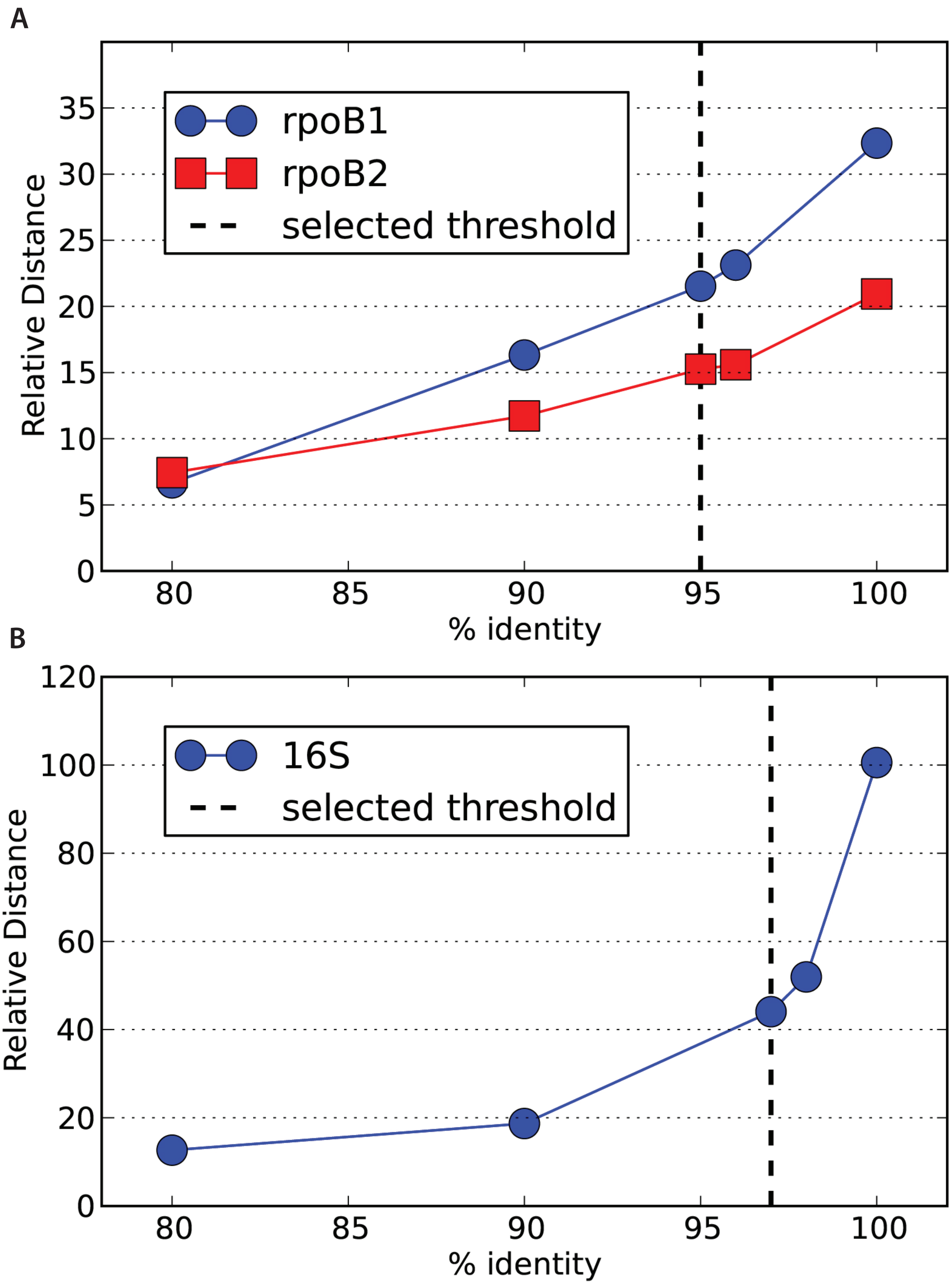
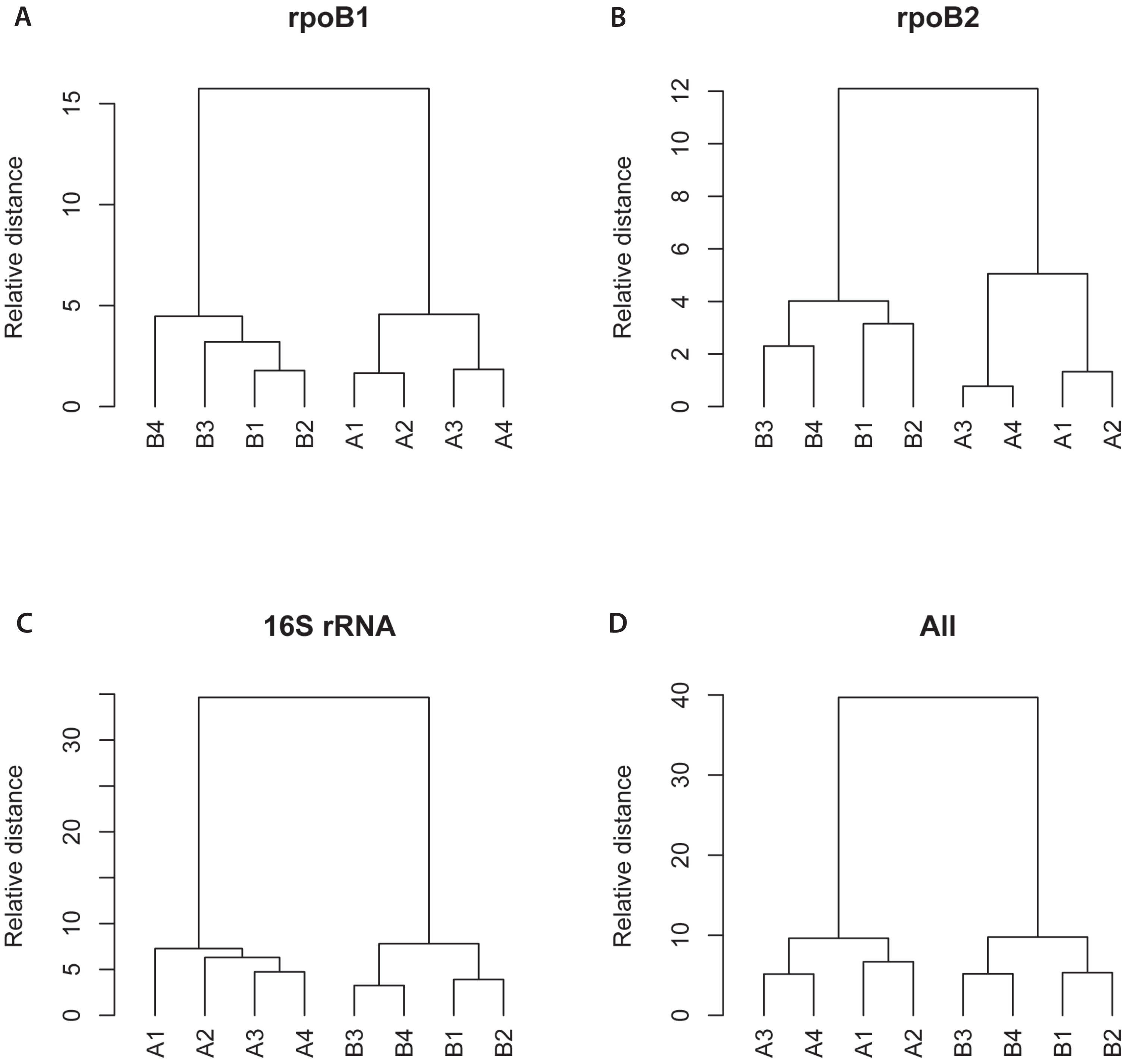

A
rpoB1
rpoB2
selected threshold
Relative Distance
0
5
10
15
20
25
30
35
80
85
90
95
100
% identity
B
16S
selected threshold
Relative Distance
0
20
40
60
80
100
120
80
85
90
95
100
% identity

Figure 4


A
rpoB1
Relative distance
0
5
10
15
B4
B3
B1
B2
A1
A2
A3
A4
B
rpoB2
Relative distance
0
2
4
6
8
10
12
B3
B4
B1
B2
A3
A4
A1
A2
C
16S rRNA
Relative distance
0
5
10
20
30
A1
A2
A3
A4
B3
B4
B1
B2
D
All
Relative distance
0
10
20
30
40
A3
A4
A1
A2
B3
B4
B1
B2