Comparison of the variation of gut bacterial diversity marker genes in public genomes: a case study of the class Clostridia

Maria Cadavid[1,2,*], **Juan S. Escobar**[2,‡] **and Laura Salazar-Jaramillo**[2,‡]

[1]Departamento de Ciencias Biológicas, Escuela de Ciencias, Universidad EAFIT, Medellin, Antioquia, Colombia, [2]Vidarium—Nutrition, Health and Wellness Research Center, Grupo Empresarial Nutresa, Medellin, Colombia, *Graduating biology student - mcadav29@eafit.edu.co, ‡Advisers

ABSTRACT Characterization of microbial communities is commonly performed using the *16S ribosomal RNA* marker gene. However, recent advances in sequencing technologies have exposed several limitations of this marker to distinguish prokaryotes at finer taxonomic levels. It has been proposed that the use of alternative, more variable marker genes counteracts *16S rRNA* disadvantages, while maintaining a cost-effective and highly informative methodology. We evaluated the suitability of *DNA kinase1* (*dnak1*) and *gyrase subunit B* (*gyrB*) genes as genetic markers for the study of community diversity of the class Clostridia. We first built a reference dataset by downloading genome assemblies classified under the class Clostridia by the GTDB database and extracted from the assemblies the sequences of the genes of interest (*dnak1*, *gyrB* and *16S rRNA*). We then estimated the variability of each gene and compared it to that of the *16S ribosomal RNA* using pairwise identity matrices. A linear regression between pairwise comparison of marker genes belonging to the same assembly allowed us to understand how much resolution is gained by using an alternative marker compared to *16S rRNA sequences*. Our results showed that both *dnaK1* and *gyrB* are more variable in their sequence than the *16S rRNA* gene and, therefore, may constitute appropriate markers to assess genetic diversity within Clostridia. Other studies evaluating different genes, bacterial groups and ecosystems, consistently confirm that the use of alternative house-keeping genes can help to elucidate finer diversity than what is observed with the traditional *16S rRNA* gene. We foresee the use of alternative, house-keeping genes as being complementary to the *16S rRNA* approach to answer questions that span different taxonomic levels and address bacterial identification, classification, phylogenetics and evolution.

KEYWORDS Marker gene; *16S ribosomal RNA (16S rRNA)*; *DNA kinase1 (dnaK1)*; *gyrase subunit B (gyrB)*, Clostridia

Introduction

An organism's genome sequence is the ultimate record of its evolutionary history (Zuckerkandl & Pauling, 1965). Therefore, the implementation of comparative analysis of molecular sequences to determine evolutionary relationships was a milestone for systematics, taxonomy and evolution. Yet, not all molecular sequences or sites in a genome evolve at the same rate: while some sites are under strong selection pressure to remain unchanged[1] and are thus conserved, other sites are variable, either because they are neutral (i.e., substitutions are approximately the same as the nucleotide mutation rate) or under selection for change[2] (i.e., higher substitutions than the nucleotide mutation rate) (Wagner, 2002). Conserved sites allow for detection of relatedness between distant species, while more variable sites serve to explain closer taxonomic relationships.

Initial work using molecular phylogenetics to understand relationships spanning the entire tree of life started in the 1970s with Carl Woese, who was in search of molecules that were common to all organisms (Woese & Fox, 1977). Given that ribosomal RNA (rRNA) performs one of the most fundamental processes inside the cell (it is key for turning DNA into proteins)

[1] Purifying selection: prevents the change of an amino acid residue at a given position because it has deleterious effects, thus favoring an excess of synonymous versus nonsynonymous substitutions (Wagner, 2002).

[2] Positive selection: more nonsynonymous substitutions than synonymous substitutions have been preserved, indicating an abundance of amino acid residue change that confers a selective advantage (Wagner, 2002).

it is present in all replicating organisms and its DNA sequence changes relatively slowly over time (Fox, Pechman, & Woese, 1977; C. R. Woese & Fox, 1977). Interestingly, the rRNA gene contains regions of both extreme conservation (i.e., regions under strong purifying selection) and hypervariability (neutral or nearly neutral) allowing to study taxonomic relationships at different levels (Fox *et al.*, 1977) and, thus, making it an excellent candidate marker to survey relatedness among organisms.

Characterization of microbial communities has thus been done traditionally using ribosomal markers, with the *16S ribosomal RNA (16S rRNA)* gene as the gold standard marker (Case *et al.*, 2007; C. R. Woese, Kandler, & Wheelis, 1990). The *16S rRNA* gene is ubiquitous in the prokaryotic domains which, together with the technical advantages and historical use, makes this marker the most widely used for taxonomical classification and separation of prokaryotes. The basic approach takes advantage of the conserved regions to design universal PCR primers to amplify sequence fragments containing the more variable regions, yet, conserved enough to be present across different taxa and allowing their discrimination (Head, Saunders, & Pickup, 1998; Carl R. Woese *et al.*, 1975). The amplified fragments are compared in terms of their nucleotide substitutions and grouped into clusters of similar sequences. Since proposed in 1994, a threshold of 97% similarity in *16S rRNA* sequences, groups them as belonging to the same species (Stackebrandt & Goebel, 1994). The historical use of the *16S rRNA* marker makes it one of the most reported and sequenced genes with extensive available databases (e.g., ribosomal RNA project: RDP, Silva, Greengenes).

Despite the extraordinary insights that have been gained through *16S rRNA* analyses to compare distant lineages, recent advances in sequencing technologies have exposed several limitations with this methodology to distinguish prokaryotes at finer taxonomic levels, such as species and strains. The reason is that for closely related bacteria, the *16S rRNA* gene is too conserved, meaning few nucleotide substitutions (Satoshi Yamamoto & Harayama, 1996). It also has the disadvantage of often being multi-copy within a single genome (Crosby & Criddle, 2003). The intragenomic copies can differ in sequence, leading to identification of multiple ribotypes for a single organism. This intragenomic heterogeneity and the variable number of copies can cause erroneous diversity and abundance estimates in a complex sample (Pei *et al.*, 2010; Sun, Jiang, Wu, & Zhou, 2013; Větrovský & Baldrian, 2013).

Other methods exist to counteract *16S rRNA* limitations. Methods such as DNA-DNA hybridization, biochemical assays and FISH (Franco-Duarte *et al.*, 2019) can be used for prokaryotic identification, however they need a pre-existing knowledge of the expected diversity and, thus, have a bias to cultured species. Considerably, many organisms are hard to culture, for example strict anaerobes, therefore, depending on the organisms of interest, these methods are not always practical. On the other hand, culture-independent sequence analysis has the advantage of not being ad hoc and, thus, retrieving unexpected diversity. An advanced sequence analysis method is shotgun metagenomic sequencing where, instead of targeting a single gene, the total genomic DNA of the microorganisms in the surveyed community is sequenced and reconstructed genome fragments are assigned to draft genomes (Sharpton, 2014). However, this approach is not feasible in many cases since the generation of metagenome data is expensive, massive and complex, especially if there is no prior knowledge about the surveyed microbial community.

Consequently, it has been suggested that the way to counterbalance *16S rRNA* disadvantages, especially for groups of closely related species, while maintaining a cost-effective and highly informative methodology, is the use of alternative, more variable, marker genes that could allow a better resolution for phylogenetic relationships of microbial communities (Nguyen, Warnow, Pop, & White, 2002; Olm *et al.*, 2020; Santos & Ochman, 2004). Sequencing a single gene implies optimizing the limited resources in fewer base pairs, as opposed to the complete genome, which leads to a greater sequencing depth per organism, and thus more informative sequences (for example to identify rarer taxa). The alternative markers are usually single-copy, house-keeping protein-coding genes (Case *et al.*, 2007; Olm *et al.*, 2020) and their efficiency can depend of the taxonomic group of interest. While such genes are under strong purifying selection, they have sites (e.g., third codon positions) that are effectively neutral and accumulate more variation than, for example, the *16S rRNA* gene, thus serving for systematic purposes.

Several studies using approaches with alternative marker genes have been done in the last years. For example, Caro-Quintero & Ochman (2015) demonstrated the utility of these markers (in their case, the gene *gyrB*) to reveal hidden diversity in Bacteroidaceae and Lachnospiraceae families. Moeller *et al.* (2016) used also the alternative marker *gyrB* to demonstrate co-speciation between Bacteroidaceae and Bifidobacteriaceae with hominids across hundreds of thousands of host generations. Guo, Cole, Brown, & Tiedje (2019) compared the variation of two single copy ribosomal protein genes, rplB and rpsC, with the *16S rRNA* for completed bacterial genomes in NCBI RefSeq. The latter demonstrated that both rplB and rpsC showed more variation than did the *16S rRNA* in the same organisms and, hence, helped to better reflect the ecology of microbial communities. These studies illustrate that the use of alternative marker genes, different to *16S rRNA*, are useful to achieve a better understanding of the diversity and taxonomic relationships between microorganisms.

The implementation of strategies with alternative marker genes can be especially useful in the context of the human gut microbiota. A characterization of the Colombian gut microbial community showed that it is particularly highly enriched in two closely related families of the class Clostridia, namely Ruminococcaceae and Lachnospiraceae (de la Cuesta-Zuluaga *et al.*, 2018). These families have members difficult to distinguish based on *16S rRNA* sequences alone, such as the polyphyletic genus Ruminococcus (exclusively associated with animal hosts) (La Reau, Meier-Kolthoff, & Suen, 2016). In addition, other taxa within these families (e.g., Faecalibacterium prausnitzii, Oscillospira) have high diversities that are not adequately assessed by *16S rRNA* sequences (De Filippis, Pasolli, & Ercolini, 2020; de la Cuesta-Zuluaga *et al.*, 2018). Importantly, it has been reported that humans whose microbiota is dominated by Ruminococcaceae members have lower risks of obesity and cardiometabolic disease, while the associations between Lachnospiraceae and health are opposite (de la Cuesta-Zuluaga *et al.*, 2018). Thus, to better understand these correlations it is of paramount importance to accurately classify sequences and assess biological diversity within these families.

Here, we carry out a case study to evaluate alternative marker genes within the class Clostridia (Bacteria: Firmicutes). This is a diverse group of strictly anaerobic to aerotolerant spore forming bacilli, including mostly gram positive and some gram negative species, abundant in the gut microbiota of humans

and non-human animals as well as free living in soil (Wells & Wilkins, 1996). Due to their anaerobic nature, this group is difficult to culture and thus to investigate. For this reason, culture-independent molecular approaches designed to study Clostridia are particularly practical and informative.

In this study, we aim to evaluate the suitability of DNA sequences from two house-keeping genes, DNA kinase1 (*dnaK1*) and gyrase subunit B (*gyrB*), as genetic markers for the study of community diversity of the class Clostridia. The genes *dnaK1* and *gyrB* codify for proteins of approximate 594 and 680 amino acids, respectively. The *dnaK1* gene codes for a molecular chaperone that refolds miss-folded proteins and helps them to reach their native conformation (Aguilar-Rodríguez *et al.*, 2016). The *gyrB* gene codes for the subunit B of the protein DNA gyrase, a type II topoisomerase, which relax or supercoil closed circular double stranded DNA in an ATP-dependent manner (Watt & Hickson, 1994). Both have been previously used as taxonomic markers (Caro-Quintero & Ochman, 2015; Olm *et al.*, 2020; Santos & Ochman, 2004; Wang, Lee, Tai, & Kasai, 2007; S. Yamamoto & Harayama, 1995; Satoshi Yamamoto & Harayama, 1996) and offer various potential advantages over standard *16S rRNA* based approaches. Both are expected to be single-copy within the genomes (Albertsen *et al.*, 2013) and are hypothesized to have a faster evolutionary rate than the *16S rRNA*, thus, they are less conserved and can resolve evolutionary relationships between closely related taxonomic groups.

Our hypothesis is that the two genes, *dnaK1* and *gyrB*, are consistently less conserved than the *16S rRNA* in their sequences within Clostridia and may therefore constitute better markers for assessing the group's real diversity at finer taxonomic levels. To test this, we characterized the genetic variability of the *dnaK1* and *gyrB* genes present in 4073 Clostridia genome assemblies available in the public Genome Taxonomy Database (GTDB) and compared it to that of the *16S rRNA* marker. In this way, we aimed to determine the suitability of using *dnaK1* and *gyrB* as genetic markers for the study of communities of the class Clostridia in the human gut microbiota.

Methods

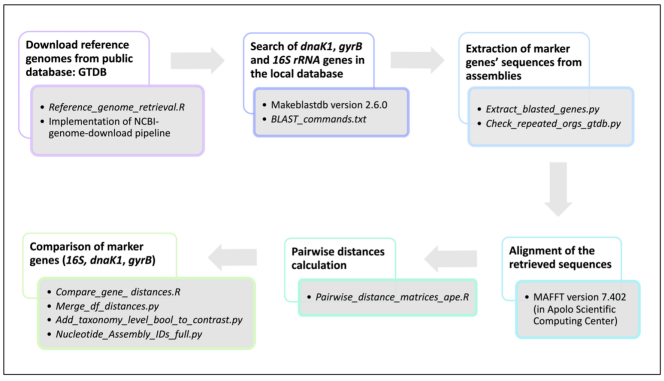

Figure 1 General methods workflow: steps (white boxes) and detail of written scripts and used programs for each purpose (gray boxes). All scripts are available in the GitHub platform following the link https://github.com/mariacadavid/Clostridia_genetic_markers

Download reference genomes from public database: GTDB

In order to build a local database of reference genomes of the class Clostridia, a search of genome assemblies was performed through the GTDB (https://gtdb.ecogenomic.org). This public database is an initiative of the Australian Research Council Laurate Fellowship to establish a standardized microbial taxonomy based on genome phylogeny (Parks *et al.*, 2020). GTDB uses genome sequences published in Genbank and RefSeq to construct a revised phylogeny. To determine our set of reference genome assemblies, we selected accession numbers of assemblies that belong to the Clostridia class according to the GTDB taxonomy, which have a high completeness (>= 95%), low contamination (<1%) and fewer than ten *16S rRNA* copies to avoid mis-assemblies (Větrovský & Baldrian, 2013). Using the selected accession numbers, the corresponding genome assembly sequences were subsequently downloaded from the National Center for Biotechnology Information (NCBI). A total of 4073 assemblies were downloaded on 28/10/2020. Within the downloaded assemblies, we had both types of data: culture independent metagenome assembled genomes (MAGs) and genome assemblies from cultured and isolated strains; both with different assembly/fragmentation levels. With this set of selected assemblies, a local database was built with the application Makeblastdb version 2.6.0 (Fig.1). (R Script for retrieval of accession numbers can be found as "Reference_genomes_retrieval.R" in the GitHub repository, Download was done following ncbi-genome-download pipeline: https://github.com/kblin/ncbi-genome-download).

Search of dnaK1, gyrB and 16S rRNA genes in the local database

Once the local database was built, we proceeded to identify and retrieve, within each assembly, the genes of interest: *dnaK1* and *gyrB*, and the widely used universal marker *16S rRNA*. The process of identifying such genes in the assemblies' sequences was performed through similarity with a query sequence. Query sequences for *dnaK1* and *gyrB* were downloaded from the UniProt database. For *dnaK1*, we selected a protein sequence classified as Chaperone protein DnaK (A0A143WYF3), gene '*dnaK1*', size: 594 amino acids, from the organism Clostridiales bacterium CHKCI006. For *gyrB*, we selected a protein sequence classified as DNA gyrase subunit B (R6D2S9), gene '*gyrB*', size: 680 amino acids, from the organism Ruminococcus sp. CAG:579. The *16S rRNA* query sequence was downloaded from NCBI, reported as '16S ribosomal RNA partial sequence' (NR_074399.1), size 1500 base pairs, from the source organism Ruminococcus albus 7 = DSM 20455.

Each query sequence was used to find the coordinates of the queried gene on each whole assembly sequence in the local database. For this, we used two Basic Local Alignment Search Tools (BLAST) provided by the NCBI (Fig.1). For the *16S rRNA* gene, we used the BLASTn program, which carries out a search of a nucleotide query sequence within a nucleotide reference database. The BLASTn was executed with 'max_target_seqs' parameter set to 5000 to indicate the number of aligned sequences to keep and 'max_hsps' parameter set to 1 to keep a single alignment for any query-subject pair. For the protein coding genes *dnaK1* and *gyrB*, we used tBLASTn, a translated version of BLAST which finds regions of local similarity between a query protein sequence compared to the six-frame translations of sequences in a nucleotide reference database (Wheeler & Bhagwat, 2007). tBLASTn was also executed with the 'max_target_seqs'

parameter set to 5000. (Bash Script for BLAST commands can be found as "BLAST_commands.txt" in the GitHub repository).

Extraction of genes' sequences

The BLAST results contained a list of records that aligned with the query. The records, here called hits, contained information of their coordinates in each genome assembly and the frame on which it is read. We accepted BLAST hits with expected values (e-value) equal to 0 as a significant threshold. We took the significant hits' coordinates to extract the nucleotide sequences corresponding to our genes of interest from the reference assemblies. The sequences corresponding to the significant hits for each gene were saved as three independent FASTA files (National Center of Biotechnology Information, 2021), one per gene (*dnaK1*, *gyrB* and *16S rRNA*) (Fig.1). Although we purposely chose house-keeping genes (*dnaK1* and *gyrB*) with one expected copy per genome, we checked for possible occurrences of multi-copies. (Python Script for the creation of new file with extracted genes can be found as "Extract_blasted_genes.py" and Python Script to check repeated hits can be found as "Check_repeated_orgs_gtdb.py" in the GitHub repository).

Alignment and pairwise distances

For each gene, we carried out an alignment with the retrieved sequences. We tested four different aligning programs: PRANK, MUSCLE, MAFFT and Clustal (Edgar, 2004; Katoh & Standley, 2013; Löytynoja, 2014; Sievers *et al.*, 2011). The alignments were run in the Apolo Scientific Computing Center at Universidad EAFIT and were visually inspected for quality check (Fig.1). During this process, the *gyrB* alignment was trimmed at the ends keeping the section between 1281 and 3955 nucleotide positions.

In order to quantify the variation for each marker, we built pairwise distance matrices from the alignments (Fig.2). To compute a distance matrix, each pair of sequences was considered separately to calculate a respective distance value. In brief, two aligned sequences were assumed to be homologous (i.e., to share a common ancestor) and each nucleotide was considered as a character. Similarities mean that the character was conserved, whereas differences were considered as derived traits that originated when substitutions, deletions or insertions occurred in the nucleotide position (Yang & Bielawski, 2000). The pairwise distances give an estimation of the degree of similarity (0) or dissimilarity (1) of two sequences from two different genome assemblies. We calculated the distances between every pair of sequences using the R 'Analyses of Phylogenetics and Evolution' (APE 5.0) package (Paradis & Schliep, 2019) with Kimura 1980 evolution model (K80). This two-parameter model assumes equal base frequencies, one transition probability and one transversion probability (Kimura, 1980).

A frequency distribution of all the pairwise distances for each gene allowed us to contrast how similar or dissimilar were the sequences of the different markers. Additionally, with the pairwise distances, we calculated the reciprocal identities, used later on, with values closer to one (1) indicating similarity between the sequences. (R Script for the pairwise distances calculations and frequency graphs can be found as "Pairwise_distance_matrices_ape.R" in the GitHub repository).

Comparison of marker genes

To understand in more detail how does the genetic variation correlate between marker genes, we compared matrices for

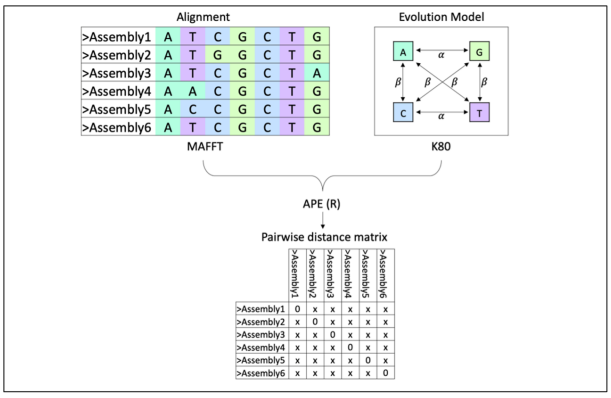

Figure 2 Workflow diagram for alignment and pairwise distances calculation.

every two genes (16S vs. *dnaK1*, 16S vs. *gyrB* and *dnaK1* vs. *gyrB*) (Fig.1). We plotted the pairwise identities of one gene against the pairwise identities of the second gene and calculated linear regressions, which resulted in one regression for the comparison of *dnaK1* vs. *16S rRNA*, a second regression for the comparison of *gyrB* vs. *16S rRNA* and a third one for the comparison of *dnaK1* vs. *gyrB*. These regressions allowed us to compare variability between two genes within the same genome assembly. (R and Python Scripts for the comparison of genes and graphs can be found as "Compare_gene_distances.R" and "Merge_df_distances.py" in the GitHub repository).

Next, we visualized how the variability of the genes behaved at different taxonomic levels, by highlighting data belonging to the same order, family, genus and species throughout the linear regression. To do this, it was necessary to associate each gene sequence to the respective taxonomic classification of the assembly of origin. To assign the corresponding taxonomies, a match was made between the nucleotide ID of the single sequence fragment where the gene was found with its corresponding assembly ID, which has an associated GTDB taxonomic classification. We calculated standard deviations of data at all taxonomic levels to better understand the dispersion of identities of both marker genes (*dnaK1* and *gyrB*) compared to that of the *16S rRNA* gene.

Finally, we also highlighted the comparisons and calculated standard deviations for data belonging to two important families that were relevant in the characterization of the Colombian microbial gut population: Ruminococcaceae and Lachnospiraceae. (Python Scripts to assign taxonomy level data can be found as "Add_taxonomy_level_bool_to_contrast.py" and "Nucleotide_Assembly_IDs_full.py" in the GitHub repository).

Results

Download of the reference dataset

A total of 4073 genome assemblies classified under the class Clostridia by GTDB were downloaded to build what was our working reference dataset. Within the reference data, we had genome assemblies with different assembly levels categorized by the NCBI as: (1) complete genome: chromosomes are gapless and have no runs of 10 or more ambiguous bases, (2) chromosome: there is sequence for one or more chromosomes; this could be a completely sequenced chromosome without gaps or a chromosome containing scaffolds or contigs with gaps between them, (3) scaffold: some sequence contigs have been connected across gaps to create scaffolds, but the scaffolds are all unplaced,

(4) contig: nothing is assembled beyond the level of sequence contigs (NCBI, https://www.ncbi.nlm.nih.gov/assembly/help/). In our data, we had 140 assemblies classified as complete genomes, 38 chromosomes, 2133 scaffolds and 1762 contigs.

Search and extraction of marker genes' sequence from reference assemblies

From these genome assemblies, we extracted by identity three genes which we aimed to compare in terms of variability: the house-keeping genes *dnaK1* and *gyrB*, and the *16S rRNA* universal marker. To retrieve the genes' sequences within the reference genome assemblies, we used BLAST. We obtained a total of 6229 BLAST hits for *dnaK1*, 5283 for *gyrB* and 5000 for *16S rRNA* (Table 1). Then, we filtered the hits by quality, keeping only hits with e-values equal to 0, to finally have a total of 4298 hits for *dnaK1*, 4050 for *gyrB* and 3826 for *16S rRNA* (Table 1).

Although we purposely chose house-keeping genes (*dnaK1* and *gyrB*) with one expected copy per genome, when extracting the genes from the reference genome assemblies we checked for possible occurrences of multi-copies. For *dnaK1*, 33 assemblies had repeated hits, this is, 33 of the reference assemblies obtained two significant hits (e-value=0) for the same gene. In a similar way, for *gyrB*, 25 assemblies had repeated hits (Table 1). It is important to note that the assemblies with repeated hits for *dnaK1* were different from those with repeated hits of *gyrB*, thus it seems not to be explained by consistently faulty assemblies.

Alignment and pairwise distances

All sequences of each gene were aligned. We tested four different aligners with similar results. Here, we worked with the alignments obtained with MAFFT version 7.402 and default parameters (gap opening penalty default = 1.53, offset (works like gap extension penalty) default = 0, maximum number of iterative refinement, default = 0), as this tool proved to be computationally efficient and yielded alignments with long homologous blocks. The alignments were visually inspected for a quality check. During this process, the *gyrB* alignment evidenced particularly low quality, misalignments and many gaps at both ends, and was therefore trimmed by the two extremes to eliminate the deficient sections (see methods). The final alignments for the *dnaK1*, *gyrB* and *16S rRNA* had lengths of respectively 2144, 2675 and 1846 nucleotides.

Based on the aforementioned alignments, we built pairwise distance matrices using the K80 model of sequence evolution to quantify sequence variation between the same marker in two reference assemblies. The frequencies of the pairwise distances for each gene were calculated, where 0 corresponded to equal sequences and 1 indicated totally unlike sequences (Fig.3). The results showed that *16S rRNA* sequences had pairwise distances skewed to 0 along the entire class Clostridia, ranging from 0 to 30 percent (Fig.3A). This means that the sequences for this gene were very similar among assemblies. In contrast, *dnaK1* and *gyrB* showed pairwise distances ranging from 0 to almost 90 percent (Fig.3B and 3C). The fact that *dnaK1* and *gyrB* covered a greater spectrum of pairwise distances for the same group of bacteria (class Clostridia) is an indication that these genes were more variable than the *16S rRNA* gene. A more variable sequence allows to distinguish between closer related organisms; thus, it could lead to a higher taxonomic resolution, capable of revealing more molecular variation between closely related groups. Next, we evaluated how the same pairwise comparison between assemblies behaved along two different marker genes.

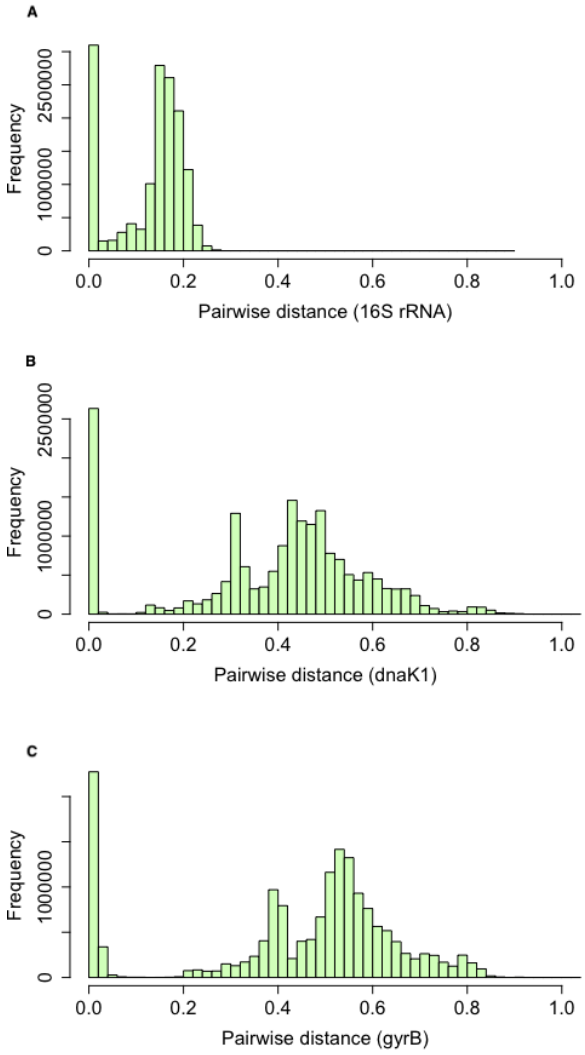

Figure 3 Frequency histograms of pairwise distances in reference assemblies for **A**.*16S rRNA*, **B**.*dnaK1* and **C**.*gyrB* genes. Pairwise distance matrices were built with APE from the alignments by quantifying differences between each pair of sequences. Two aligned sequences were assumed to be homologous, and each nucleotide was considered as a character. Similarities mean that the character was conserved, whereas differences were considered as derived traits that originated when substitutions, deletions or insertions occurred in the nucleotide position. The pairwise distances give an estimation of the degree of similarity (0) or dissimilarity (1) of two sequences from two different genome assemblies.

Comparison of marker genes

Due to the different assembly levels present in our reference genome assemblies, not all of them contained the three genes of interest within the assembled contiguous segment. Some of the assemblies had the three markers (*dnaK1*, *gyrB*, *16S rRNA*) present in a single sequence segment, but other assemblies contained two or only one of them. As our intention was to compare the variability of different genes that came from the same assembly, we worked with a different subset of assemblies for every two genes to compare, which were sets that contained at least the two genes in question. In 355 assemblies, *16S rRNA* and *dnaK1* were simultaneously present, 245 for *16S rRNA* and *gyrB*, and 179 for *dnaK1* and *gyrB* (Table 2).

To estimate how different the variability of one gene was compared to another, we plotted the pairwise identities of the marker gene against the pairwise identities of the *16S rRNA* sequences for assemblies classified under the same taxonomic rank (Fig.4A and 4B). This allowed us to understand and how much resolution is gained by using an alternative marker compared to *16S rRNA* sequences. We also plotted this data to compare the variability between the two proposed marker genes, *dnaK1* and *gyrB* (Fig.4C), and quantified these relationships by calculating linear regressions. We found that within the group studied (class Clostridia), the regression between *dnaK1* and *gyrB* was y=-9.9+x, R^2=0.87, which reflects strong correlation and points out that both have a similar variation. In contrast, we obtained the following regressions for *16S rRNA* with *dnaK1* (Fig.4A) and *16S rRNA* with *gyrB* (Fig.4B), respectively: y=-100+1.9x, R^2=0.5 and y=-160+2.5x, R^2=0.67. This reflects greater variation of both alternative genes in comparison to the traditional ribosomal marker.

The pervious results confirm that, at the class level analyzed here, the house-keeping genes *dnaK1* and *gyrB* have similar sequence identities in most genome pairs and that *16S rRNA* gene identities were higher than both *dnaK1* and *gyrB*. To test our departing hypothesis that less conserved genes, such as *dnaK1* and *gyrB* relative to *16S rRNA*, constitute better markers for identifying finer taxonomic groups, we highlighted how the same comparisons behaved at different taxonomic levels (Fig.5, 6 and 7). This analysis supports the assumption that the finer the taxonomic level, the greater the identity of all genes. Yet, it is interesting to note that the dispersion of identities of both marker genes (*dnaK1* and *gyrB*) compared to that of the *16S rRNA* genes was always greater, at all evaluated taxonomic ranks (Table 2).

Finally, we also highlighted the comparisons within two important, previously mentioned, families within Clostridia: Ruminococcaceae and Lachnospiraceae. We noticed, in both, a similar behavior to the rest of the families included in the analysis (Fig.8).

Discussion

In this study, we aimed to evaluate the suitability of the genes DNA kinase1 (*dnaK1*) and gyrase subunit B (*gyrB*) as genetic markers for characterizing the community diversity of the class Clostridia. We hypothesized that these genes were consistently less conserved than the *16S rRNA* gene and, therefore, constituted appropriate markers to assess the genetic diversity within the group. We found that *dnaK1* and *gyrB* have higher variability compared to that of *16S rRNA* within Clostridia. Our results also indicated that assemblies that are conventionally assigned to a given bacterial species using the *16S rRNA* at a 97% nucleotide identity threshold have an average nucleotide identity of 84.3% in their *dnaK1* gene and 82.5% in their *gyrB* gene. Furthermore, assemblies with 99% overall *16S rRNA* identity, generally classified as strains, have a *dnaK1* nucleotide identity of 88.1% and a *gyrB* nucleotide identity of 87.5%. Thus, using either *dnaK1* or *gyrB* one could identify greater genetic variation than that provided by the *16S rRNA* gene. This pattern was consistent across all taxonomic ranks within Clostridia, ranging from class to species.

The interest in evaluating the scope of alternative markers has increased recently, with several studies comparing specific marker gene's variability to that of the *16S rRNA* and systematically concluding that the alternative genes are more powerful to recover fine-level diversity. For instance, Caro-Quintero & Ochman (2015) evaluated the variability of *gyrB* vs. *16S rRNA* as a proof of concept in wild gorillas' fecal microbiome samples. These samples were selected because they were reported to be enriched in Lachnospiraceae and completely depleted in Bacteroidaceae by previous *16S rRNA* analyses, hence, it constituted a good test for assessing the suitability of the alternative marker at taxonomic levels lower than family. They amplified, from the gorilla fecal samples, a fragment of the *gyrB* gene from members of the two bacterial families of interest (Bacteroidaceae and Lachnospiraceae) and compared clustering results with those obtained with the *16S rRNA* complete gene sequence. For both families, they were able to define more clusters (i.e., higher richness) using *gyrB* than with *16S rRNA*. To estimate how much resolution was gained by using *gyrB* compared to 16S RNA a linear regression was calculated between the identities of the targeted *gyrB* region and the full-length *16S rRNA* for organisms classified at the same taxonomic rank (comparable to figure 4B). Their results showed that organisms with $\geq$97% overall *16S rRNA* identity, had a *gyrB* nucleotide identity of $\geq$88% (Caro-Quintero & Ochman, 2015). This last result resembles what we obtained targeting a different microbial group with the same marker: an average nucleotide identity of 82.5% in *gyrB* for organisms with 97% *16S rRNA* average identity within the Clostridia dataset.

Similarly, Moeller *et al.* (2016) evaluated *gyrB* to profile strain diversity within the gut microbiomes of humans, chimpanzees, bonobos and gorillas. For their case, the use of a fine-scale resolution marker was essential to understand phylogenetic relationships between closely related microorganisms, which ultimately would enable to test for co-speciation between gut bacteria and hominids. They were able to demonstrate, with the use of *gyrB*, co-speciation between gut bacteria of the families Bacteroidaceae and Bifidobacteriaceae with hominids across hundreds of thousands of host generations (Moeller *et al.*, 2016).

In the same way, Guo *et al.* (2019) compared the variation of two different single-copy ribosomal protein genes, rplB and rpsC, with the *16S rRNA* in complete bacterial genomes available in NCBI RefSeq. Within their microbial groups of interest from plant rhizospheres (orders Rhizobiales and Pseudomonadales, genera Rhizobium, and Pseudomonas), they showed that rplB and rpsC genes had larger variations among the genomes than the *16S rRNA* within their corresponding order and genus. When taking into account genomes that belonged to the same genus, they found that *16S rRNA* had an average identity of 95.2% while rplB and rpsC 87.2% and 90.3%, respectively. In their analyses, rplB and rpsC yielded significantly higher alpha diversities than the *16S rRNA*. Furthermore, using these alternative markers, they were able to separate microbial communities from three different plants' rhizospheres. Their conclusion confirmed that these faster evolving marker genes offer increased

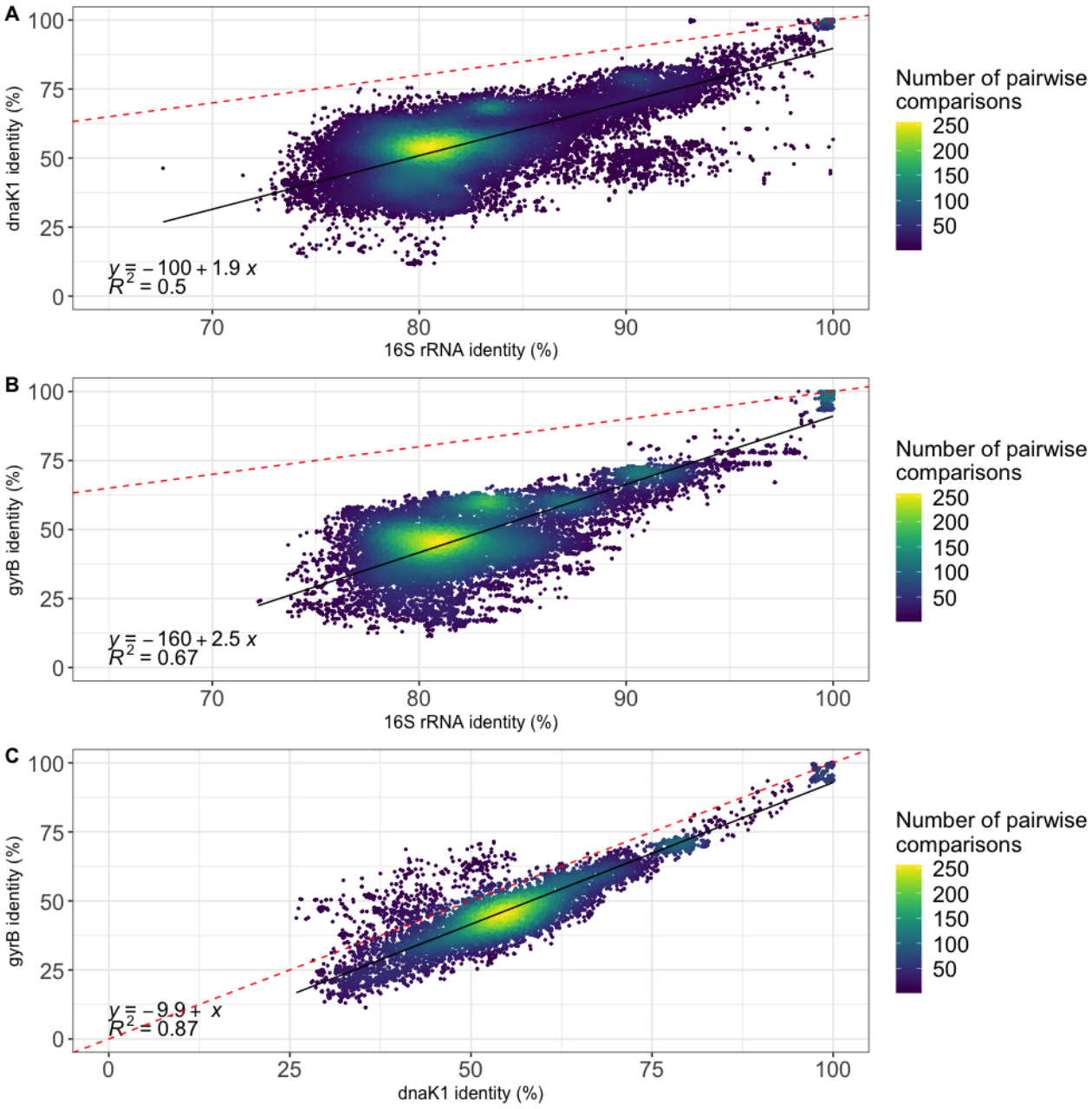

Figure 4 Association of pairwise comparisons among Clostridia genome assemblies using *16S rRNA*, *dnaK1* and *gyrB* genes. Each dot on the plots represents a pair of assemblies, the colors indicate the density of pairwise comparisons with those values. The red dashed line is y = x. Data below the y=x line indicates the gene on the X axis is more conserved than the gene on the Y axis. The black line is a regression that allows to quantify the differences in variability. **A**.Association between *16S rRNA* and *dnaK1*. Note that a *16S rRNA* sequence identity value of 97%, which is conventionally used to delineate bacterial species, corresponds to 84% nucleotide sequence identity for *dnaK1*. **B**.Association between *16S rRNA* and *gyrB*. Note that a *16S rRNA* sequence identity value of 97% corresponds to 82.5% nucleotide sequence identity for *dnaK1*. **C**.Association between *gyrB* and *dnaK1*.

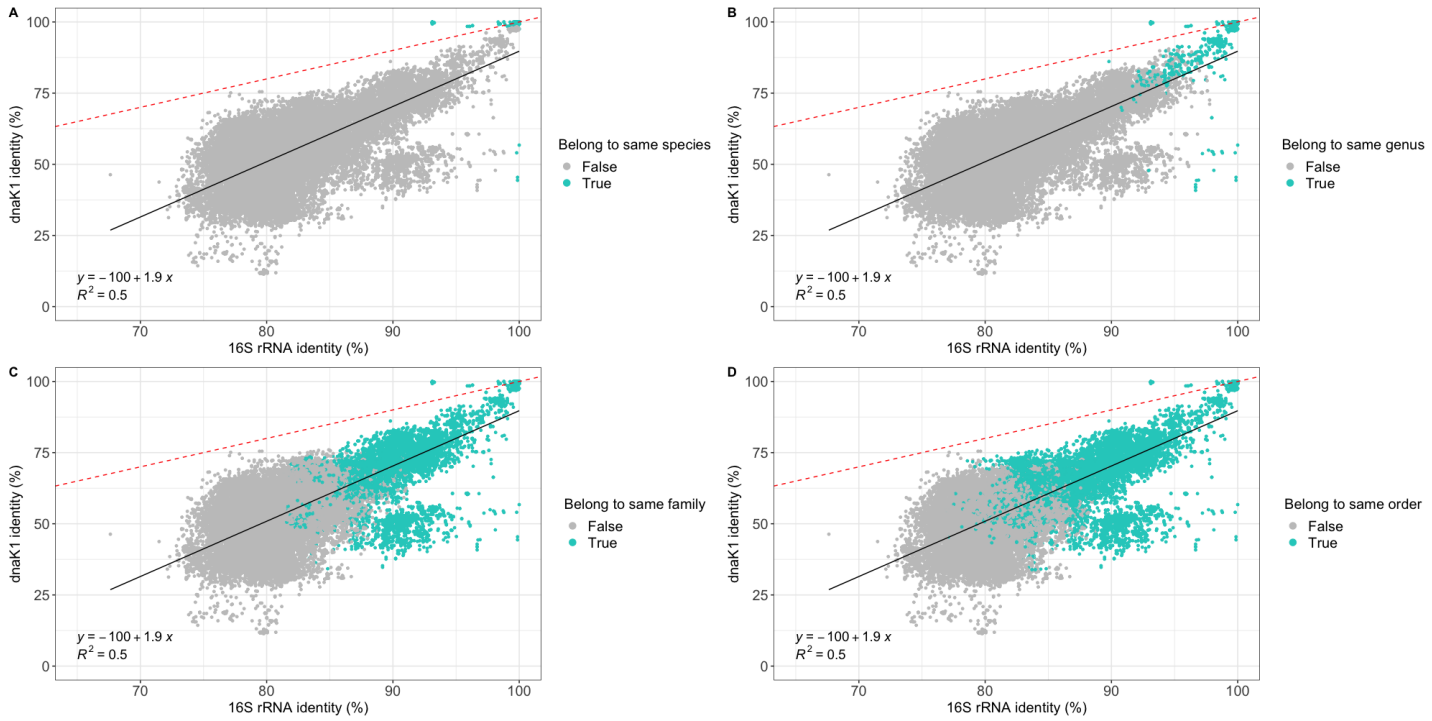

Figure 5 Association between the degree of sequence identity of *16S rRNA* and *dnaK1* genes among Clostridia genome assemblies. Each dot on the plots represents a pair of assemblies, the red dashed line is y = x and the black line is a regression that allows to quantify the differences in variability. Color blue depicts the same taxonomic level. **A**.species. **B**.genus. **C**.family. **D**.order.

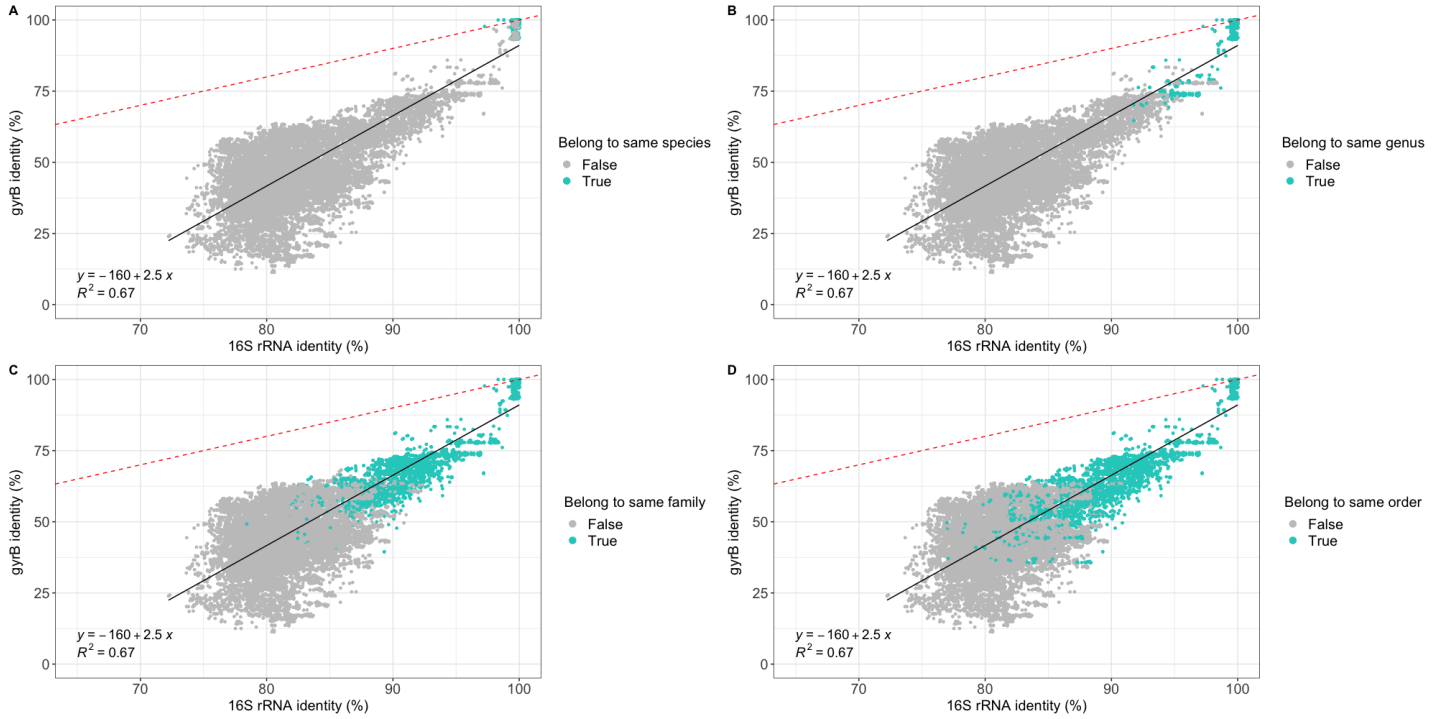

Figure 6 Association between the degree of sequence identity of *16S rRNA* and *gyrB* genes among Clostridia genome assemblies. Each dot on the plots represents a pair of assemblies, the red dashed line is y = x and the black line is a regression that allows to quantify the differences in variability. Color blue depicts the same taxonomic level. **A**.species. **B**.genus. **C**.family. **D**.order.

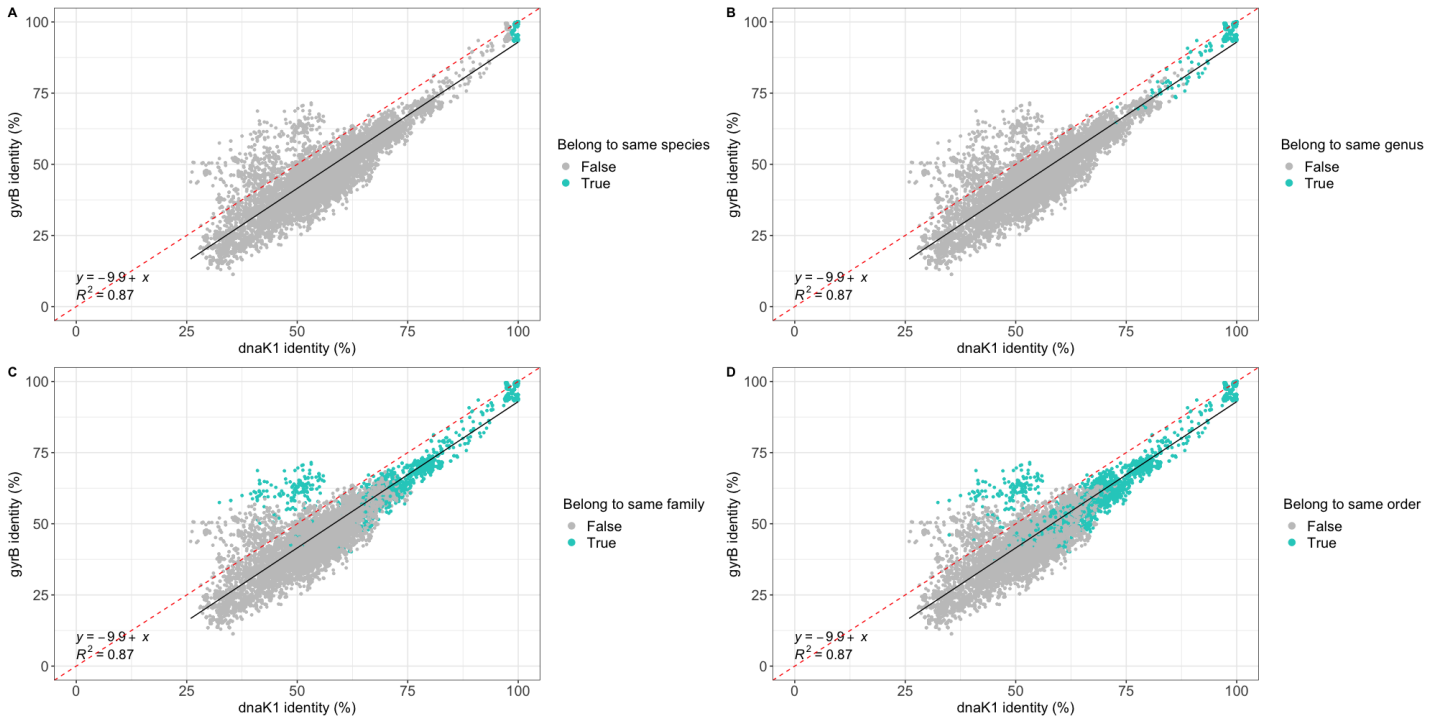

Figure 7 Association between the degree of sequence identity of *dnaK1* and *gyrB* genes among Clostridia genome assemblies. Each dot on the plots represents a pair of assemblies, the red dashed line is y = x and the black line is a regression that allows to quantify the differences in variability. Color blue depicts the same taxonomic level. **A**.species. **B**.genus. **C**.family. **D**.order.

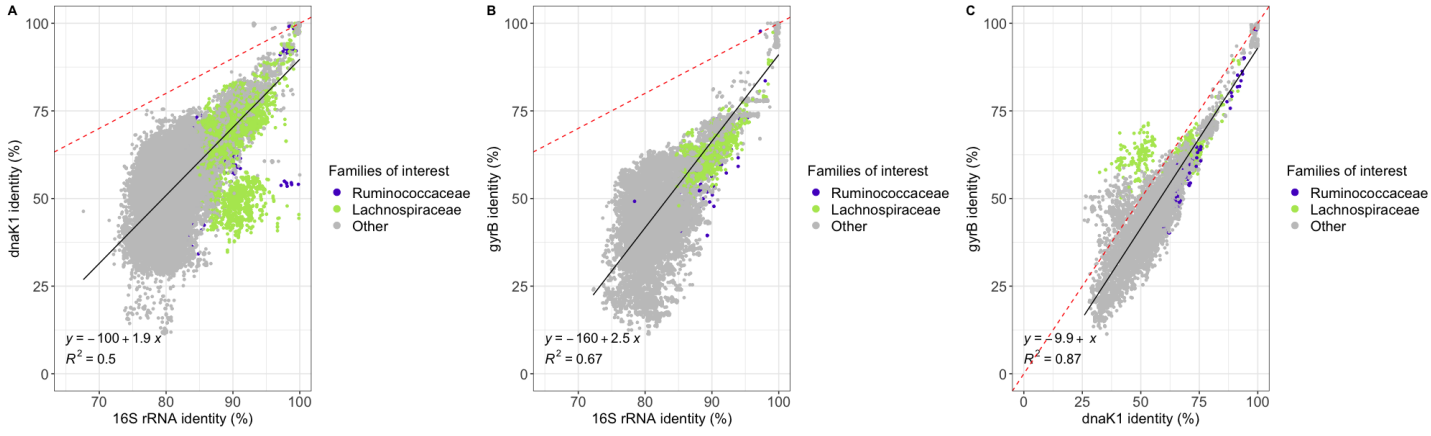

Figure 8 Association of pairwise comparisons among Clostridia genome assemblies using *16S rRNA*, *dnaK1* and *gyrB* genes highlighting data from Ruminococcaceae and Lachnospiraceae families. Each dot on the plots represents a pair of assemblies, the red dashed line is y = x and the black line is a regression that allows to quantify the differences in variability. **A**.Association between *16S rRNA* and *dnaK1*. **B**.Association between *16S rRNA* and *gyrB*. **C**.Association between *gyrB* and *dnaK1*.

resolution to reflect the ecology of microbial communities (Guo *et al.*, 2019).

Despite the well-known fact that different parts of the genome evolve at different rates, there seems to be a consistent pattern of divergence of (some) house-keeping genes with respect to 16S. Olm *et al.* (2020) calculated gene identity thresholds for species delimitation (i.e., the average percentage of identity corresponding to the same species) for over 100 single-copy marker genes based on genomes deposited in RefSeq, including our genes of interest (*dnaK1* and *gyrB*). A long-standing question in microbiology has been to understand to what extent the concept of species, as discrete categories, applies to prokaryotes. In their recent work, Olm *et al.* support the idea that distinct discrete sequences groups or "microbial species" are defined by a 95% whole genome average nucleotide identity (ANI), as originally proposed with respect to DNA-DNA hybridization (Goris *et al.*, 2007; Olm *et al.*, 2020). Moreover, this value did not seem to be in a continuous range but showed a pattern of discrete clusters of genomes.

With this "microbial species" definition, Olm *et al.* (2020) showed that many specific genetic markers can act as effective proxies for whole genome ANI values and, consequently, have high species discrimination power, while some others cannot. This study showed how the *16S rRNA* marker is not as efficient as protein coding single-copy genes because it is not variable enough (Olm *et al.*, 2020). Moreover, the threshold of 97% identity in *16S rRNA* sequences, that is commonly used for operational taxonomical unit (OTU) clustering, seems inaccurate and not representative of microbial species discrimination (Edgar, 2018; Olm *et al.*, 2020). In fact, bacteria of different species can resemble up to 99% in their *16S rRNA* sequence (Olm *et al.*, 2020). Interestingly, among the +100 genes evaluated by Olm and colleagues, both of the genes we evaluated here (*dnaK1* and *gyrB*) showed high species discrimination power. According to them, a threshold reflecting the average percentage identity corresponding to the same species was for *dnaK1* and *gyrB* of 97.7% and 96.7%, respectively, as opposed to 99.5% for *16S rRNA* (Olm *et al.*, 2020). In parallel, our results, despite not being directly comparable since they exclusively consider the Clostridia class, also showed, for assemblies classified under the same species, mean nucleotide identities over 90% for *dnaK1*, *gyrB* and *16S rRNA*, with identities for *16S rRNA* on average higher than those of the other two genes. In addition, Olm *et al.* effectively demonstrated that within their dataset, *dnaK1* had an average copy number of 1.01 per genome and was present in 68.32% of their reference genomes. Similarly, *gyrB* had an average copy number of 1.002 and was recoverable from 57.85% of their reference genomes. On the other hand, *16S rRNA* was significantly less often recoverable from their reference genomes and was multicopy in 56.3% of them (Olm *et al.*, 2020), emphasizing the advantages of these alternative markers.

Together, our study and those mentioned above consistently confirm that the use of alternative markers can help to elucidate finer diversity than what is observed with the traditionally used ribosomal marker *16S rRNA*. In our case, we obtained this pattern when testing two genes in a specific taxonomic group. However, it seems to be a generalized pattern with several other genes, in different microbial groups, different environmental contexts, samples and even in different taxonomic ranks (Caro-Quintero & Ochman, 2015; Case *et al.*, 2007; Guo *et al.*, 2019; Olm *et al.*, 2020; Vos, Quince, Pijl, de Hollander, & Kowalchuk, 2012).

Although the above arguments favor the use of alternative markers, they do not necessarily suggest that *16S rRNA* based methods should be avoided. As mentioned earlier, this ribosomal marker has historically allowed extraordinary insights in understanding the general diversity of prokaryotic domains. Its downside (i.e., multiple intragenomic copies and not enough variability for fine level identification) is clear but its advantages are imperative: it is ubiquitous and allows the use of universal primers, which has served immensely for wide diversity approaches. Further, it is definitely key in elucidating diversity in unfamiliar environments at broad taxonomic levels. It is obvious that the choice of a methodology depends enormously on the scientific question to be answered. In other words, *16S rRNA* approach and marker gene techniques serve for different purposes, *16S rRNA* acts like a wide-angle lens on microbial diversity while alternative marker genes, such as *dnaK1* and *gyrB*, act like microscopes for specific groups. These two approaches can even complement each other for a comprehensive study of diversity.

In summary, for our case, where we were interested in the tools to assess the fine diversity of a specific taxonomical group, choosing a methodology based on protein-coding single-copy marker genes, is the most reasonable and informative approach. However, it is worth mentioning that these genes may not be as ubiquitous as *16S rRNA* across prokaryotes, the design of universal primers may not be possible, requiring a careful design of primers to target the microbial group of interest. Methodologies for this targeted primer design are already available (phyloTAGs) (Caro-Quintero & Ochman, 2015).

To conclude, our results validate the departing hypothesis that either the use of *dnaK1* or *gyrB* as marker genes for the Clostridia class can be more informative to assess diversity at fine taxonomic levels than the traditional ribosomal marker *16S rRNA*. Moreover, several studies have shown that this pattern is consistent among other alternative markers within different interest groups and systematically elucidate greater diversity. The implementation of these cost-effective methodologies can lead to efficiently resolve close taxonomic relationships and reclassify polyphyletic groups (e.g., the genus Ruminococcus within Clostridia (La Reau *et al.*, 2016)). The current availability of massive data is making it possible to tackle these questions in greater detail. Finally, we foresee analysis of protein-coding genes as being complementary to the *16S rRNA* approach to answering questions that span different taxonomic levels and address bacterial identification, classification, phylogenetics and evolution. Complementing both methods may help elucidate the biology and function of members of this and other groups, which could have key implications for animal health, in particular, human health and, therefore, may even hold promise for profit in areas such as personalized medicine and nutrition.

Acknowledgments

Special thanks to all members of Vidarium—Nutrition Health and Wellness Research Center for my fruitful training as a professional. The authors thank Jacobo de la Cuesta-Zuluaga for his help with the genome retrieval protocol and for his support providing important references.

Table 1 Review data from the search of *dnaK1*, *gyrB* and *16S rRNA* genes in 4073 reference genome assemblies.

Gene	Query length	BLAST parameters	Hits filtered by e-value	Repeated hits*
dnaK1	594 amino acids	max_target_seqs=5000	4298	33
gyrB	680 amino acids	max_target_seqs=5000	4050	25
16S rRNA	1500 nucleotides	max_target_seqs=5000 max_hsps=1	3826	0

*Number of sequence segments with repeated hits

Table 2 Standard deviations (SD) as a measure of dispersion of identities of gene comparisons (*16S rRNA* vs. *dnaK1*, *16S rRNA* vs. *gyrB*, *dnaK1* vs. *gyrB*) at all taxonomic levels and two families of interest (Ruminococcaceae and Lachnospiraceae).

16S rRNA* vs. *dnaK1

355 assemblies with presence of both genes in within the assembled contiguous segment

Pairwise comparisons	Total	16S rRNA identity SD	dnaK1 identity SD
Belong to same class	62835	4.809	13.193
Belong to same order	12031	4.864	13.726
Belong to same family	8972	4.122	14.127
Belong to same genus	1832	2.035	7.946
Belong to same species	1092	1.065	3.039
Belong to same family Ruminococcaceae	300	4.656	11.526
Belong to same family Lachnospiraceae	4186	2.623	11.562

16S rRNA* vs. *gyrB

245 assemblies with presence of both genes in within the assembled contiguous segment

Pairwise comparisons	Total	16S rRNA identity SD	gyrB identity SD
Belong to same class	29871	5.562	16.780
Belong to same order	6200	5.281	16.141
Belong to same family	5197	4.632	14.351
Belong to same genus	1777	1.339	6.768
Belong to same species	1329	0.194	2.376
Belong to same family Ruminococcaceae	105	4.053	12.269
Belong to same family Lachnospiraceae	816	2.688	6.351

dnak1* vs. *gyrB

179 assemblies with presence of both genes in within the assembled contiguous segment

Pairwise comparisons	Total	dnaK1 identity SD	gyrB identity SD
Belong to same class	15931	14.951	16.519
Belong to same order	3432	14.457	15.699
Belong to same family	2874	14.018	14.561
Belong to same genus	915	4.302	5.790
Belong to same species	590	0.438	2.649
Belong to same family Ruminococcaceae	105	10.269	12.915
Belong to same family Lachnospiraceae	741	10.672	6.124

References

Aguilar-Rodríguez, J., Sabater-Muñoz, B., Montagud-Martínez, R., Berlanga, V., Alvarez-Ponce, D., Wagner, A., & Fares, M. A. (2016). The molecular chaperone dnak is a source of mutational robustness. Genome Biology and Evolution, 8(9), 2979–2991. https://doi.org/10.1093/gbe/evw176

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature Biotechnology, 31(6), 533–538. https://doi.org/10.1038/nbt.2579

Caro-Quintero, A., & Ochman, H. (2015). Assessing the unseen bacterial diversity in microbial communities. Genome Biology and Evolution, 7(12), 3416–3425. https://doi.org/10.1093/gbe/evv234

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., & Kjelleberg, S. (2007). Use of *16S rRNA* and rpoB genes as molecular markers for microbial ecology studies. Applied and Environmental Microbiology, 73(1), 278–288. https://doi.org/10.1128/AEM.01177-06

Crosby, L. D., & Criddle, C. S. (2003). Understanding bias in microbial community analysis techniques due to rrn operon copy number heterogeneity. BioTechniques, 34(4), 790–802. https://doi.org/10.2144/03344rr01

De Filippis, F., Pasolli, E., & Ercolini, D. (2020). Newly Explored Faecalibacterium Diversity Is Connected to Age, Lifestyle, Geography, and Disease. Current Biology, 30(24), 4932-4943.e4. https://doi.org/10.1016/j.cub.2020.09.063

de la Cuesta-Zuluaga, J., Corrales-Agudelo, V., Velásquez-Mejía, E. P., Carmona, J. A., Abad, J. M., & Escobar, J. S. (2018). Gut microbiota is associated with obesity and cardiometabolic disease in a population in the midst of Westernization. Scientific Reports, 8(1), 1–14. https://doi.org/10.1038/s41598-018-29687-x

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics, 34(14), 2371–2375. https://doi.org/10.1093/bioinformatics/bty113

Fox, G. E., Pechman, K. R., & Woese, C. R. (1977). Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. International Journal of Systematic Bacteriology, 27(1), 44–57. https://doi.org/10.1099/00207713-27-1-44

Franco-Duarte, R., Černáková, L., Kadam, S., Kaushik, K. S., Salehi, B., Bevilacqua, A., … Rodrigues, C. F. (2019). Advances in chemical and biological methods to identify microorganisms—from past to present. Microorganisms, 7(5). https://doi.org/10.3390/microorganisms7050130

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology, 57(1), 81–91. https://doi.org/10.1099/ijs.0.64483-0

Guo, J., Cole, J. R., Brown, C. T., & Tiedje, J. M. (2019). Title: Comparing faster evolving rplB and rpsC versus SSU rRNA for improved microbial community resolution. Preprint BioRxiv, 1–28.

Head, I. M., Saunders, J. R., & Pickup, R. W. (1998). Microbial evolution, diversity, and ecology: A decade of ribosomal RNA analysis of uncultivated microorganisms. Microbial Ecology, 35(1), 1–21. https://doi.org/10.1007/s002489900056

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution, 16(2), 111–120. https://doi.org/10.1007/BF01731581

La Reau, A. J., Meier-Kolthoff, J. P., & Suen, G. (2016). Sequence-based analysis of the genus Ruminococcus resolves its phylogeny and reveals strong host association. Microbial Genomics, 2(12), e000099. https://doi.org/10.1099/mgen.0.000099

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. In Methods in Molecular Biology (Vol. 1079, pp. 155–170). https://doi.org/10.1007/978-1-62703-646-7_10

Moeller, A. H., Caro-Quintero, A., Mjungu, D., Georgiev, A. V., Lonsdorf, E. V., Muller, M. N., … Ochman, H. (2016). Cospeciation of gut microbiota with hominids. Science, 353(6297), 380–382. https://doi.org/10.1126/science.aaf3951

National Center of Biotechnology Information. (2021). BLAST Topics. Retrieved April 24, 2021, from https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp

Nguyen, N.-P., Warnow, T., Pop, M., & White, B. (2002). A perspective on *16S rRNA* operational taxonomic unit clustering. Npj Biofilms and Microbiomes, 57(6), 10–13. https://doi.org/10.1038/npjbio

Olm, M. R., Crits-christoph, A., Diamond, S., Lavy, A., Carnevali, P. B. M., & Banfield, J. F. (2020). Bacterial Species Boundaries. MSystems, 5(1), e00731-19.

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics, 35(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Parks, D. H., Chuvochina, M., Chaumeil, P. A., Rinke, C., Mussig, A. J., & Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. Nature Biotechnology, 38(9), 1079–1086. https://doi.org/10.1038/s41587-020-0501-8

Pei, A. Y., Oberdorf, W. E., Nossa, C. W., Agarwal, A., Chokshi, P., Gerz, E. A., … Pei, Z. (2010). Diversity of *16S rRNA* genes within individual prokaryotic genomes. Applied and Environmental Microbiology, 76(12), 3886–3897. https://doi.org/10.1128/AEM.02953-09

Santos, S. R., & Ochman, H. (2004). Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. Environmental Microbiology, 6(7), 754–759. https://doi.org/10.1111/j.1462-2920.2004.00617.x

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. Frontiers in Plant Science, 5(JUN), 1–14. https://doi.org/10.3389/fpls.2014.00209

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., … Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology, 7(539). https://doi.org/10.1038/msb.2011.75

Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic note: A place for DNA-DNA reassociation and *16S rRNA* sequence analysis in the present species definition in bacteriology. International Journal of Systematic Bacteriology, 44(4), 846–849. https://doi.org/10.1099/00207713-44-4-846

Sun, D. L., Jiang, X., Wu, Q. L., & Zhou, N. Y. (2013). Intragenomic heterogeneity of *16S rRNA* genes causes overestimation of prokaryotic diversity. Applied and Environmental Microbiology, 79(19), 5962–5969. https://doi.org/10.1128/AEM.01282-13

Větrovský, T., & Baldrian, P. (2013). The Variability of the *16S rRNA* Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLoS ONE, 8(2), 1–10. https://doi.org/10.1371/journal.pone.0057923

Vos, M., Quince, C., Pijl, A. S., de Hollander, M., & Kowalchuk, G. A. (2012). A comparison of rpoB and *16S rRNA* as markers in pyrosequencing studies of bacterial diversity. PLoS ONE, 7(2), 1–8. https://doi.org/10.1371/journal.pone.0030600

Wagner, A. (2002). Selection and gene duplication: A view from the genome. Genome Biology, 3(5), 3–5. https://doi.org/10.1186/gb-2002-3-5-reviews1012

Wang, L. T., Lee, F. L., Tai, C. J., & Kasai, H. (2007). Comparison of *gyrB* gene sequences, *16S rRNA* gene sequences and DNA-DNA hybridization in the Bacillus subtilis group. International Journal of Systematic and Evolutionary Microbiology, 57(8), 1846–1850. https://doi.org/10.1099/ijs.0.64685-0

Watt, P. M., & Hickson, I. D. (1994). Structure and function of type II DNA topoisomerases. Biochemical Journal, 303(3), 681–695. https://doi.org/10.1042/bj3030681

Wells, C. L., & Wilkins, T. D. (1996). Clostridia: Sporeforming Anaerobic Bacilli. In S. Baron (Ed.), Medical Microbiology (4th ed.). Galveston: University of Texas Medical Branch at Galveston.

Wheeler, D., & Bhagwat, M. (2007). BLAST QuickStart: example-driven web-based BLAST tutorial. In Methods in molecular biology (Clifton, N.J.) (Vol. 395). https://doi.org/10.1007/978-1-59745-514-5_9

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America, 74(11), 5088–5090. https://doi.org/10.1073/pnas.74.11.5088

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences of the United States of America, 87(12), 4576–4579. https://doi.org/10.1073/pnas.87.12.4576

Woese, Carl R., Fox, G. E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., … David, S. (1975). Conservation of primary structure in 16S ribosomal RNA. Nature, 254, 83–86.

Yamamoto, S., & Harayama, S. (1995). PCR amplification and direct sequencing of *gyrB* genes with universal primers and their application to the detection and taxonomic analysis of Pseudomonas putida strains. Applied and Environmental Microbiology, 61(3), 1104–1109. https://doi.org/10.1128/aem.61.3.1104-1109.1995

Yamamoto, Satoshi, & Harayama, S. (1996). Phylogenetic analysis of Acinetobacter strains based on the nucleotide sequences of *gyrB* genes and on the amino acid sequences of their products. International Journal of Systematic Bacteriology, 46(2), 506–511. https://doi.org/10.1099/00207713-46-2-506

Yang, Z., & Bielawski, J. R. (2000). Statistical methods for detecting molecular adaptation. Trends in Ecology and Evolution, 15(12), 496–503. https://doi.org/10.1016/S0169-5347(00)01994-7

Zuckerkandl, E., & Pauling, L. (1965). Molecules as documents of history. Journal of Theoretical Biology, 8(2), 357–366.