RESEARCH ARTICLE

REVISED Concatenated 16S rRNA sequence analysis improves bacterial taxonomy [version 3; peer review: 2 approved]

Bobby Paul

Department of Bioinformatics, Manipal School of Life Sciences, Manipal Academy of Higher Education, Manipal, Karnataka, 576104, India

v3 **First published:** 19 Dec 2022, **11**:1530
https://doi.org/10.12688/f1000research.128320.1
Second version: 03 Apr 2023, **11**:1530
https://doi.org/10.12688/f1000research.128320.2
Latest published: 01 Sep 2023, **11**:1530
https://doi.org/10.12688/f1000research.128320.3

Abstract

Background: Microscopic, biochemical, molecular, and computer-based approaches are extensively used to identify and classify bacterial populations. Advances in DNA sequencing and bioinformatics workflows have facilitated sophisticated genome-based methods for microbial taxonomy although sequencing of the 16S rRNA gene is widely employed to identify and classify bacterial communities as a cost-effective and single-gene approach. However, the 16S rRNA sequence-based species identification accuracy is limited because of the occurrence of multiple copies of the 16S rRNA gene and higher sequence identity between closely related species. The availability of the genomes of several bacterial species provided an opportunity to develop comprehensive species-specific 16S rRNA reference libraries.

Methods: Sequences of the 16S rRNA genes were retrieved from the whole genomes available in the Genome databases. With defined criteria, four 16S rRNA gene copy variants were concatenated to develop a species-specific reference library. The sequence similarity search was performed with a web-based BLAST program, and MEGA software was used to construct the phylogenetic tree.

Results: Using this approach, species-specific 16S rRNA gene libraries were developed for four closely related *Streptococcus* species (*S. gordonii*, *S. mitis*, *S. oralis*, and *S. pneumoniae*). Sequence similarity and phylogenetic analysis using concatenated 16S rRNA copies yielded better resolution than single gene copy approaches.

Conclusions: The approach is very effective in classifying genetically closely related bacterial species and may reduce misclassification of bacterial species and genome assemblies.

Keywords

bacterial nomenclature, bacterial taxonomy, concatenated phylogeny, species-specific barcode reference library

Open Peer Review

Approval Status ✓ ✓

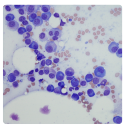
	1	2
version 3 (revision) 01 Sep 2023	✓ view	✓ view
version 2 (revision) 03 Apr 2023	✓ view	? view
version 1 19 Dec 2022	? view	

1. **Siddaramappa Shivakumara**, Institute of Bioinformatics and Applied Biotechnology, Bengaluru, India

2. **Wellyzar Sjamsuridzal**, Universitas Indonesia, Depok, Indonesia

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Manipal Academy of Higher Education gateway.



This article is included in the Cell & Molecular Biology gateway.

Corresponding author: Bobby Paul (bobby.paul@manipal.edu)

Author roles: Paul B: Conceptualization, Data Curation, Formal Analysis, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Open access funding was provided by Manipal Academy of Higher Education, Manipal.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Paul B. **Concatenated 16S rRNA sequence analysis improves bacterial taxonomy [version 3; peer review: 2 approved]** F1000Research 2023, **11**:1530 https://doi.org/10.12688/f1000research.128320.3

First published: 19 Dec 2022, **11**:1530 https://doi.org/10.12688/f1000research.128320.1

REVISED **Amendments from Version 2**

Added limitations of this approach in the conclusion section.

Any further responses from the reviewers can be found at the end of the article

Introduction

The genomic region encoding the 16S ribosomal RNA (16S rRNA) is extensively studied, and used to identify and classify bacterial species. The 16S rRNA is a conserved component of the small subunit (30S) of the prokaryotic ribosome. The gene encoding the 16S rRNA is ~1500 base pair (bp) long, and it consists of nine variable regions (Reller *et al.* 2007; Chakravorty *et al.* 2007; Sabat *et al.* 2017). The sequence of the 16S rRNA gene has been extensively used as a molecular marker in culture-independent methods to identify and classify diverse bacterial communities (Clarridge 2004; Johnson *et al.* 2019). Bacterial 16S rRNA sequences are currently being used to study the evolution, phylogenetic relationships, and environmental abundance of various taxa (Vetrovsky and Baldrian 2013; Srinivasan *et al.* 2015; Peker *et al.* 2019).

Although 16S rRNA sequence analyses are the mainstay of taxonomic studies of bacteria, there are some limitations. For example, the 16S rRNA gene has poor discriminatory power at the species level (Winand *et al.* 2020), and the copy number per genome can vary from 1 to 15 or even more (Vetrovsky and Baldrian 2013; Winand *et al.* 2020). The variable copies of this gene within a genome makes distinct data for a species. Therefore, gene copy normalization (GCN) may be necessary prior to sequence analysis. However, GCN may not improve the 16S rRNA sequence analyses in all scenarios, and comprehensive, species-specific catalogues of 16S rRNA gene copies may be necessary (Starke *et al.* 2021). Furthermore, intra-species variations in the 16S rRNA gene copies were observed in several bacterial genome assemblies (Paul *et al.* 2019). Only a few bacterial species contain identical 16S rRNA gene copies, and sequence diversity increases with increasing copy numbers of 16S rRNA genes (Vetrovsky and Baldrian 2013). The high levels of similarity of the 16S rRNA gene across some bacterial species poses a major challenge for taxonomic studies using bioinformatics methods (Deurenberg *et al.* 2017; Peker *et al.* 2019).

Factors such as purity of bacterial cultures, quality of the purified DNA samples, and potential DNA chimeras should be carefully considered while sequencing and analysis of 16S rRNA genes (Janda and Abbott 2007; Church *et al.* 2020). Sequencing errors can lead to misidentification of bacteria and phylogenetic anomalies (Alachiotis *et al.* 2013). Other concerns include sequence ambiguities, gaps generated during DNA sequencing and sequence comparisons, and choosing the appropriate algorithm (local or global) for sequence alignment. Since the local alignment algorithm is extensively used for sequence similarity-based comparisons, it is important to carefully consider whether a single variable region or a combination of variable regions of the 16S rRNA gene would be ideal for bacterial classification (Janda and Abbott 2007; Johnson *et al.* 2019; Winand *et al.* 2020). Using erroneous 16S rRNA sequences as references and improper bioinformatics workflows can mislead bacterial identification. Further, the growth of bioinformatics and genetic data has led to the current genome-based microbial classification. However, the success rate of these approaches are highly dependent on the skill of data analyst personnel in next generation sequencing technologies, computational tools, operation of high performance computing systems. Researchers without sufficient experience or skill in such technologies may also mislead the bacterial taxonomy (Baltrus 2016).

Other methods for bacterial identification include the sequencing and analysis of the polymerase chain reaction (PCR) amplified ~4.5 kb 16S–23S rRNA regions (Benitez-Paez and Sanz 2017; Sabat *et al.* 2017; Kerkhof *et al.* 2017). However, the 16S–23S rRNA sequence-based method is less practical application due to the lack of appropriate reference sequence databases and reliable tools/methods for sequence analysis (Sabat *et al.* 2017). Recent advances in bioinformatics workflows (Winand *et al.* 2020; Schloss 2020) and reference databases such as SILVA, EzBioCloud (Quast *et al.* 2013; Yoon *et al.* 2017) have further improved 16S rRNA-based bacterial taxonomy. However, these approaches are not completely reliable due to misclassification of some bacterial species and erroneous genome assemblies (Steven *et al.* 2017; Martínez-Romero *et al.* 2018; Mateo-Estrada *et al.* 2019; Bagheri *et al.* 2020).

The entire 16S rRNA gene (~1500 bp) can be amplified and sequenced using the conventional or high throughput sequencing methods. However, many 16S rRNA sequence-based bacterial identification studies do not seem to include all of these nine variable regions (Stackebrandt *et al.* 2021). Due to the large volume of whole-genome data that is being produced by high throughput sequencing technologies, there is an urgent need to translate the genomic data for convenient microbiome analyses that ensure clinical practitioners can readily understand and quickly implement (Church *et al.* 2020). This study aimed to develop a workflow for accurate identification of bacteria using concatenated,

species-specific 16S rRNA sequences. It was hoped that the species-specific libraries would yield much better resolution in sequence similarity- and phylogeny-based bacterial classification.

Methods

Estimation of variations in intra-genomic 16S rRNA gene copies

It has been reported that sequence alignment of 16S rRNA gene copies at the intra-genomic level shows a higher degree of variability in species belonging to the *Firmicutes* and *Proteobacteria* (Vetrovsky and Baldrian 2013; Ibal *et al.* 2019). Therefore, this study used eight 16S rRNA gene copies (Underlying data: Supplementary data 1 (Paul 2022)) retrieved from the complete genome of *Enterobacter asburiae* strain ATCC 35953 (NZ_CP011863.1). To estimate intra-genomic variability between these 16S rRNA gene copies, BLAST+ 2.13.0 (RRID:SCR_004870; Altschul *et al.* 1990) and Clustal Omega 1.2.4 (RRID:SCR_001591; Sievers *et al.* 2011) sequence alignment algorithms were used. Previous studies suggested unweighted pair group method with arithmetic averages (UPGMA) algorithm for the phylogenetic analysis of 16S rRNA genes (Clarridge 2004; Caporaso *et al.* 2011). Hence, phylogenetic analysis of these 16S rRNA gene copies were performed using the UPGMA method (Maximum Composite Likelihood; 500 bootstrap replicates) provided in the MEGA software (version 11; RRID: SCR_000667; Kumar *et al.* 2018).

Construction of species-specific concatenated 16S rRNA reference libraries

Previous studies have reported that the genes encoding 16S rRNA from several bacterial species share >99% sequence identity (Deurenberg *et al.* 2017; Peker *et al.* 2019). Therefore, the 16S rRNA-based methods failed to correctly identify bacterial species that are genetically closely related (Deurenberg *et al.* 2017; Devanga-Ragupathi *et al.* 2018). It has been reported that 16S rRNA-based methods cannot distinguish between *Streptococcus mitis* and *Streptococcus pneumoniae* due to the high sequence similarity (Reller *et al.* 2007; Lal *et al.* 2011). Hence, the study decided to choose the 16S rRNA gene copies from four closely related species of *Streptococcus.*

More than 552,575 whole-genome sequences are currently (Aug 2023) available for bacterial species in the Genome database (RRID:SCR_002474; https://www.ncbi.nlm.nih.gov/genome). Many of these genomes were sequenced using high throughput sequencing technologies such as Illumina/Ion-Torrent (short read sequencing) and PacBio/Nanopre (long read sequencing). Furthermore, most of these whole-genome sequences were obtained after a hybrid assembly of short and long read sequence data. This extensive, high throughput data can be effectively used to develop advanced genome-based methods for microbial systematics. Although the genomic data is available in four levels (contig, scaffold, chromosome, and complete), this study used only the complete genomes to retrieve 16S rRNA genes.

To develop species-specific barcode reference libraries, this study retrieved full-length 16S rRNA genes from 16 complete genome sequences belonging to four *Streptococcus* species (*S. gordonii*, *S. mitis*, *S. oralis,* and *S. pneumoniae*). Details of the dataset used to develop species-specific concatenated reference libraries are provided in Table 1, and the sequences are provided in the underlying data (Supplementary data 2 (Paul 2022)). Sequences were trimmed beyond the universal primer pair (fD1-5′-GAG TTT GAT CCT GGC TCA-3′ and rP2-5′-ACG GCT AAC TTG TTA CGA CT-3′, which are used for full-length 16S rDNA amplification, Weisburg *et al.* 1991) to maintain uniform length. To perform multiple sequence alignment and identify the intra-species parsimony informative (Parsim-info) variable sites, the MEGA 11 software was used. A species-specific barcode reference library that covers the entire Parsim-info variable sites was constructed by concatenating four 16S rRNA gene copies from four different strains of a species. The rationale for the selection of four copies for constructing a species-specific barcode reference library was: (i) a maximum of four variations can be found at a single site, and (ii) earlier studies have shown that the mean 16S rRNA copies per genome is four (Vetrovsky and Baldrian 2013).

Demonstration of concatenated 16S rRNA in sequence similarity and phylogeny

This study analyzed a few cases to demonstrate (i) the classical sequence similarity and (ii) phylogenetic analysis using concatenated species-specific 16S rRNA reference libraries. The study used nine 16S rRNA gene copies (sequenced using the Sanger method) showing higher sequence similarity to the 16S rRNA genes of multiple species of *Streptococcus* were retrieved from GenBank database (RRID:SCR_002760). The web-based BLAST2 (version 2.13.0) program for aligning two or more sequences was used to estimate the maximum score, total alignment score, and sequence identity of these nine 16S rRNA sequences selected. For the sequence similarity search, a single copy of the 16S rRNA (sequenced using the Sanger method or retrieved from a whole-genome assembly) can be considered as 'Query sequence'. The concatenated species-specific reference libraries need to be provided in the text area for 'Subject sequence'. However, to perform phylogenetic analysis, it is mandatory that the target sequence (length = n bp) be concatenated four times (length = 4 × n bp). Phylogenetic analysis was performed for single gene copies and concatenated approach using UPGMA method as indicated above.

Table 1. Details of whole genome assemblies used for the development of concatenated 16S rRNA reference libraries. One copy of 16S rRNA gene from each strain is used for the concatenation.

Species	Strains	Genome accession number	No. of 16S rRNA gene copies	Sequencing platform	Species-specific library name	Library length (bp)	No. of Parsim-info sites
S. gordonii	FDAARGOS 1454	CP077224.1	4	PacBio; Illumina	*S.gordonii*-Ref-I	6076	7
	NCTC7868	LR134291.1	4	PacBio			
	KCOM 1506	CP012648.1	5	Illumina			
	NCTC9124	LR594041.1	4	PacBio			
S. mitis	B6	NC_013853.1	4	NA	*S.mitis*-Ref-I	6033	10
	KCOM 1350	CP012646.1	3	Illumina			
	SVGS 061	CP014326.1	4	PacBio; Illumina			
	NCTC 12261	CP028414.1	4	PacBio			
S. oralis	NCTC 11427	LR134336.1	4	PacBio	*S.oralis*-Ref-I	6038	24
	34	CP079724.1	4	Illumina; Nanopore			
	FDAARGOS 886	CP065706.1	4	PacBio; Illumina			
	F0392	CP034442.1	4	PacBio			
S. pneumoniae	475	CP046355.1	4	PacBio	*S.pneumoniae*-Ref-I	6032	6
	NU83127	AP018936.1	4	Nanopore; Illumina			
	NCTC7465	LN831051.1	4	PacBio			
	6A-10	CP053210.1	4	PacBio			

Results

Intra-genomic 16S rRNA variations in *E. asburiae*

Historically, sequences of the 16S rRNA genes have been used to identify known and new bacterial species. However, efficiency of PCR-based amplification, poor discrimination at the species level, multiple polymorphic 16S rRNA gene copies, and improper bioinformatics workflows for the data analysis can impact the identification. The genome of *E. asburiae* contains eight copies of the 16S rRNA gene. Analysis using Clustal Omega (global alignment) and BLAST (local alignment) showed that the sequences of these eight alleles had average identities of 99.29 and 99%, respectively (Table 2). Therefore, choosing the appropriate algorithm/tool is critical for the estimation of sequence identities and sequence-based species delineation. For analyzing sequence pairs that are highly identical, global sequence alignment

Table 2. Percent identity of eight intra genomic 16S rRNA regions from *Enterobacter asburiae* strain ATCC 35953 (NZ_CP011863.1). Percent identity given below the diagonal line is calculated with Clustal Omega software (Mean identity: 99.29%) and those above the diagonal line were calculated with the BLASTN program (Mean identity: 99.00%). Genome coordinates of 16S rRNA copies: R1: 2686082–2687660 (1579 bp); R2: 3148265–3149814 (1550 bp); R3: 3313470–3315019 (1550 bp); R4: 3583942–3585481 (1540 bp); R5:3684745–3686294 (1550 bp); R6: 3771751–3773300 (1550 bp); R7: 3968538–3970087 (1550 bp); R8: 4647650–4649199 (1550 bp).

16S rRNA copies	R1	R2	R3	R4	R5	R6	R7	R8
R1		98.10	98.04	97.47	98.04	97.47	97.59	98.04
R2	99.10		99.74	99.23	99.94	99.29	99.48	99.94
R3	98.97	99.74		99.23	99.68	99.03	99.23	99.81
R4	98.90	99.41	99.41		99.16	98.52	98.71	99.29
R5	99.03	99.94	99.68	99.35		99.23	99.42	99.87
R6	98.39	99.29	99.03	98.70	99.23		99.68	99.23
R7	98.58	99.48	99.23	98.89	99.42	99.68		99.42
R8	99.03	99.94	99.81	99.48	99.87	99.23	99.42	

Genomic coordinates	400	421	422	435	551	552	561	562	835	932	933	1103	1105	1106	1107	1121	1123	1125	1315	1316	1501	1508
R1: 2686082-2687660	C	C	A	A	T	C	G	A	G	G	T	C	A	C	A	T	T	G	C	G	T	C
R2: 3148265-3149814	C	T	G	A	T	C	G	A	G	T	G	C	A	C	A	T	T	G	C	G	C	T
R3: 3313470-3315019	T	T	G	G	T	C	G	A	A	T	G	C	A	C	A	T	T	G	C	G	C	C
R4: 3583942-3585481	C	T	G	G	T	C	G	A	G	T	G	C	A	C	A	T	T	G	C	G	-	C
R5: 3684745-3686294	C	T	G	A	T	C	G	A	G	T	G	C	A	C	A	T	T	G	T	G	C	T
R6: 3771751-3773300	C	T	G	A	A	G	C	T	G	T	G	T	C	T	G	C	A	A	C	G	C	T
R7: 3968538-3970087	C	T	G	A	T	C	G	A	G	T	G	T	C	T	G	C	A	A	C	A	C	T
R8: 4647650-4649199	C	T	G	G	T	C	G	A	G	T	G	C	A	C	A	T	T	G	C	G	C	T

Figure 1. Clustal Omega based multiple sequence alignment of eight intra genomic 16S rRNA gene copies from *Enterobacter asburiae* strain ATCC 35953 (NZ_CP011863.1) showing 22 variable sites. According to Chakravorty *et al.* (2007), the nine variable regions of 16S rRNA gene spanned nucleotides 69-99, 137-242, 433-497, 576-682, 822-879, 986-1043, 1117-1173, 1243-1294, and 1435-1465 for V1 to V9 respectively.

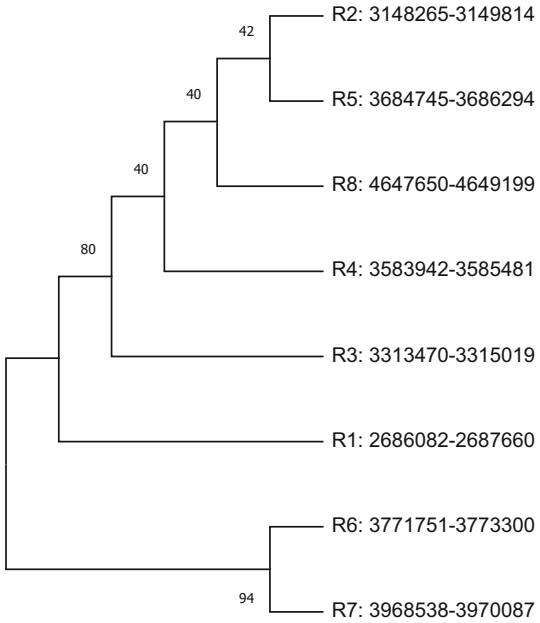


Figure 2. Phylogenetic tree of eight intra genomic 16S rRNA gene copies from *Enterobacter asburiae* strain ATCC 35953 (NZ_CP011863.1). The node label denotes the coordinate of 16S rRNA regions in the genome.

algorithms seem to be more appropriate because they consider all the nucleotides for the estimation of sequence identity. Clustal Omega based multiple sequence alignment of the eight alleles of the 16S rRNA gene in the genome of *E. asburiae* showed 22 variable sites (Figure 1). These results show that the computational analysis using a single gene copy makes different results for species harbouring variable copies of this gene.

The evolutionary relationship between species is usually represented using a phylogenetic tree based on the analysis of a single gene, multiple genes, or whole genomes. However, bacterial identification and classification is mainly based on the phylogenetic analysis of single copies of 16S rRNA genes. A phylogenetic tree was constructed to understand how variations in the sequences of the eight alleles of the 16S rRNA gene in the genome of *E. asburiae* influence species delineation (Figure 2). These results indicate that the intra-genomic variations in 16S rRNA copies may mislead the bacterial taxonomy in single gene copy approaches.

Species-specific concatenated 16S rRNA libraries

This study selected four species of *Streptococcus* (*S. gordonii*, *S. mitis*, *S. oralis*, and *S. pneumoniae*) to construct species-specific concatenated reference libraries based on 16S rRNA gene sequences obtained from complete genomes. Four variable copies of the 16S rRNA gene from a species are required to construct a species-specific concatenated reference library. The details of species-specific libraries are listed in Table 1 and the sequences are provided in the underlying data

(Supplementary data 3 (Paul 2022)). Analysis using the sequences of 16S rRNA genes showed 24, 10, 7, and 6 Parsim-info variable sites for *S. oralis*, *S. mitis*, *S. gordonii,* and *S. pneumoniae*, respectively. The intra-species Parsim-info variable sites were located in both the conserved and variable regions of the 16S rRNA gene (Supplementary data 4 (Paul 2022)).

The study used full-length 16S rRNA gene copies from four different strains to highlight the variations at the species level. However, a large number of partial 16S rRNA gene sequences are available in the public genetic databases. Further, many researchers are amplifying only few variable regions of the 16S rRNA gene. In such cases, a species-specific concatenated reference library can be constructed using partial sequences. Intra-species variations in the sequences of 16S rRNA gene copies influence the sequence-based bacterial identification. Therefore, concatenation of the sequences of 16S rRNA gene provides much better resolution compared to analysis using sequences from a single copy of the 16S rRNA gene.

Demonstration of concatenated 16S rRNA based species identification

This study compared sequences of nine 16S rRNA genes from different species of *Streptococcus* (Table 3) against the species-specific concatenated reference libraries constructed. The analysis showed that the concatenated sequences provide much better resolution in sequence similarity search and phylogenetic analysis. The sequence accession numbers GU470907.1 and KF933785.1 classified as *S. mitis* showed a higher maximum and total alignment score with concatenated 16S rRNA library of *S. oralis* than *S. mitis* (Table 3). Two sequences (OM368574.1 classified as *S. mitis* and OM368578.1 classified as *S. pneumoniae*) showed same score against the four reference libraries constructed. Based on the maximum total alignment score these two sequences are belonging to *S. pneumoniae*, however, they classified as two separate species. Interestingly, the sequence GU470907.1 classified as *S. mitis* showed 100% identity with *S. oralis* reference library with a total alignment score of 10936.

The study plotted two phylogenetic tree to highlight the difference in single gene copy approach and concatenated approach. Figure 3 represent the single gene copy approach, shows phylogenetic tree of the nine 16S rRNA gene sequences selected along with the gene copies used for the construction of four concatenated species-specific reference libraries. The inclusion of misclassified sequences and intra-species variations in 16S rRNA copies may mislead the phylogenetic tree inference. Figure 4 shows the phylogenetic relationship of nine selected sequences with four concatenated species-specific reference libraries constructed. The concatenated GU470907.1 sequence showed a phylogenetic relationship with *S. oralis* and sequence OM368574.1 was genetically related to *S. pneumoniae*. Phylogenetic analysis showed that three sequences AM157428 (*S. mitis*), KF933785 (*S. mitis*), and AM157442 (*S. pneumoniae*) stayed separately and might be other species than the four species tested. Furthermore, two sequences AJ295848 and NR_028664 classified as *S. mitis* showed significant similarity with concatenated 16S rRNA reference library of *S. mitis*. Similarly, sequence NR_117719 (*S. oralis*) showed phylogenetic relationship with reference library of *S. oralis* and OM368578 (*S. pneumoniae*) with *S. pneumoniae* reference library. These results further confirm that species-specific concatenated 16S rRNA reference libraries provide much better taxonomic resolution. Therefore, this study recommends concatenated sequences of 16S rRNA genes for sequence similarity- and phylogeny-based species identification.

Discussion

Sequencing and analysis of the 16S rRNA encoding region is a conventional and robust method for identifying and classifying bacterial species. The barcode gene is widely used in sequence similarity, phylogeny, and metagenome-based species identification. However, the accuracy of bacterial taxonomy based on 16S rRNA barcode regions is limited by the intra-genomic heterogeneity of multiple 16S rRNA gene copies and significant sequence identity of this gene among closely related taxa. Furthermore, identification of closely related species using sequences of the 16S rRNA gene is a challenge, and it may lead to species misidentification (Boudewijns *et al.* 2006; Church *et al.* 2020). About 15% of the bacterial genomes have only a single copy of the 16S rRNA gene, and only a minority of bacterial genomes contain identical 16S rRNA gene copies (Vetrovsky and Baldrian 2013). The 16S rRNA gene copies can vary from 1 to 15 in a genome, and the copy number is taxon specific (Vetrovsky and Baldrian 2013). Sequence diversity increases with the increasing 16S rRNA copy numbers. The 16S rRNA sequence variation can even be found at intra-genomic level or in different strains of a species. Amplification of a limited number of variable regions cannot achieve the same taxonomic resolution as that of the entire gene (Johnson *et al.* 2019). Usage of misclassified 16S rRNA sequences as a reference and inappropriate bioinformatics workflows can also mislead the taxonomic assignment. To overcome these challenges, it is important to translate high throughput microbial genomic data into meaningful, actionable information that clinicians can readily understand and quickly implement for bacterial identification. Hence, the study intended to develop a species-specific catalogue of concatenated 16S rRNA gene copies that can yield better inference in sequence similarity and phylogenetic analysis.

Table 3. Similarity of selected sequences against the concatenated species-specific 16S rRNA reference libraries.

GenBank Accession Number	Species	*S. gordonii*-Ref-I			*S. mitis*-Ref-I			*S. oralis*-Ref-I			*S. pneumoniae*-Ref-I		
		Max Score	Total Score	Identity (%)	Max Score	Total Score	Identity (%)	Max Score	Total Score	Identity (%)	Max Score	Total Score	Identity (%)
AJ295848.1	*S. mitis*	2495	9967	96.45	2769	11027	99.80	2758	10851	99.67	2752	10982	99.60
AM157428.1	*S. mitis*	2462	9845	96.05	2724	10866	99.27	2702	10685	99.01	2708	10805	99.07
NR_028664.1	*S. mitis*	2499	9991	96.45	2776	10979	99.87	2750	10864	99.54	2724	10888	99.27
GU470907.1	*S. mitis*	2536	10096	96.91	2715	10796	99.14	2787	10936	100	2091	10716	98.87
KF933785.1	*S. mitis*	2466	9832	96.06	2667	10593	98.54	2673	10650	98.61	2632	10502	98.15
OM368574.1	*S. mitis*	2475	9896	96.24	2754	10968	99.67	2732	10814	99.40	2760	10990	99.73
OM368578.1	*S. pneumoniae*	2475	9896	96.24	2754	10968	99.67	2732	10814	99.40	2760	10990	99.73
AM157442.1	*S. pneumoniae*	2470	9863	96.12	2702	10779	99.01	2715	10726	99.14	2702	10777	99.01
NR_117719.1	*S. oralis*	2531	10074	96.84	2710	10774	99.07	2787	10925	100	2697	10739	98.94

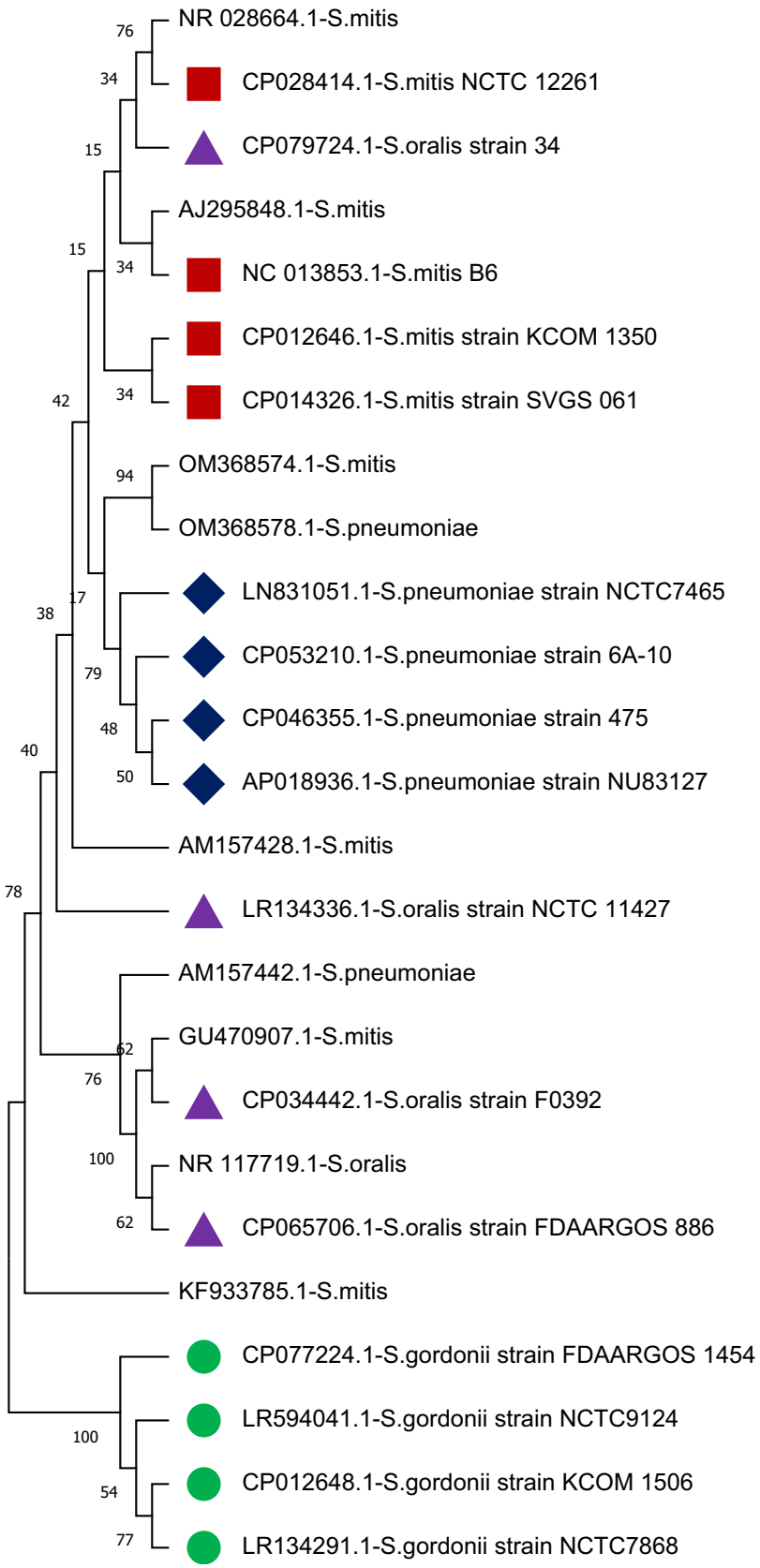


Figure 3. Phylogenetic analysis of randomly selected nine 16S rRNA sequences classified as *Streptococcus* species and sequences used for species-specific reference library. The phylogenetic tree plotted using single copy approach. The node name highlighted in shapes (◆, ●, ▲, ■) represents the sequences which are used for the construction of four concatenated species-specific reference libraries.

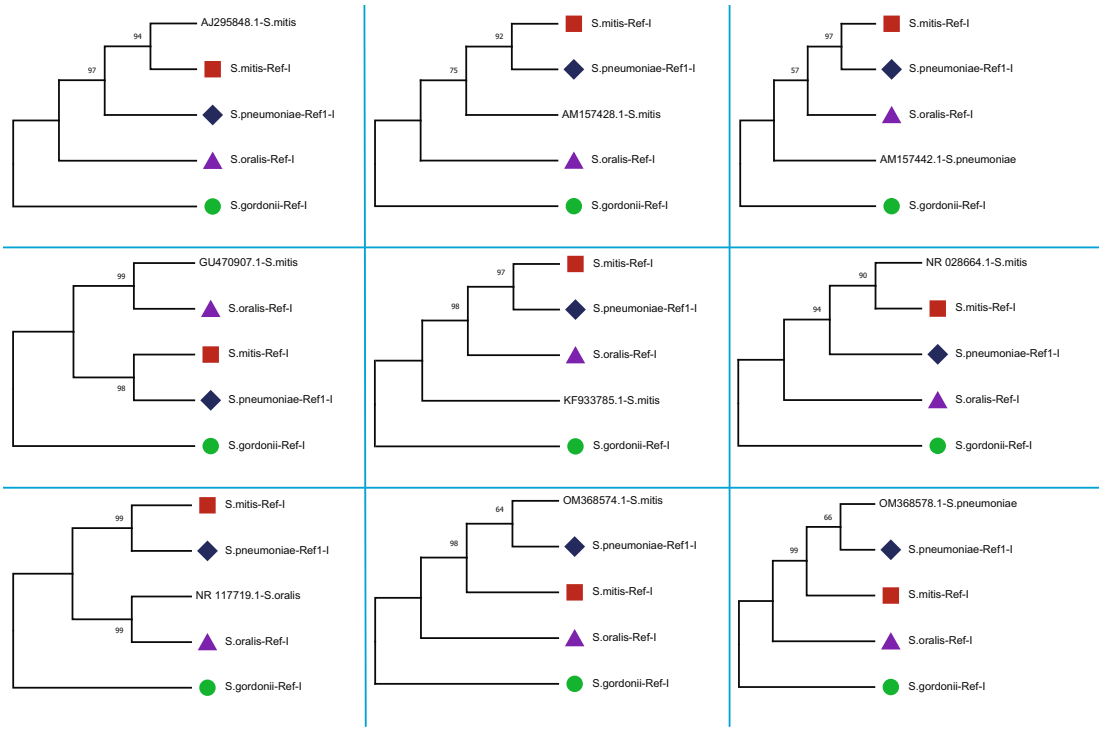


Figure 4. Phylogenetic tree constructed using concatenated 16S rRNA approach. The randomly selected nine 16S rRNA sequences classified as *Streptococcus* species were compared with four species-specific reference libraries constructed. The node name highlighted in shapes (◆, ●, ▲, ■) represents the four species-specific reference libraries.

Several bioinformatics resources are extensively used for the 16S rRNA sequence analysis and bacterial identification. However, several researchers report the sequence similarity derived through a local alignment algorithm. Earlier reports have suggested that the species belonging to the taxa Gammaproteobacteria show higher intra-species variability (Vetrovsky and Baldrian 2013). Hence, the study estimated the percent identity of intra-genomic 16S rRNA gene copies of *E. asburiae* using local and global alignment algorithms. The reference genome of *E. asburiae* has eight 16S rRNA gene copies in its genome. The BLAST and Clustal sequence alignment algorithms yielded marginally varying results for the intra-genomic 16S rRNA gene copies. Local alignment algorithms may not consider base mismatches at the ends of sequences when calculating percent identity, while global alignment algorithms consider entire sequences. Therefore, global sequence alignment is best for estimating intra and inter-species identity for single gene copies. However, BLAST can calculate the total alignment score with multiple paralogue regions. Hence, web-based BLAST2 is suggested for estimating the sequence similarity using concatenated barcode reference libraries.

The GenBank (Leray *et al.* 2019) and NCBI 16S RefSeq databases for bacteria (Winand *et al.* 2020) are reliable for species-level identification and classification. However, few earlier studies have highlighted the misclassification of species and genome assemblies in public genetic databases (Parks *et al.* 2018; Varghese *et al.* 2015). For example, the 16S rRNA sequence accession number (Ac. No.) LT707617.1 shows the organism as *Streptococcus mitis.* Conventional BLAST-based sequence similarity search shows the highest identity of 99.60% with *S. mitis* 16S rRNA sequence (Ac. No. AB002520.1). However, the 16S rRNA sequence (Ac. No. LT707617.1) did not show significant similarity with other 16S rRNA reference sequences available for *S. mitis.* Furthermore, the sequence also shows 99.44% identity with reference 16S rRNA sequences of *S. gordonii.* Hence, the study performed a sequence alignment of the sequence (Acc. No. LT707617.1) against species-specific concatenated 16S rRNA reference libraries for *S. gordonii* (*S.gordonii*-Ref-I), and *S. mitis* (*S.mitis*-Ref-I). The alignment resulted in a significant identity of 99.44% with *S.gordonii*-Ref-I (2279 maximum and 9041 total alignment score) than *S.mitis*-Ref-I (97.13% identity with 2119 maximum and 8449 total alignment score). Single copy BLAST results may show only a minor fraction of the difference in percent identity and maximum or total alignment score for closely related species. However, sequence similarity estimation using species-specific concatenated reference libraries shows marginal difference in total alignment score, as it is aligned against four copies. Hence, 16S rRNA analysis with a species-specific concatenated barcode reference library will give better accuracy for bacterial classification than approaches using a single copy.

Several 16S rRNA sequences show 100% identity with multiple species, which is the major challenge in sequence-based species identification. For example, the 16S rRNA sequence from *S. mitis* (Accession. No. GU470907.1; 1522 bp) shares 100% identity with the 16S rRNA gene from *S. oralis* strain ATCC 35037 genome (Ac. No. CP034442.1). Hence, the sequence (GU470907.1) aligned against the species-specific concatenated reference libraries for *S. oralis* (*S.oralis*-Ref-I), and *S. mitis* (*S.mitis*-Ref-I). The result showed 100% identity with *S. oralis* (2787 maximum and 10936 total alignment score), and 99.14% identity with *S. mitis* (2715 maximum and 10796 total alignment score). Further, a phylogenetic tree of GU470907.1 (1509 × 4 = 6036 bp) with reference libraries *S.mitis*-Ref-I, and *S.oralis*-Ref-I was plotted. The UPGMA-based phylogenetic tree showed that the *S. mitis* (GU470907.1) sequence is more closely related to *S. oralis* than *S. mitis* (Figure 4). Concatenated 16S rRNA-based estimation of sequence similarity and a phylogenetic inference provides better resolution than single-gene approaches. These results show that the concatenated 16S rRNA approach is very effective in discriminating genetically closely related bacterial species. Furthermore, other studies have also highlighted that the phylogenetic tree inferred from vertically inherited protein sequence concatenation provided higher resolution than those obtained from a single copy (Ciccarelli *et al.* 2006; Thiergart *et al.* 2014).

Recent phylogenetic studies using concatenated multi-gene sequence data highlighted the importance of incorporating variations in gene histories, which will improve the traditional phylogenetic inferences (Devulder *et al.* 2005; Johnston *et al.* 2019). Furthermore, a single type of analysis should not be relied upon, instead, and to a certain extent, integrated bioinformatics approaches can avoid misclassification. As a cost-effective approach, the study combined substantial variations in 16S rRNA gene copies from a species to examine the performance of the single gene concatenation approach. Analyses using a concatenated 16S rRNA gene approach have the following advantages: (i) the gene is present in all the bacterial species, (ii) the gene is weakly affected by horizontal gene transfer and mutation, (iii) the approach is very cost-effective, (iv) there is a large volume of reference genomic data available for several bacterial species, (v) it is effective in discriminating closely related bacterial species, (vi) the analyses can be performed in a computer with minimum configuration, and (vii) the analyses can be employed with available tools for sequence similarity and molecular phylogeny.

Conclusions

The concatenated 16S rRNA analyses showed that:

- Full-length 16S rRNA gene amplification provides better accuracy than inference based on partial gene sequences with a limited number of variable regions.
- Full-length 16S rRNA gene copies from whole-genome assemblies (in 'complete' stage) should be used rather than partial sequences available from the public genetic databases to construct species-specific concatenated 16S rRNA libraries and further downstream analysis.
- To avoid mismatches in the sequence alignment, trim the bases beyond the primer ends and correct the base-call errors prior to the analysis.
- Estimation of mean 16S rRNA identity at the intra-species level helps to classify the species having a higher degree of intra-genomic 16S rRNA heterogeneity.
- Four distinct 16S rRNA gene copies cover all the Parsim-Info variable sites and these can be used to construct a concatenated species-specific reference library.
- The total alignment score can be considered if the query sequence shows more or less the same percent identity with multiple species.
- It is not prudent to rely only on sequence similarity; the final decision must be based on the phylogenetic inference.
- Species-specific concatenated 16S rRNA gene libraries are recommended for sequence similarity and phylogenetic analysis.
- The limitation of the approach is that developing a species-specific reference library requires 16S rRNA copies from at least four whole genome assemblies.

Data availability

Underlying data

Zenodo: Underlying data for 'Concatenated 16S rRNA sequence analysis improves bacterial taxonomy', https://doi.org/10.5281/zenodo.7758747 (Paul 2022).

This project contains the following underlying data:

- Supplementary data 1: The 16S rRNA copies retrieved from the whole genome of *Enterobacter asburiae* strain ATCC 35953.
- Supplementary data 2: Full-length 16S rRNA gene copies retrieved from 16 genome assemblies belonging to four *Streptococcus* species (*S. gordonii*, *S. mitis*, *S. oralis*, and *S. pneumoniae*).
- Supplementary data 3: Species-specific concatenated 16S rRNA libraries constructed for four *Streptococcus* species (*S. gordonii*, *S. mitis*, *S. oralis*, and *S. pneumoniae*).
- Supplementary data 4: Intra-species Parsim-info variable sites in the 16S rRNA gene from for four *Streptococcus* species (*S. gordonii*, *S. mitis*, *S. oralis*, and *S. pneumoniae*).

Accession numbers

GenBank: *Streptococcus gordonii* strain FDAARGOS 1454 chromosome, complete genome. Accession number CP077224.1. https://www.ncbi.nlm.nih.gov/nuccore/CP077224.1

GenBank: *Streptococcus gordonii* strain NCTC7869, chromosome 1, complete genome. Accession number LR134291.1. https://www.ncbi.nlm.nih.gov/nuccore/LR134291.1

GenBank: *Streptococcus gordonii* strain KCOM 1506 (=ChDC B679), complete genome. Accession number CP012648.1. https://www.ncbi.nlm.nih.gov/nuccore/CP012648.1

GenBank: *Streptococcus gordonii* strain NCTC9124, chromosome 1, complete genome. Accession number LR594041.1. https://www.ncbi.nlm.nih.gov/nuccore/LR594041.1

GenBank: *Streptococcus mitis* B6, complete genome. Accession number NC_013853.1. https://www.ncbi.nlm.nih.gov/nuccore/NC_013853.1

GenBank: *Streptococcus mitis* strain KCOM 1350 (= ChDC B183), complete genome. Accession number CP012646.1. https://www.ncbi.nlm.nih.gov/nuccore/CP012646.1

GenBank: *Streptococcus mitis* strain SVGS_061 chromosome, complete genome. Accession number CP014326.1. https://www.ncbi.nlm.nih.gov/nuccore/CP014326.1

GenBank: *Streptococcus mitis* NCTC 12261 chromosome, complete genome. Accession number CP028414.1. https://www.ncbi.nlm.nih.gov/nuccore/CP028414.1

GenBank: *Streptococcus oralis* strain NCTC11427, chromosome 1, complete genome. Accession number LR134336.1. https://www.ncbi.nlm.nih.gov/nuccore/LR134336.1

GenBank: *Streptococcus oralis* strain 34 chromosome, complete genome. Accession number CP079724.1. https://www.ncbi.nlm.nih.gov/nuccore/CP079724.1

GenBank: *Streptococcus oralis* strain FDAARGOS_886 chromosome, complete genome. Accession number CP065706.1. https://www.ncbi.nlm.nih.gov/nuccore/CP065706.1

GenBank: *Streptococcus oralis* subsp. *dentisani* strain F0392 chromosome, complete genome. Accession number CP034442.1. https://www.ncbi.nlm.nih.gov/nuccore/CP034442.1

GenBank: *Streptococcus pneumoniae* strain 475 chromosome, complete genome. Accession number CP046355.1. https://www.ncbi.nlm.nih.gov/nuccore/CP046355.1

GenBank: *Streptococcus pneumoniae* NU83127 DNA, complete genome. Accession number AP018936.1. https://www.ncbi.nlm.nih.gov/nuccore/AP018936.1

GenBank: *Streptococcus pneumoniae* NCTC7465, chromosome 1, complete genome. Accession number LN831051.1. https://www.ncbi.nlm.nih.gov/nuccore/LN831051.1

GenBank: *Streptococcus pneumoniae* strain 6A-10 chromosome, complete genome. Accession number CP053210.1. https://www.ncbi.nlm.nih.gov/nuccore/CP053210.1

GenBank: *Streptococcus mitis* strain 127R, partial 16S rRNA gene. Accession number AJ295848.1. https://www.ncbi.nlm.nih.gov/nuccore/AJ295848.1

GenBank: *Streptococcus mitis* clone 2C4, 16S rRNA gene. Accession number AM157428.1. https://www.ncbi.nlm.nih.gov/nuccore/AM157428.1

GenBank: *Streptococcus mitis* strain NS51, partial 16S rRNA gene. Accession number NR_028664.1. https://www.ncbi.nlm.nih.gov/nuccore/NR_028664.1

GenBank: *Streptococcus mitis* bv. 2 strain F0392, partial 16S rRNA gene. Accession number GU470907.1. https://www.ncbi.nlm.nih.gov/nuccore/GU470907.1

GenBank: *Streptococcus mitis* strain ChDC B553, partial 16S rRNA gene. Accession number KF933785. https://www.ncbi.nlm.nih.gov/nuccore/KF933785.1

GenBank: *Streptococcus mitis* strain FC6528, partial 16S rRNA gene. Accession number OM368574.1. https://www.ncbi.nlm.nih.gov/nuccore/OM368574.1

GenBank: *Streptococcus pneumoniae* strain FC6532, partial 16S rRNA gene. Accession number OM368578.1. https://www.ncbi.nlm.nih.gov/nuccore/OM368578.1

GenBank: *Streptococcus pneumoniae* clone 4V4, 16S rRNA gene. Accession number AM157442. https://www.ncbi.nlm.nih.gov/nuccore/AM157442.1

GenBank: *Streptococcus oralis* subsp. *dentisani* strain 7747, partial 16S rRNA gene. Accession number NR_117719. https://www.ncbi.nlm.nih.gov/nuccore/NR_117719.1

GenBank: *Enterobacter asburiae* strain ATCC 35953 chromosome, complete genome. Accession number NZ_CP011863. https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP011863.1

GenBank: *Streptococcus mitis* strain HAC11, isolate #11, partial 16S rRNA gene. Accession number LT707617. https://www.ncbi.nlm.nih.gov/nuccore/LT707617.1

GenBank: *Streptococcus mitis* strain NCTC 3165, MAFF 911479, 16S rRNA gene. Accession number AB002520.1. https://www.ncbi.nlm.nih.gov/nuccore/AB002520.1

Acknowledgements

The author would like to thank DST-FIST, the Government of India, TIFAC-CORE in Pharmacogenomics and Manipal Academy of Higher Education (MAHE), Manipal for the support and facilities provided.

References

Alachiotis N, Vogiatzi E, Pavlidis P, *et al.*: **Chromatogate: a tool for detecting base mis-calls in multiple sequence alignments by semi-automatic chromatogram inspection.** *Comput. Struct. Biotechnol. J.* 2013; **6**: e201303001.
PubMed Abstract | Publisher Full Text | Free Full Text

Altschul SF, Gish W, Miller W, *et al.*: **Basic local alignment search tool.** *J. Mol. Biol.* 1990; **215**: 403–410.
Publisher Full Text

Bagheri H, Severin AJ, Rajan H: **Detecting and correcting misclassified sequences in the large-scale public databases.** *Bioinformatics.* 2020; **36**: 4699–4705.
PubMed Abstract | Publisher Full Text | Free Full Text

Baltrus DA: **Divorcing strain classification from species names.** *Trends Microbiol.* 2016; **24**: 431–439.
Publisher Full Text

Benitez-Paez A, Sanz Y: **Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinIONTM portable Nanopore sequencer.** *Gigascience.* 2017; **6**: 1–12.
PubMed Abstract | Publisher Full Text | Free Full Text

Boudewijns M, Bakkers JM, Sturm PDJ, *et al.*: **16S rRNA gene sequencing and the routine clinical microbiology laboratory: A perfect marriage?** *J. Clin. Microbiol.* 2006; **44**: 3469–3470.
PubMed Abstract | Publisher Full Text | Free Full Text

Caporaso JG, Lauber CL, Walters WA, *et al.*: **Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample.** *Proc. Natl. Acad. Sci. U S A.* 2011; **108**: 4516–4522.
Publisher Full Text

Chakravorty S, Helb D, Burday M, *et al.*: **A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria.** *J. Microbiol. Methods.* 2007; **69**: 330–339.
Publisher Full Text

Church DL, Cerutti L, Gürtler A, *et al.*: **Performance and application of 16S rRNA gene cycle sequencing for routine identification of bacteria in the clinical microbiology laboratory.** *Clin. Microbiol. Rev.* 2020; **33**: e00053–e00019.
PubMed Abstract | Publisher Full Text | Free Full Text

Ciccarelli FD, Doerks T, von Mering C, *et al.*: **Toward automatic reconstruction of a highly resolved tree of life.** *Science.* 2006; **311**: 1283–1287.
Publisher Full Text

Clarridge JE: **Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases.** *Clin. Microbiol. Rev.* 2004; **17**: 840–862.
PubMed Abstract | Publisher Full Text | Free Full Text

Deurenberg RH, Bathoorn E, Chlebowicz MA, *et al.*: **Application of next generation sequencing in clinical microbiology and infection prevention.** *J. Biotechnol.* 2017; **243**: 16–24.
PubMed Abstract | Publisher Full Text

Devanga-Ragupathi NK, Muthuirulandi SDP, Inbanathan FY, *et al.*: **Accurate differentiation of *Escherichia coli* and *Shigella* serogroups: challenges and strategies.** *New Microbes New Infect.* 2018; **21**: 58–62.
PubMed Abstract | Publisher Full Text | Free Full Text

Devulder G, de Montclos MP, Flandrois JP: **A multigene approach to phylogenetic analysis using the genus *Mycobacterium* as a model.** *Int. J. Syst. Evol. Microbiol.* 2005; **55**: 293–302.
Publisher Full Text

Ibal JC, Pham HQ, Park CE, *et al.*: **Information about variations in multiple copies of bacterial 16S rRNA genes may aid in species identification.** *PLoS One.* 2019; **14**: e0212090.
PubMed Abstract | Publisher Full Text | Free Full Text

Janda JM, Abbott SL: **16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls.** *J. Clin. Microbiol.* 2007; **45**: 2761–2764.
PubMed Abstract | Publisher Full Text | Free Full Text

Johnson JS, Spakowicz DJ, Hong BY, *et al.*: **Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis.** *Nat. Commun.* 2019; **10**: 5011–5029.
PubMed Abstract | Publisher Full Text | Free Full Text

Johnston PR, Quijada L, Smith CA, *et al.*: **A multigene phylogeny toward a new phylogenetic classification of *Leotiomycetes*.** *IMA Fungus.* 2019; **10**: 1.
PubMed Abstract | Publisher Full Text | Free Full Text

Kerkhof LJ, Dillon KP, Haggblom MM, *et al.*: **Profiling bacterial communities by MinION sequencing of ribosomal operons.** *Microbiome.* 2017; **5**: 116.
PubMed Abstract | Publisher Full Text | Free Full Text

Kumar S, Stecher G, Li M, *et al.*: **MEGA X: Molecular evolutionary genetics analysis across computing platforms.** *Mol. Biol. Evol.* 2018; **35**: 1547–1549.
PubMed Abstract | Publisher Full Text | Free Full Text

Lal D, Verma M, Lal R: **Exploring internal features of 16S rRNA gene for identification of clinically relevant species of the genus *Streptococcus.*** *Ann. Clin. Microbiol. Antimicrob.* 2011; **10**: 28.
PubMed Abstract | Publisher Full Text | Free Full Text

Leray M, Knowlton N, Ho SL, *et al.*: **GenBank is a reliable resource for 21st century biodiversity research.** *Proc. Natl. Acad. Sci. USA.* 2019; **116**: 22651–22656.
PubMed Abstract | Publisher Full Text | Free Full Text

Liu Y, Lai Q, Shao Z: **Genome analysis-based reclassification of *Bacillus weihenstephanensis* as a later heterotypic synonym of *Bacillus mycoides*.** *Int. J. Syst. Evol. Microbiol.* 2018; **68**: 106–112.
PubMed Abstract | Publisher Full Text

Martínez-Romero E, Rodríguez-Medina N, Beltrán-Rojel M, *et al.*: **Genome misclassification of *Klebsiella variicola* and *Klebsiella quasipneumoniae* isolated from plants, animals and humans.** *Salud Publica Mex.* 2018; **60**: 56–62.
PubMed Abstract | Publisher Full Text

Mateo-Estrada V, Grana-Miraglia L, Lopez-Leal G, *et al.*: **Phylogenomics reveals clear cases of misclassification and genus-wide phylogenetic markers for *Acinetobacter*.** *Genome Biol. Evol.* 2019; **11**: 2531–2541.
PubMed Abstract | Publisher Full Text | Free Full Text

Parks DH, Waite DW, Skarshewski A, *et al.*: **A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.** *Nat. Biotechnol.* 2018; **36**: 996–1004.
PubMed Abstract | Publisher Full Text

Paul B, Dixi G, Murali TS, *et al.*: **Genome-based taxonomic classification.** *Genome.* 2019; **62**: 45–52.
Publisher Full Text

Paul B: **Concatenated 16S rRNA sequence analysis improves bacterial taxonomy.** 2022.
Publisher Full Text

Peker N, Garcia-Croes S, Dijkhuizen B, *et al.*: **A comparison of three different bioinformatics analyses of the 16S-23S rRNA encoding region for bacterial identification.** *Front. Microbiol.* 2019; **10**: 620.
PubMed Abstract | Publisher Full Text | Free Full Text

Quast C, Pruesse E, Yilmaz P, *et al.*: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.** *Nucleic Acids Res.* 2013; **41**: D590–D596.
PubMed Abstract | Publisher Full Text | Free Full Text

Reller LB, Weinstein MP, Petti CA: **Detection and identification of microorganisms by gene amplification and sequencing.** *Clin. Infect. Dis.* 2007; **44**: 1108–1114.
Publisher Full Text

Sabat AJ, van Zanten E, Akkerboom V, *et al.*: **Targeted next-generation sequencing of the 16S-23S rRNA region for culture-independent bacterial identification increased discrimination of closely related species.** *Sci. Rep.* 2017; **7**: 1–12.
PubMed Abstract | Publisher Full Text | Free Full Text

Schloss PD: **Reintroducing mothur: 10 Years Later.** *Appl. Environ. Microbiol.* 2020; **86**: e02343–e02319.
PubMed Abstract | Publisher Full Text | Free Full Text

Sievers F, Wilm A, Dineen D, *et al.*: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.** *Mol. Syst. Biol.* 2011; **7**: 539.
Publisher Full Text

Srinivasan R, Karaoz U, Volegova M, *et al.*: **Use of 16S rRNA gene for identification of a broad range of clinically relevant bacterial pathogens.** *PLoS One.* 2015; **10**: e0117617.
PubMed Abstract | Publisher Full Text | Free Full Text

Stackebrandt E, Mondotte JA, Fazio LL, *et al.*: **Authors need to be prudent when assigning names to microbial isolates.** *Arch. Microbiol.* 2021; **203**: 5845–5848.
PubMed Abstract | Publisher Full Text

Starke R, Pylro VS, Morais DK: **16S rRNA gene copy number normalization does not provide more reliable conclusions in metataxonomic surveys.** *Microb. Ecol.* 2021; **81**: 535–539.
PubMed Abstract | Publisher Full Text | Free Full Text

Steven B, Hesse C, Soghigian J, *et al.*: **Simulated rRNA/DNA ratios show potential to misclassify active populations as dormant.** *Appl. Environ. Microbiol.* 2017; **83**: e00696–e00617.
PubMed Abstract | Publisher Full Text | Free Full Text

Thiergart T, Landan G, Martin WF: **Concatenated alignments and the case of the disappearing tree.** *BMC Evol. Biol.* 2014; **14**: 212–266.
PubMed Abstract | Publisher Full Text | Free Full Text

Varghese NJ, Mukherjee S, Ivanova N, *et al.*: **Microbial species delineation using whole genome sequences.** *Nucleic Acids Res.* 2015; **43**: 6761–6771.
PubMed Abstract | Publisher Full Text | Free Full Text

Vetrovsky T, Baldrian P: **The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses.**

PLoS One. 2013; **8**: e57923.
PubMed Abstract | Publisher Full Text | Free Full Text

Weisburg WG, Barns SM, Pelletier DA, *et al.*: **16S ribosomal DNA amplification for phylogenetic study.** *J. Bacteriol.* 1991; **173**: 697–703.
PubMed Abstract | Publisher Full Text | Free Full Text

Winand R, Bogaerts B, Hoffman S, *et al.*: **Targeting the 16S rRNA gene for bacterial identification in complex mixed samples: Comparative evaluation of second (Illumina) and third (Oxford Nanopore technologies) generation sequencing technologies.** *Int. J. Mol. Sci.* 2020; **21**: 298.
PubMed Abstract | Publisher Full Text | Free Full Text

Yoon SH, Ha SM, Kwon S, *et al.*: **Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies.** *Int. J. Syst. Evol. Microbiol.* 2017; **67**: 1613–1617.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 3

Reviewer Report 08 September 2023

https://doi.org/10.5256/f1000research.155249.r203269

✔ **Siddaramappa Shivakumara**
Institute of Bioinformatics and Applied Biotechnology, Bengaluru, Karnataka, India

The revised version [version 3] addressed the minor concerns and is further improved. I approve the same for indexing.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bacterial Genomics, Bacterial Taxonomy, Comparative Genomics, Pathogenomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 08 September 2023

https://doi.org/10.5256/f1000research.155249.r203270

✔ **Wellyzar Sjamsuridzal**
Department of Biology, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, West Java, Indonesia

Dear Authors,

I have read the revised version 3 of the manuscript and I found all of my concerns have been responded to. Therefore, I approve this version for indexing.

Thank you.
Wellyzar S.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Microbial Systematics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 21 August 2023

https://doi.org/10.5256/f1000research.144651.r189677

?

Wellyzar Sjamsuridzal
[1] Department of Biology, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, West Java, Indonesia
[2] Department of Biology, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, West Java, Indonesia

Are the conclusions drawn adequately supported by the results?

Partly: The conclusion should elaborate more by considering the limitation of the study. Developing a species-specific reference library for all bacteria using concatenated 16S rRNA gene copy variants is a challenging task.

Here are a few considerations regarding its feasibility for species identification:

Concatenating 16S rRNA gene copy variants for sequence similarity estimation and phylogenetic inference can offer improved resolution compared to single-gene approaches. However, careful design, data interpretation, and consideration of alternative approaches are still required to account for potential biases and supplementary analyses when necessary.

1. The 16S rRNA gene is commonly used for bacterial identification, but it may not capture the full diversity of all bacterial species. While it can be effective for many bacteria, there are certain groups that may have unique or divergent sequences not adequately represented by the 16S rRNA gene.

2. Some bacterial species can have significant intraspecies genetic variation, including

variations in the 16S rRNA gene. Concatenating multiple gene copy variants can help account for this variability to a certain degree, but it may not capture all the genetic diversity within a species.

3. The 16S rRNA gene is suitable for identifying bacterial species at a broader taxonomic level. However, for more precise differentiation or identification, additional genetic markers or techniques may be required, such as whole-genome sequencing or multilocus sequence typing.

4. Concatenating multiple gene copy variants from a large number of bacteria would result in a substantial increase in the complexity and size of the reference library. Managing and interpreting such a comprehensive library would require significant computational resources and expertise.

In summary, while concatenating 16S rRNA gene copy variants can be a valuable approach for bacterial identification, developing a species-specific reference library for all bacteria using this method may have limitations. It is important to consider the diversity and variability within bacterial species and explore additional techniques for more precise identification when necessary.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Microbial Systematics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 28 Aug 2023

Bobby Paul

Thank you for the critical comments. I have included the limitation of this approach in the conclusion section, and the manuscript has been updated. The study used four genetically related species for the demonstration. Concatenated 16S rRNA analysis yielded better resolution than a single copy. Yes, developing concatenated species-specific 16S rRNA reference libraries for entire bacterial populations is challenging. The study developed reference libraries for four species. More than 552,575 whole-genome sequences are currently (Aug 2023) available for bacterial species in the Genome database. Hopefully, these datasets can be used for the development of species-specific reference libraries for remaining bacterial species. Further, we are in the process of developing reference libraries for other bacterial species and hosting a web server for easy analysis.
Here is the response to the specific queries.

1. The gene encoding the 16S rRNA is ~1500 base pair (bp) long, present in all bacterial populations, and the gene consists of nine variable regions. In sequence analysis, the variable regions are differentiating species. Hence, the proposed approach is applicable to delineate all the bacterial populations. However, as it is ~1500 bp, sequencing of this gene is not sufficient to estimate the genome-wide difference. Estimations such as average nucleotide identity (ANI), Genome BLAST Distance Phylogeny approach can be applied to whole genome sequences.

2. Genetic differences can be found between strains within a species, as well as between intra-genomic 16S rRNA gene copies. The approach proposed in this manuscript is exclusively to avoid misclassification due to genetic variations in the intra-species 16S rRNA gene copies. The 16S rRNA sequence analysis is not sufficient to estimate the entire genetic difference. Several whole genome-based approaches exist for genome-wide diversity estimation and bacterial taxonomy. The 16S rRNA sequencing is considered a cost-effective, gold-standard single gene approach for species identification. Hence, the method is more appropriate for bacterial species identification. The rate of mutations and lateral gene transfer is high in bacterial populations. However, 16S rRNA is less affected by these genetic changes. Further, factors such as genome size, genetic rearrangements, and unique genomic regions can slightly affect whole genome based species identification. However, concatenated 16S rRNA analysis won't be affected by these factors.

3. Approaches to bacterial taxonomy have improved from microscopic examination to genome-based methods over the years. However, few earlier studies have highlighted the misclassification of species and genome assemblies in public genetic databases (Parks et al. 2018; Varghese et al. 2015). The manuscript explains the proof of concept. Whether concatenation of multiple copies of 16S rRNA can improve bacterial taxonomy. The study used four closely related *Streptococcus* species (*S. gordonii*, *S. mitis, S. oralis*, and *S. pneumoniae*), which is difficult to differentiate with current genome based approaches for bacterial taxonomy. The results show the proposed concatenated 16S rRNA analysis can easily differentiate bacterial species, even sharing high genetic similarity. Although, estimations such as average nucleotide identity (ANI), genome-to-genome distance calculation, Genome BLAST Distance Phylogeny approach can be applied to whole genome sequences.

4. Thank you. The study was completed using a laptop that has a 4GB of RAM. Whole genome sequence data available in the public genetic databases can be used to develop reference libraries for remaining bacterial species. The analyses can be employed with currently available tools for sequence similarity and molecular phylogeny. Therefore, interpreting results should not pose a challenge for researchers. However, we are in the process of hosting a web server for more easy and appropriate analysis of global bacterial populations.

Kindly respond if you have any other queries or concerns.

Kind Regards
Bobby Paul

Competing Interests: Nil

Reviewer Report 11 April 2023

https://doi.org/10.5256/f1000research.144651.r168593

Siddaramappa Shivakumara
[1] Institute of Bioinformatics and Applied Biotechnology, Bengaluru, Karnataka, India
[2] Institute of Bioinformatics and Applied Biotechnology, Bengaluru, Karnataka, India

The author has addressed most of my technical and non-technical (e.g., writing style and presentation) concerns. Version 2 is considerably improved compared to Version 1 in terms of readability, flow, and impact. However, I still believe there is scope to further improve the Discussion section. Nevertheless, I approve Version 2 without any further reservations.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bacterial Genomics, Bacterial Taxonomy, Comparative Genomics, Pathogenomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 10 February 2023

https://doi.org/10.5256/f1000research.140896.r158444

?

Siddaramappa Shivakumara
[1] Institute of Bioinformatics and Applied Biotechnology, Bengaluru, Karnataka, India
[2] Institute of Bioinformatics and Applied Biotechnology, Bengaluru, Karnataka, India
[3] Institute of Bioinformatics and Applied Biotechnology, Bengaluru, Karnataka, India

The manuscript entitled “Concatenated 16S rRNA sequence analysis improves bacterial taxonomy [version 1]” by Bobby Paul is generally well written and reports interesting results. The methods are appropriate and the analyses meet the quality standards.

However, there are some concerns about the clarity and presentation of results. Major revisions in the text are required to improve clarity and comprehensibility. My suggestions/critiques for revising the text in different sections of the manuscript are provided. Author’s attention has also been drawn to technical concerns, which need to be carefully addressed.

The Discussion section is rather lengthy; it should be shortened and include only relevant results/studies (present and previous). I would be happy to provide detailed critiques on the Discussion section after the manuscript goes through the first round of revision.

Is the work clearly and accurately presented and does it cite the current literature?
Some statements/results are missing citations and references. These have been highlighted and need to be revised.

Are the conclusions drawn adequately supported by the results?
Some conclusions are unsubstantiated, and require relevant data and proper interpretation. These have been highlighted and need to be revised.

Introduction

"The 16S ribosomal RNA (16S rRNA) encoding region is extensively studied to identify and classify bacterial species. The 16S rRNA is a conserved component of the 30S small subunit of a prokaryotic ribosome"
Please revise this as
"The DNA region encoding the 16S ribosomal RNA (16S rRNA) is extensively studied, and used to identify and classify bacterial species. The 16S rRNA is a conserved component of the small subunit (30S) of the prokaryotic ribosome".

The gene is ~1500 base pair (bp) long, and it consists of nine variable regions.
Please revise this as
"The gene encoding the 16S rRNA is base pair (bp) long, and it consists of nine variable regions"

"For decades, the sequence of the 16S rRNA gene has been used as a potential molecular marker in culture-independent methods to identify and classify diverse bacterial communities (Clarridge, 2004; Johnson *et al*. 2019)"
Please revise this as
"The sequence of the 16S rRNA gene has been extensively used as a molecular marker in culture-independent methods to identify and classify diverse bacterial communities"

"The 16S rRNA sequences are currently being used as an accurate and rapid method to study bacterial evolution, phylogenetic relationships, populations in an environment, and quantification of abundant taxa"
Please revise this as
"Bacterial 16S rRNA sequences are currently being used to study the evolution, phylogenetic relationships, and environmental abundanceof various taxa".

"Despite the wide range of applications, a few shortcomings limit the accuracy of results derived through the 16S rRNA sequence analysis. One such aspect is that the 16S rRNA gene has poor discriminatory power at the species level (Winand *et al*. 2020), and the copy number can vary from 1 to 15 or even more (Vetrovsky and Baldrian, 2013; Winand *et al*. 2020)."
Please revise this as
"Although 16S rRNA sequence analyses are the mainstay of taxonomic studies of bacteria, there are some limitations. For example, the 16S rRNA gene has poor discriminatory power at the species level (Winand *et al*. 2020), and the copy number is highly variable (Vetrovsky and Baldrian, 2013; Winand *et al*. 2020). ".

"The presence of multiple variable copies of this gene makes distinct data for a species."
The above statement is a little confusing, and should be revised. Their presence in itself may not "make distinct data".

"Hence, gene copy normalization (GCN) is necessary prior to sequence analysis."

Please revise this as
"Therefore, gene copy normalization (GCN) may be necessary prior to sequence analysis.

"However, studies show that the GCN approach does not improve the 16S rRNA sequence analyses in real scenarios and suggests a comprehensive species-specific catalogue of gene copies (Starke *et al*. 2021)".
Please revise this as
"However, GCN may not improve the 16S rRNA sequence analyses in all scenarios, and comprehensive, species-specific catalogues of 16S rRNA gene copies may be necessary (Starke *et al*. 2021).

"Secondly, the intra-genomic variations between the 16S rRNA gene copies were observed in several bacterial genome assemblies (Paul *et al*. 2019)."
Please revise this as
"Furthermore, intra-species variations in the number of copies of the16S rRNA gene were observed in several bacterial genome assemblies"

"Only a minority of the bacterial genomes harbor identical 16S rRNA gene copies, and sequence diversity increases with increasing copy numbers (Vetrovsky and Baldrian, 2013). "
Please revise this as
"Only a few bacterial species contain identical 16S rRNA gene copies, and sequence diversity increases with increasing copy numbers of 16S rRNA genes".

"Further, currently available 16S rRNA-based bioinformatics approaches are not always amenable to classify bacterium at the species level due to high inter-species sequence similarities (Peker *et al* . 2019; Deurenberg *et al*. 2017)."
Please revise this as
"The high levels of similiarty of the 16S rRNA genes across some bacterial species poses a major challege for taxonomic studies using bioinformatics methods".

“A few other issues are also related to the sequencing and bioinformatics analysis of 16S rRNA gene regions. These include the purity of bacterial isolates, the quality of isolated DNA, and the possibility of chimeric molecules (Janda and Abbott, 2007; Church *et al*. 2020)”
Please revise this as
Factors such as purity of bacterial cultures, quality of the purified DNA samples, and potential DNA chimeras should be carefully considered while sequencing and analysis of 16S rRNA genes.

“Base-call errors can also mislead the sequence identity and phylogenetic inferences (Alachiotis *et al*. 2013).
Please revise this as
“Sequencing errors can lead to misidentification of bacteria and phylogenetic anomalies”.

“The other concerns on sequence-based analysis, comparison, and species identification include the number of base ambiguities processed, gaps generated during sequence comparison, and algorithm (local or global) used for the sequence alignment.”
Please revise this as
“Other concerns include sequence ambiguities, gaps generated during DNA sequencing and sequence comparisons, and choosing the appropriate algorithm (local or global) for sequence

alignment.”.”

“The local alignment algorithm is extensively used for sequence similarity-based species identification. Several studies were conducted to identify the best variable region or combination of variable regions for bacterial classification, and a consensus remains to be implemented (Janda and Abbott, 2007; Johnson *et al*. 2019; Winand *et al*. 2020).”
Please revise this as
“Since the local alignment algorithm is extensively used for sequence similarity-based comparisons, it is important to carefully consider whether a single variable region or a combination of variable regions of the 16S rRNA gene would be ideal for bacterial classification.”

“Usage of misclassified sequence as a reference and improper bioinformatics workflows mislead the bacterial taxonomy.
Please revise this as
“Using erroneous 16S rRNA sequences as references and improper bioinformatics workflows can mislead bacterial identification”.

"Further, the growth of bioinformatics and genetic data has placed genome-based microbial classification with researchers with little or no taxonomic experience, which may also mislead the bacterial taxonomy (Baltrus, 2016)”.
Kindly revise the above statement because it is too complex and confusing.

“A few bacterial identification systems with high resolution have been developed using the sequence of polymerase chain reaction (PCR) amplified □4.5 kb long 16S–23S rRNA regions (Benítez-Páez and Sanz, 2017; Sabat *et al*. 2017; Kerkhof *et al*. 2017).”
Please revise this as
“Other methods for bacterial identification include the sequencing and analysis of the polymerase chain reaction (PCR) amplified □4.5 kb 16S–23S rRNA regions (Benítez-Páez and Sanz, 2017; Sabat *et al*. 2017; Kerkhof *et al*. 2017).”

“However, these approaches have a few limitations, such as the lack of reference 16S–23S rRNA sequence databases and complementary bioinformatics resources for reliable species identification (Sabat *et al*. 2017)”.
Please revise this as
“However, the 16S–23S rRNA sequence-based method is less practical due to the lack of appropriate reference sequence databases and reliable tools/methods for sequence analysis”

“The recent advancements in bioinformatics workflows (Winand *et al*. 2020; Schloss, 2020) and reference databases such as SILVA, EzBioCloud (Quast *et al*. 2013; Yoon, 2017) improved 16S rRNA-based bacterial taxonomy. However, a few recent genome-based studies highlighted the misclassification incidences in bacterial species and genome assemblies (Steven *et al*. 2017; Martínez-Romero, *et al*. 2018; Mateo-Estrada *et al*. 2019; Bagheri *et al*. 2020)”.
The first part here talks about 16S rRNA-based bacterial taxonomy, and the second part talks about genome assemblies. It is not clear what the connection is. Kindly revise these two sentences. A suggestion for revision is shown below (although even this suggestion fails to convey the connection):
“Although improvements in reference databases (such as SILVA and EzBioCloud) and bioinformatics workflows have facilitated 16S rRNA-based bacterial taxonomy, recent genome-

based studies have indicated that incidences of misclassification of bacterial species and erroneous genome assemblies."

"Nowadays, conventional and high throughput sequencers can amplify all the nine variable regions of the 16S rRNA gene".
Please revise as
"The entire 16S rRNA gene (~1500 bp) can be amplified and sequenced using the conventional or high throughput sequencing methods".

"Although, many 16S rRNA-based bacterial identification studies lack a complete set of variable regions (Stackebrandt *et al*. 2021)".
Please revise as
"However, many 16S rRNA sequence-based bacterial identification studies do not seem to include all of these nine variable regions (Stackebrandt *et al*. 2021)".

"The classical and high throughput sequencing technologies produce a large volume of whole-genome data. There is an urgent need to translate the genomic data for convenient microbiome analyses that ensure clinical practitioners can readily understand and quickly implement it (Church *et al*. 2020)".
Please revise as
"Due to the large volume of whole-genome data that is being produced by high throughput sequencing technologies, there is an urgent need to develop methods for analyzing this data for accurate identification of bacteria. It is also important to envisage that clinical practitioners would be able to understand these methods and implement them quickly (Church *et al*. 2020)".

"Hence, the study intended to demonstrate a workflow to develop species-specific concatenated 16S rRNA reference libraries and its analysis. The species-specific libraries can yield better resolution in sequence similarity and phylogeny based bacterial classification approaches".
Please revise as
"This study aimed to develop a workflow for accurate identification of bacteria using concatenated, species-specific 16S rRNA sequences. It was hoped that the species-specific libraries would yield much better resolution in sequence similarity- and phylogeny-based bacterial classification".

Methods
Estimation of variations in intra-genomic 16S rRNA gene copies
"Sequence alignment of 16S rRNA copies at the intra-genomic level shows a higher degree of variability in species belonging to the *Firmicutes* and *Proteobacteria* (Vetrovsky and Baldrian, 2013; Ibal *et al*. 2019)".
Please revise this as
"It has been reported that sequence alignment of 16S rRNA alleles at the intra-genomic level shows a higher degree of variability in species belonging to the *Firmicutes* and *Proteobacteria* (Vetrovsky and Baldrian, 2013; Ibal *et al*. 2019)".

"Hence, the study used eight 16S rRNA copies (Underlying data: Supplementary data 1 (Paul, 2022)) retrieved from the whole genome of *Enterobacter asburiae* strain ATCC 35953 (NZ_CP011863.1)".
Please revise this as
"Therefore, this study used eight 16S rRNA alleles (Underlying data: Supplementary data 1 (Paul,

2022)) retrieved from the complete genome of *Enterobacter asburiae* strain ATCC 35953 (NZ_CP011863.1)".

"The BLAST+ 2.13.0 (RRID:SCR_004870; Altschul *et al*. 1990) and Clustal Omega 1.2.4 (RRID:SCR_001591; Sievers *et al*. 2011) sequence alignment algorithms were used to estimate intra-genomic variability between the 16S rRNA gene copies".
Please revise this as
"To estimate intra-genomic variability between these 16S rRNA alleles, BLAST+ 2.13.0 (RRID:SCR_004870; Altschul *et al*. 1990) and Clustal Omega 1.2.4 (RRID:SCR_001591; Sievers *et al*. 2011) sequence alignment algorithms were used".

"Phylogenetic relatedness between intra-genomic 16S rRNA copies were estimated using the Maximum Likelihood method (Tamura-Nei model; 500 bootstrap replicates) with MEGA software (version 11; RRID: SCR_000667; Kumar *et al*. 2018)."
Please revise this as
"Phylogenetic analysis of these 16S rRNA alleles were performed using the maximum likelihood method (Tamura-Nei model; 500 bootstrap replicates) and the MEGA software (version 11; RRID: SCR_000667; Kumar *et al*. 2018)".

Construction of species-specific concatenated 16S rRNA reference libraries

"Previous studies have reported that several bacterial species share more than 99% sequence identity in the 16S rRNA encoding region".
Please revise this as
"Previous studies have reported that the genes encoding 16S rRNA from several bacterial species share >99% sequence identity". [ALSO PROVIDE A REFERENCE HERE]

"Hence, the 16S rRNA-based bacterial identification methods failed to discriminate such genetically related species (Deurenberg *et al*. 2017; Devanga-Ragupathi *et al*. 2018)".
Please revise this as
"Therefore, the 16S rRNA-based methods failed to correctly identify bacterial species that are genetically closely related (Deurenberg *et al*. 2017; Devanga-Ragupathi *et al*. 2018)".

"It has been reported that *Streptococcus mitis* and *Streptococcus pneumoniae* are almost indistinguishable from each other based on the sequence similarity of their 16S rRNA regions (Reller *et al*. 2007; Lal *et al*. 2011)".
Please revise this as
It has been reported that 16S rRNA-based methods cannot distinguish between *Streptococcus mitis* and *Streptococcus pneumoniae* due to the high sequence similarity (Reller *et al*. 2007; Lal *et al*. 2011).

"To develop species-specific barcode reference libraries, the study used 16S rRNA gene copies from whole-genome assemblies of four closely related species of *Streptococcus* (*S. gordonii*, *S. mitis* , *S. oralis,* and *S. pneumoniae*)".
Please DELETE the above, because it is repeated below.

"More than 463,000 whole-genome assemblies are currently available for prokaryotes at the Genome database (RRID:SCR_002474; https://www.ncbi.nlm.nih.gov/genome)".

Please revise this as
'More than 463,000 whole-genome sequences are currently (please provide a date here) available for prokaryotes (please specify if this is only for bacteria, or also includes archaea) in the Genome database (RRID:SCR_002474; https://www.ncbi.nlm.nih.gov/genome)".

"Most microbial genomes were sequenced with high throughput sequencing technologies such as Illumina/Ion-Torrent (short read sequencing) and PacBio/Nanopre (long read sequencing").
Please revise this as
"Many of these genomes were sequenced using high throughput sequencing technologies such as Illumina/Ion-Torrent (short read sequencing) and PacBio/Nanopre (long read sequencing)".
"Further, many of these whole-genome assemblies are derived through a hybrid assembly of short and long read sequence data".
Please revise this as
"Furthermore, most of these whole-genome sequences were obtained after a hybrid assembly of short and long read sequence data".

"The large volume of high throughput data can be effectively used to develop advanced genome-based approaches for microbial systematics".
Please revise this as
"This extensive, high throughput data can be effectively used to develop advanced genome-based methods for microbial systematics."

The genomic data is available in four assembly completion levels (contig, scaffold, chromosome, and complete). However, the study used only the genomes assemblies in the 'complete' stage to retrieve 16S rRNA gene copies.
Please revise this as
Although the genomic data is available in four levels (contig, scaffold, chromosome, and complete), this study used only the complete genomes to retrieve 16S rRNA genes.

"The study retrieved full-length 16S rRNA gene copies from 16 genome assemblies belonging to four *Streptococcus* species (*S. gordonii*, *S. mitis*, *S. oralis,* and *S. pneumoniae*").
Please revise this as
"To develop species-specific barcode reference libraries, this study retrieved full-length 16S rRNA genes from 16 complete genome sequences belonging to four *Streptococcus* species (*S. gordonii*, *S. mitis*, *S. oralis,* and *S. pneumoniae*)".

"The detailed information on the dataset used to develop species-specific concatenated reference libraries is provided in Table 1 and the sequences are provided in the underlying data (Supplementary data 2 (Paul, 2022))"
Please revise this as
"Details of the dataset used to develop species-specific concatenated reference libraries are provided in Table 1, and the sequences are provided in the underlying data (Supplementary data 2 (Paul, 2022))"

"To maintain equal length, sequences were trimmed out beyond the universal primer pair fD1-5'-GAG TTT GAT CCT GGC TCA-3' and rP2-5'-ACG GCT AAC TTG TTA CGA CT-3' (Weisburg *et al*. 1991) for full-length 16S rDNA amplification".
Please revise this as

"Sequences were trimmed beyond the universal primer pair (fD1-5'-GAG TTT GAT CCT GGC TCA-3' and rP2-5'-ACG GCT AAC TTG TTA CGA CT-3', which are used for full-length 16S rDNA amplification, Weisburg *et al*. 1991) to maintain uniform length [PLEASE MENTION THIS IN BASE PAIRS]"

"The study used MEGA 11 software to perform multiple sequence alignment and identify the intra-species parsimony informative (Parsim-info) variable sites".
Please revise this as
"To perform multiple sequence alignment and identify the intra-species parsimony informative (Parsim-info) variable sites, the MEGA 11 software was used".

"A species-specific barcode reference library covering entire Parsim-info variable sites was constructed by concatenating four 16S rRNA gene copies representing four different strains of a species".
Please revise this as
"A species-specific barcode reference library that covers the entire Parsim-info variable sites was constructed by concatenating four 16S rRNA gene copies from four different strains of a species".

"The rationale behind the selection of four copies for a species-specific barcode reference library is: (i) a maximum of four variations can be found on a single site, and (ii) earlier studies have shown that the mean 16S rRNA copies per genome is four (Vetrovsky and Baldrian, 2013)".
Please revise this as
"The rationale for the selection of four copies for constructing a species-specific barcode reference library was: (i) a maximum of four variations can be found at a single site, and (ii) earlier studies have shown that the mean 16S rRNA copies per genome is four (Vetrovsky and Baldrian, 2013)".

Demonstration of concatenated 16S rRNA in sequence similarity and phylogeny

"The study analyzed a few cases to demonstrate the classical sequence similarity and phylogenetic analysis using concatenated species-specific 16S rRNA reference libraries".
Please revise this as
"This study analyzed a few cases to demonstrate (i) the classical sequence similarity and (ii) phylogenetic analysis using concatenated species-specific 16S rRNA reference libraries".

"The study used nine Sanger sequenced 16S rRNA gene copies showing higher sequence similarity with multiple species of *Streptococcus* retrieved from the GenBank database (RRID:SCR_002760)".
Please revise this as
"Nine 16S rRNA gene copies (sequenced using the Sanger method) showing higher sequence similarity to the 16S rRNA genes of multiple species of *Streptococcus* were retrieved from GenBank database (RRID:SCR_002760)".

"The web based BLAST2 (version 2.13.0) program for aligning two or more sequences was used to estimate the maximum score, total alignment score, and sequence identity".
Please revise this as
"The web based BLAST2 (version 2.13.0) program for aligning two or more sequences was used to estimate the maximum score, total alignment score, and sequence identity of these nine 16S rRNA".

"A single copy of the 16S rRNA region derived through Sanger sequencing or retrieved from a whole-genome assembly can be considered as 'Query sequence".
Please revise this as
"A single 16S rRNA gene (sequenced using the Sanger method or retrieved from a whole-genome assembly was the 'Query sequence"'. [IT IS NOT CLEAR WHETHER THIS GENE WAS FROM *Streptococcus.* PLEAASE CLARIFY THIS]

"The concatenated species-specific reference libraries must be provided in the 'Subject sequence' section".
Please revise this as
"The concatenated species-specific reference libraries were provided in the 'Subject sequence' window".

"To perform phylogenetic analysis, it is mandatory that the target sequence (length = n bp) has to be concatenated four times (length = 4 × n bp), appending next to the last base".
Please revise this as
"To perform phylogenetic analysis, it is mandatory that the target sequence (length = n bp) be concatenated four times (length = 4 × n bp)".

"Phylogenetic relatedness was estimated using the Maximum Likelihood method (Tamura-Nei model; 500 bootstrap replicates) with MEGA 11 software".
Please revise this as
"Phylogenetic analysis was performed as indicated above/"

Results
Intra-genomic 16S rRNA variations in *Enterobacter asburiae*

Historically, the 16S rRNA gene sequences were used to identify known and new bacterial species.
Please revise this as
Historically, sequences of the 16S rRNA genes have been used to identify known and new bacterial species.

However, this method is impacted by several factors such as amplification efficiency, poor discriminatory power at the species level, multiple polymorphic 16S rRNA gene copies, and improper bioinformatics workflows for the data analysis.
Please revise this as
However, efficiency of PCR-based amplification, poor discrimination at the species level, multiple polymorphic 16S rRNA gene copies, and improper bioinformatics workflows for the data analysis can impact the identification. [PLEASE PROVIDE A REFERENCE HERE]

The *E. asburiae* genome had eight 16S rRNA gene copies that showed a mean identity of 99.29% in sequence alignment using Clustal Omega (global alignment), whereas BLAST (local alignment) analysis resulted in an average of 99% identity between the copies (Table 2).
Please revise this as
The genome of *E. asburiae* contains eight copies of the 16S rRNA gene. Analysis using Clustal Omega (global alignment) and BLAST (local alignment) showed that the sequences of these eight alleles had average identities of 99.29 and 99%, respectively (Table 2).

Hence, the selection of an appropriate algorithm has a significant role in the estimation of percent identity, and a vital role in sequence-based species delineation.
Please revise this as
Therefore, choosing the appropriate algorithm/tool is critical for the estimation of sequence identities and sequence-based species delineation.

Global sequence alignment programs generally perform better for highly identical sequence pairs, and the algorithm considers all the bases for the estimation of sequence identity. The multiple sequence alignment showed 22 variable sites in 16S rRNA gene copies of the *E. asburiae* genome (Figure 1).
Please revise this as
For analyzing sequence pairs that are highly identical, global sequence alignment programs/tools seem to be more appropriate because they consider all the nucleotides for the estimation of sequence identity.

The multiple sequence alignment showed 22 variable sites in 16S rRNA gene copies of the *E. asburiae* genome (Figure 1).
Please revise this as
Multiple sequence alignment [PLEASE MENTION THE TOOL/ALGORITHM HERE, AND ALSO IN THE LEGEND OF FIGURE 1] of the sequences of the eight alleles of the 16S rRNA gene in the genome of *E. asburiae* showed 22 variable sites (Figure 1).

The evolutionary relationship between species is usually represented in a phylogenetic tree drawn using a single barcode gene, multiple genes, or whole genomes.
Please revise this as
The evolutionary relationship between species is usually represented using a phylogenetic tree based on the analysis of a single gene, multiple genes, or whole genomes.

However, bacterial species nomenclature is mainly designated based on the confidence obtained from the phylogenetic tree derived through single copy 16S rRNA analysis.
Please revise this as
However, bacterial identification and classification is mainly based on the phylogenetic analysis of single copies of 16S rRNA genes

To highlight how the intra-genomic 16S rRNA variations influence the species delineation, a phylogenetic tree was constructed using eight 16S rRNA gene copies of *E. asburiae* reference genome showing multiple nodes (Figure 2).
Please revise this as
A phylogenetic tree was constructed to understand how variations in the sequences of the eight alleles of the 16S rRNA gene in the genome of *E. asburiae* influence species delineation (Figure 2).

The sequence similarity and phylogeny-based analysis indicate that the intra-genomic variations in 16S rRNA copies may mislead the bacterial taxonomy in single gene copy approaches.
IT IS NOT CLEAR TO ME HOW THIS ANALYSIS "indicate(s) that the intra-genomic variations in 16S rRNA copies may mislead the bacterial taxonomy in single gene copy approaches". **KINDLY ELABORATE ON THE SAME.**

Species-specific concatenated 16S rRNA libraries

The study selected four *Streptococcus* species (*S. gordonii*, *S. mitis*, *S. oralis*, and *S. pneumoniae*) to construct species-specific concatenated 16S rRNA reference libraries.
Please revise this as
This study selected four species of *Streptococcus* (*S. gordonii*, *S. mitis*, *S. oralis*, and *S. pneumoniae*) to construct species-specific concatenated reference libraries based on 16S rRNA gene sequences obtained from complete genomes.

The study used 16S rRNA copies retrieved from four whole genome assemblies in the 'complete' stage to construct a species-specific barcode library.
Please DELETE the above sentence.

Four copies of the 16S rRNA gene are required to construct the concatenated library for a species.
Please revise this as
Sequences from four copies of the 16S rRNA gene are required to construct a concatenated library for a species [PLEASE PROVIDE A REFERENCE HERE].
The details of constructed species-specific libraries are listed in Table 1 and the sequence is provided in the underlying data (Supplementary data 3 (Paul, 2022)).
Please revise this as
The details of species-specific libraries are listed in Table 1 and the sequences are provided in the underlying data (Supplementary data 3 (Paul, 2022)). THIS REFERENCE IS INCOMPLETE IN THE LIST, AND THE LINK IS NOT WORKING. PLEASE CHECK ANC CORRECT.
ALSO, IN THE TITLE OF TABLE 1, "for the development of concatenated" SHOULD BE REVISED AS "for the construction of concatenated"

The 16S rRNA sequence analysis shows 24 Parsim-info variable sites for *S. oralis*, 11 variations in *S. mitis*, seven variations in *S. gordonii,* and six variations found in *S. pneumoniae.*
Please revise this as
Analysis using the sequences of 16S rRNA genes showed 24, 11, 7, and 6 Parsim-info variable sites for *S. oralis*, *S. mitis*, *S. gordonii,* and *S. pneumoniae*, respectively. [PLEASE PROVIDE/SHOW THE DATA FOR THIS]

The observed intra-species Parsim-info variable sites are residing on both conserved and variable regions of the 16S rRNA gene.
Please revise this as
The intra-species [PLEASE CHECK IF THIS SHOULD BE "inter-species"] Parsim-info variable sites were located in both the conserved and variable regions of the 16S rRNA gene. [PLEASE PROVIDE/SHOW THE DATA FOR THIS]

The study used full-length 16S rRNA copies from four different strains to highlight the variations at the species level.
Please revise this as
This study used full-length sequences of 16S rRNA genes from four different species to check the variations at the species level.

However, a large volume of partial 16S rRNA sequences are available in the public genetic databases.
Please revise this as

However, a large number of partial sequences of 16S rRNA genes are available in the public genetic databases.

In such cases, a species-specific concatenated 16S rRNA reference library can be developed with partial sequences.
Please revise this as
In such cases, a species-specific concatenated reference library can be constructed using partial sequences.

Intra-species variation on 16S rRNA gene copies influences the sequence based bacterial taxonomy.
Please revise this as
Intra-species [PLEASE CHECK IF THIS SHOULD BE "inter-species"] variations in the sequences of 16S rRNA gene copies influences the sequence-based bacterial identification. [PLEASE PROVIDE/SHOW THE DATA FOR THIS, OR SUBSTANTIATE THE CONCLUSION]

Hence, the concatenated 16S rRNA approach yields better resolution than single copy analysis in classical sequence similarity and phylogeny based species identification approaches.
Please revise this as
Therefore, concatenation of the sequences of 16S rRNA genes/alleles provides much better resolution compared to analysis using sequences from a single copy of the 16S rRNA gene. PLEASE NOTE: IN THE ABSENCE OF DATA, THIS STATEMENT REMAINS UNSUBSTANTIATED.

Demonstration of concatenated 16S rRNA based species identification

The study compared nine 16S rRNA sequences representing *Streptococcus* species (Table 3) with species-specific concatenated reference libraries.
Please revise this as
This study compared sequences of nine 16S rRNA genes from different species of *Streptococcus* (Table 3) using species-specific concatenated reference libraries.

Concatenated sequence analysis gives better resolution in sequence similarity search and phylogenetic analysis.
Please revise this as
Concatenated sequences provide much better resolution in sequence similarity search and phylogenetic analysis.

The sequence accession numbers GU470907.1 and KF933785.1 classified as *S. mitis* showed a higher maximum and total alignment score with *S. oralis* than *S. mitis* (Table 3).
Please revise this as
Two sequences (accession numbers GU470907.1 and KF933785.1) from *S. mitis* had a higher maximum and total alignment score to sequences from *S. oralis* than *S. mitis* (Table 3). PLEASE PROVIDE THE ACCESSION NUMBERS FROM S. ORALIS THAT PRODUCED THIS RESULT.

Whereas the sequence (OM368574.1; classified as *S. mitis*) showed a higher sequence alignment score with *S. pneumoniae.*
Please revise this as
Furthermore, yet another sequence from *S. mitis* (accession number OM368574.1) had a higher

alignment score to sequences from *S. pneumoniae.* PLEASE PROVIDE THE ACCESSION NUMBERS FROM S. PNEUMONIAE THAT PRODUCED THIS RESULT.

Figure 3A shows a maximum likelihood tree of the nine 16S rRNA gene sequences with four concatenated species-specific reference libraries.
Please revise this as
The maximum likelihood phylogenetic tree based on four concatenated species-specific reference libraries and the sequences of nine 16S rRNA genes is shown in Figure 3A. [THE LEGEND IS INSUFFICIENT; PLEASE PROVIDE MORE DETAILS IN THE LEGEND. ALSO, FIGURE 3B HAS NOT BEEN CALLED IN THE RESULTS SECTION, IT HAS BEEN CALLED DIRECTLY IN THE DISCUSSION SECTION]

The concatenated GU470907.1 and KF933785.1 sequences showed a phylogenetic relationship with *S. oralis* and sequence OM368574.1 was genetically related to *S. pneumoniae.*
Please revise this as
[BASED ON THE ACCESSION NUMBERS, GU470907.1 and KF933785.1 CANNOT BE CONCATENATED SEQUENCES, PLEASE CHECK AND CORRECT]
Two sequences (accession numbers GU470907.1 and KF933785.1) from *S. mitis* appeared to be phylogenetically closer to *S. oralis*, and yet another sequence from *S. mitis* (accession number OM368574.1) from *S. mitis* was closer to *S. pneumoniae.* [PLEASE NOTE: THE BOOTSTRAP VALUES ARE RATHER LOW; THEREFORE THE RESULTS NEED TO BE INTERPRETED CAREFULLY]

These results indicate that the species-specific concatenated 16S rRNA reference libraries have great potential in the taxonomic classification.
Please revise this as
These results further confirm that species-specific concatenated 16S rRNA reference libraries provide much better taxonomic resolution.

Hence, the study suggests the usage of concatenated variable 16S rRNA copies for sequence similarity and phylogeny-based species identification.
Please revise this as
Therefore, this study recommends using concatenated sequences of 16S rRNA genes for sequence similarity- and phylogeny-based species identification.

A species-specific reference library with concatenated 16S rRNA gene copies provides better resolution in phylogenetic analysis than the single copy inference.
THE SENTENCE ABOVE MAY BE DELETED BECAUSE A SIMILAR STATEMENT HAS ALREADY BEEN MADE.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bacterial Genomics, Bacterial Taxonomy, Comparative Genomics, Pathogenomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 22 Mar 2023

Bobby Paul

Dear Sir,

Thank you very much for your critical review and valuable suggestions for manuscript improvement. I have revised the manuscript by addressing all the suggestions and resubmitted it for your consideration. I sincerely hope that the modified manuscript is sufficiently improved, and kindly respond if you have any further questions or comments.

Thank you

Kind Regards
Bobby Paul

Competing Interests: Nil