Difficulty in achieving high identification accuracy and high amplification efficiency in bacterial 16S rRNA amplicon metagenomics

Wenfa Ng

Department of Biomedical Engineering, National University of Singapore
Email: ngwenfa771@hotmail.com

Abstract: Bacterial phylogenetics has largely been determined via 16S rRNA gene sequencing and phylogenetic tree reconstruction. Observed utility of this approach has driven the popularity of the 16S rRNA gene amplicon metagenomics method for profiling and identifying diverse microbes from specific habitats. This work sought to develop universal primers for amplifying the 16S rRNA gene from a consortium of disparate microbial species. Using multiple sequence alignment of the 16S rRNA gene of a variety of microbes, the resulting highly conserved region of the consensus sequence was used for design of universal polymerase chain reaction (PCR) primers for 16S rRNA gene. Application of the universal primers in simulated PCR reveals poor amplification efficiency where only 12 species out of 31 generated an amplicon. BLAST analysis of the resulting amplicons reveals a classification error of 50%. More significantly, analysis of the amplicon length indicates variable read length ranging from 81 to 122 base pair compared to the predicted read length of 100 base pairs. This suggests that the 16S rRNA gene harbours significant hitherto underappreciated sequence diversity, and may have unknown alternative splicing and recombination mechanisms. Overall, results from this study suggests that primer design for 16S rRNA amplicon metagenomics may be application and habitat specific, where it is difficult to design universal primers for all bacterial species. Conceptually, this meant that there may be sequence co-evolution in 16S rRNA gene for microbial species in the habitat where environmental and nutritional conditions impact on 16S rRNA gene structure and sequence. In essence, 16S rRNA gene may habour epigenetics signals at the gene level.

Keywords: 16S rRNA gene; phylogenetics; amplicon metagenomics; bacterial species; gene structure and sequence

Subject areas: systems biology; bioinformatics; molecular biology; evolutionary biology; biochemistry

---

Phylogenetics traces the evolutionary relationships between organisms in the same or across habitats.[1] At the molecular level, efforts to delineate phylogenetic relationships between different microbial species typically rests on the 16S rRNA gene.[2] [3] Indeed, 16S rRNA gene is the anchor for the current three domains of life phylogenetic tree, and is the bedrock of phylogenetics effort.[4]

This work starts off with the hypothesis of whether it is possible to use the highly conserved regions of the 16S rRNA gene to design universal primers for the polymerase chain reaction (PCR) amplification of amplicons for subsequent sequencing or for direct identification via quantitative polymerase chain reaction (qPCR). Such an effort may result in universal primers useful for profiling the microbial diversity in many habitats via amplicon metagenomics or direct identification via Sanger sequencing of the amplicon.

Efforts were first expended to download the 16S rRNA gene sequence of a variety of disparate microbial species on the de Silva database portal linked to the European Nucleotide Archive. Using an in-house MATLAB phylogenetic analysis software, multiple sequence alignment was conducted for the downloaded 16S rRNA gene sequences to yield a consensus sequence. Visual inspection of the multiple sequence alignment reveals

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1931.v1

stretches of the gene where there is high level of sequence conservation across different species. This region was used to design the universal primers.

Next, universal primers designed were used in another in-house MATLAB simulated PCR software to conduct computational biology experiments examining whether the primers could amplify a large number of amplicons from different microbial species, and how well the amplicons could be used in accurate identification of the microbial species.

Table 1. Identification accuracy and amplification efficiency of universal 16S rRNA primer designed in this study for a simulated microbiota.

Bacterial species	Amplicon length	Identity of amplicon	Percent similarity	Error
Shigella flexneri, strain WAB1966	81	Shigella flexneri	100	
Shigella dysenteriae, strain: B11-1	0			
Salmonella enterica, strain: NBRC 12529	101	Salmonella enterica subsp. enterica serovar Agona strain R21.1368	100	
Salmonella bongori strain JEO 4162	102	Salmonella enterica subsp. enterica serovar Agona strain R21.1368	100	Error in classification
Campylobacter lari, strain:A2	0			
Campylobacter fetus MGH 97-2126	0			
Campylobacter coli, strain: K80	0			
Campylobacter jejuni strain LMG 9217	0			
Campylobacter upsaliensis gene strain: TMUC1534	0			
Proteus vulgaris gene, strain: X10-1	119	Proteus vulgaris	100	
Aeromonas hydrophila	98	Aeromonas encheleia	100	Error in classification
Citrobacter freundii, strain: SSCT56	100	Citrobacter freundii	100	
Staphylococcus aureus JCM 2413	0			
Serratia liquefaciens gene, strain: NBRC 12979	101	Dickeya dianthicola	100	Error in classification
Vibrio parahaemolyticus clone Vp23	97	Vibrio harveyi	100	Error in classification
Clostridium perfringens strain: ATCC 13124	0			
Escherichia coli	122	Escherichia coli	100	
Klebsiella oxytoca gene	90	Salmonella enterica subsp. enterica serovar Agona strain R21.1368	100	Error in classification
Listeria monocytogenes strain CECT 4032	0			
Candida albicans strain JCM 1542	0			
Arcobacter butzleri strain ED-1	0			
Pseudomonas aeruginosa	100	Pseudomonas aeruginosa	100	
Enterococcus faecalis	0			
Giardia intestinalis	0			
Yersinia enterocolitica	94	Yersinia kristensenii	100	Error in classification
Bacteroides fragilis strain: JCM 11019	0			
Cryptosporidium parvum isolate: Sakha103	0			
Entamoeba histolytica isolate EH_IQ1	0			
Helicobacter hepaticus	0			
Helicobacter cinaedi gene strain: PAGU0597	0			
Candidatus Helicobacter heilmannii	0			

Table 1 shows the identification accuracy and amplification efficiency of the universal primers designed in this study for different microbial species in a simulated microbiota. Results show poor amplification efficiency where only 12 out of 31 microbial species yielded a 16S rRNA gene amplicon. More importantly, 50% of the amplified amplicons could be used for accurate identification of the original microbial species. This suggests that the concept of universal primers for disparate microorganisms may need a rethink, and there is hitherto unknown gene architecture and genetic recombination mechanisms in 16S rRNA gene that impact on the results we observed.

Specifically, Table 1 also reveals that the amplicon read length for 16S rRNA gene ranges from 81 to 122 base pair for a target 100 base pair region. This suggests that there is substantial sequence diversity and genetic recombination such as alternative splicing occurring in the 16S rRNA gene of different microbial species which has eluded our understanding of the gene thus far. Implications for this observation include: (i) the architecture of the 16S rRNA gene may be affected by environmental and nutritional conditions where similarities in these conditions lead to similar gene architecture and read length, (ii) 16S rRNA gene of microbial species in the same habitat may have similar gene architecture and read length, and (iii) epigenetic effects may be encoded in the 16S rRNA gene at the gene level.

Overall, universal primers for amplifying the 16S rRNA gene in amplicon metagenomics is highly sought after. Results from this work reveals that the concept of universal primers in microbial amplicon metagenomics may need a rethink especially for microbes in divergent environments with different nutritional conditions. Observations of poor amplification efficiencies, poor identification accuracies, and variable amplicon read lengths point to a complex 16S rRNA gene architecture with a larger than anticipated sequence space. A complex 16S rRNA gene architecture would exert tight constraints on

 doi:10.20944/preprints202305.1931.v1

yielding a similar read length amplicon for identification, while an expanded sequence space meant that regions with high sequence conservation may be too small for yielding amplicon of sufficient length to afford accurate identification of microbial species. Collectively, 16S rRNA gene is a complex gene with epigenetic encodings, and its use in amplicon metagenomics may require care such as using it to identify species in a particular habitat where microbes with similar metabolic demands and environmental adaptations reside. Such metabolic and environmental adaptations would select for particular 16S rRNA gene architecture and sequence space as the molecule plays an important role in affecting the stability of the large and small ribosome macromolecular complex, which in turn, affects protein synthesis rate and cell growth rate.

Conflicts of interest: The author declares no conflicts of interest.

Funding: No funding was used in this work.

References

1. Kapli, P., Yang, Z. & Telford, M. J. Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* **21**, 428–444 (2020).
2. Hassler, H. B. *et al.* Phylogenies of the 16S rRNA gene and its hypervariable regions lack concordance with core genome phylogenies. *Microbiome* **10**, 104 (2022).
3. Yang, B., Wang, Y. & Qian, P.-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* **17**, 135 (2016).
4. Forterre, P. The universal tree of life: an update. *Front. Microbiol.* **6**, (2015).