A Comprehensive Approach for Profiling Human Oral Microbiome with NGS 16S rRNA Data

989

T. Chen[1], **J. McCafferty**[1] and F.E. Dewhirst[1,2]

[1]*The Forsyth Institute, Cambridge, Massachusetts,*

[2]*Harvard School of Dental Medicine, Boston, Massachusetts*

Forsyth
Pioneering Discoveries
Profound Change

VE RI TAS

Abstract

Objectives: Next generation sequencing platforms such as those based on the Illumina or Roche 454, have become cost-effective and commonly applied to study human oral microbiome by deep 16S rRNA gene sequencing of microbial samples. New and existing bioinformatics software pipelines are being developed and modified for analyzing the sequence reads and interpreting the microbial richness and diversity of the samples. Different software and analyzing procedures may result in different conclusions even on the same dataset. In this report, we evaluate pipelines and proposed a comprehensive approach that is optimized specifically for analyzing the V3-V4 region of 16S rRNA short read sequences derived from human oral samples.

Methods: We isolated the V3-V4 hyper-variable region of the 16S rRNA gene from the full length reference sequences contained in the HOMD v13.2 reference database. The reads were clustered into OTUs and taxonomy was assigned in order to demonstrate the achievable resolution expected from sequencing this region.

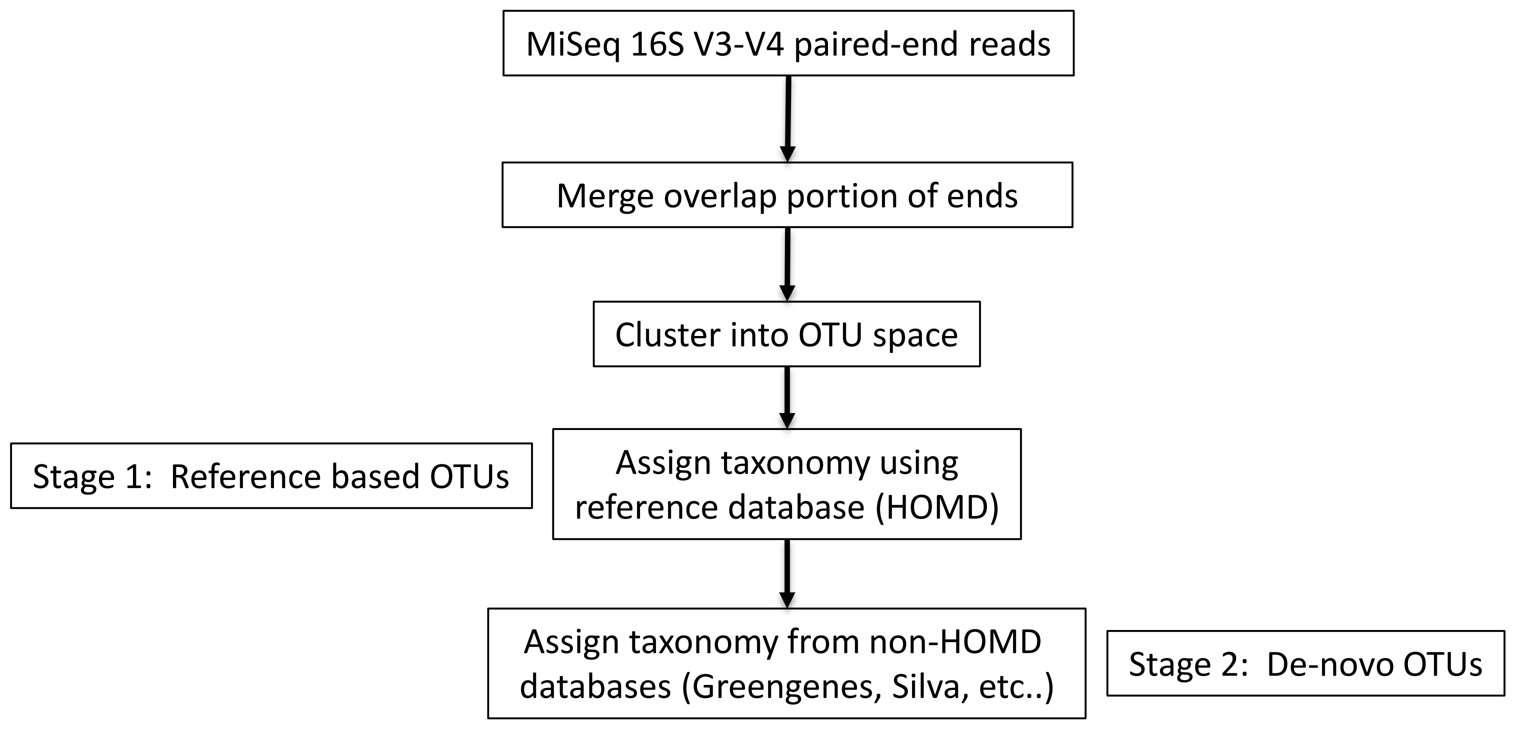
Results: The human oral microbiome has been well-characterized and full length 16S rRNA gene sequences of the most abundant species are available as references. We propose the use of two-stage, open-ended reference-base OTU calling pipeline: 1) reference-based OTU calling using HOMD 16S rRNA references and taxonomy inferred from the HOMD taxonomy; 2) *de-novo* OTU calling of the reads not mapped in stage 1 and taxonomy inferred from non-HOMD references, such as GreenGenes or Silva databases.

Conclusion: The proposed two-stage approach for NGS 16S rRNA data is comprehensive in mapping the reads to known human oral taxa as well as in discovering novel taxa in the oral microbial samples. Taxonomic resolution to the species level is difficult using 16S rRNA short reads. We have demonstrated this problem using the V3-V4 region of the 16S gene. Where species resolution is fuzzy, reporting sub-groups of potential species will allow researchers the ability to further investigate species of interest with species specific probes.

Introduction

Next generation sequencing platforms such as those based on the Illumina or Roche 454, have become cost-effective and commonly applied to study human oral microbiome by deep 16S rRNA gene sequencing of microbial samples. New and existing bioinformatics software pipelines are being developed and modified for analyzing the sequence reads and interpreting the microbial richness and diversity of the samples. Different software and analyzing procedures may result in different conclusions even on the same dataset. In this report, we evaluate pipelines and proposed a comprehensive approach that is optimized specifically for analyzing the V3-V4 region of 16S rRNA short read sequences derived from human oral samples.

Methods



MiSeq 16S V3-V4 paired-end reads

Merge overlap portion of ends

Cluster into OTU space

Stage 1: Reference based OTUs

Assign taxonomy using reference database (HOMD)

Assign taxonomy from non-HOMD databases (Greengenes, Silva, etc..)

Stage 2: De-novo OTUs

Fig1. Recommended 16S RNA analysis pipeline for studying human oral microbiome.

Conclusions

The proposed two-stage approach for NGS 16S rRNA data is comprehensive in mapping the reads to known human oral taxa as well as in discovering novel taxa in the oral microbial samples. Taxonomic resolution to the species level is difficult using 16S rRNA short reads. We have demonstrated this problem using the V3-V4 region of the 16S gene. Where species resolution is fuzzy, reporting sub-groups of potential species will allow researchers the ability to further investigate species of interest with species specific probes.

References

Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner ACR, Yu W-H, Lakshmanan A, Wade WG. 2010. The human oral microbiome. *J. Bacteriol.* 192:5002-5017.

T Magoc and S Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:21 (2011).

J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 2010.

Edgar,RC, Haas,BJ, Clemente,JC, Quince,C, Knight,R (2011) UCHIME improves sensitivity and speed of chimera detection, *Bioinformatics*

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 72:5069-72.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.

Results

Adjusting the sequence identity level at which to cluster short reads will affect the resolution of taxonomic assignments. We have demonstrated this by using a 99% sequence identity threshold to cluster the V3-V4 region HOMD V13.2 reference sequences. Resolution down to the species level can be achieved in ~84% of the V3-V4 reference sequences (Fig2). Sequences that fail to cluster into OTUs with unique taxon will cluster into groups containing similar taxa higher up in rank. While a majority of species can be identified at this level there are still some problematic species that are too similar to separate (Fig3). The genus of *Streptococcus* is one example (Table 1). We were able to identify 9 different species while the rest were clustered into 3 sub-groups at the genus level with the largest sub-group containing 31 species.

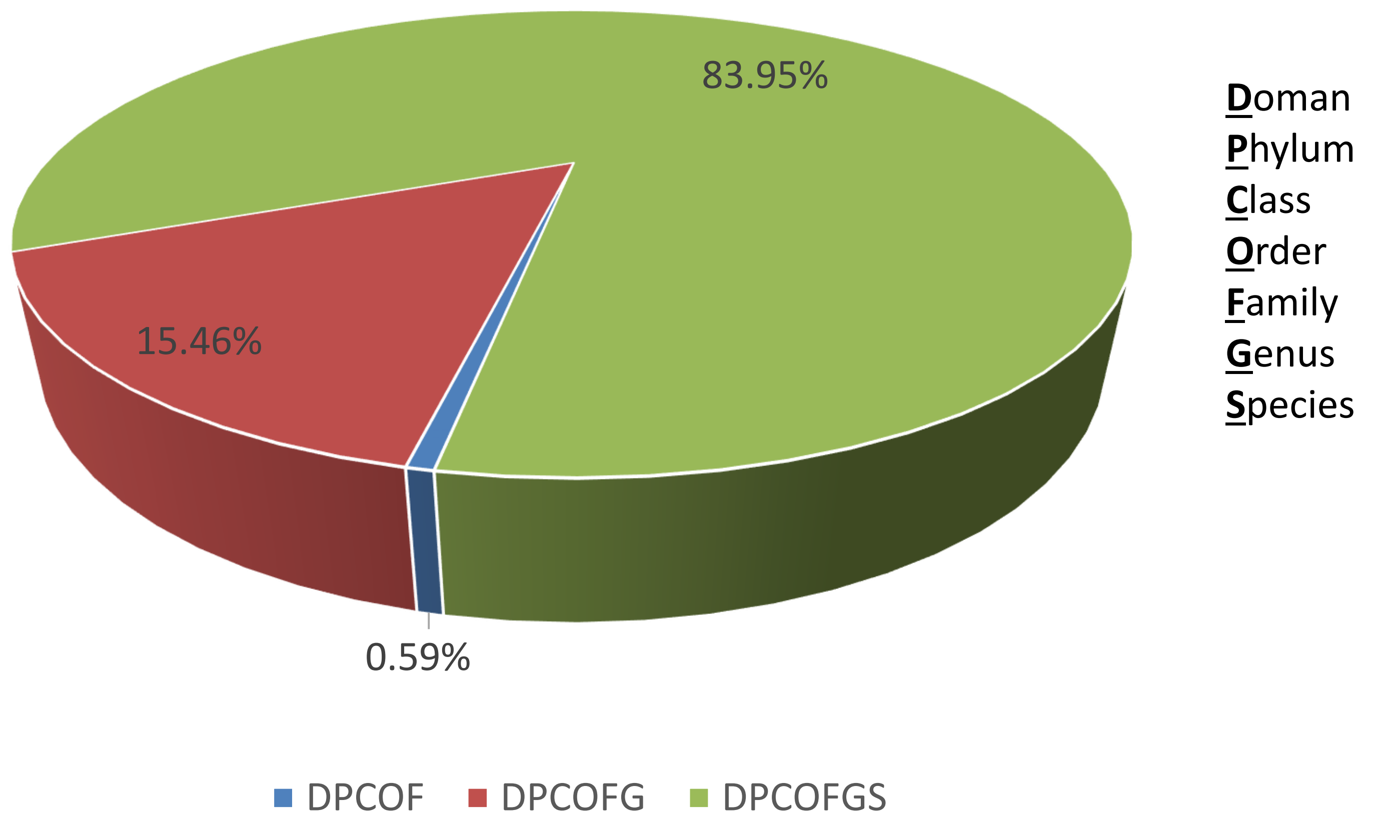
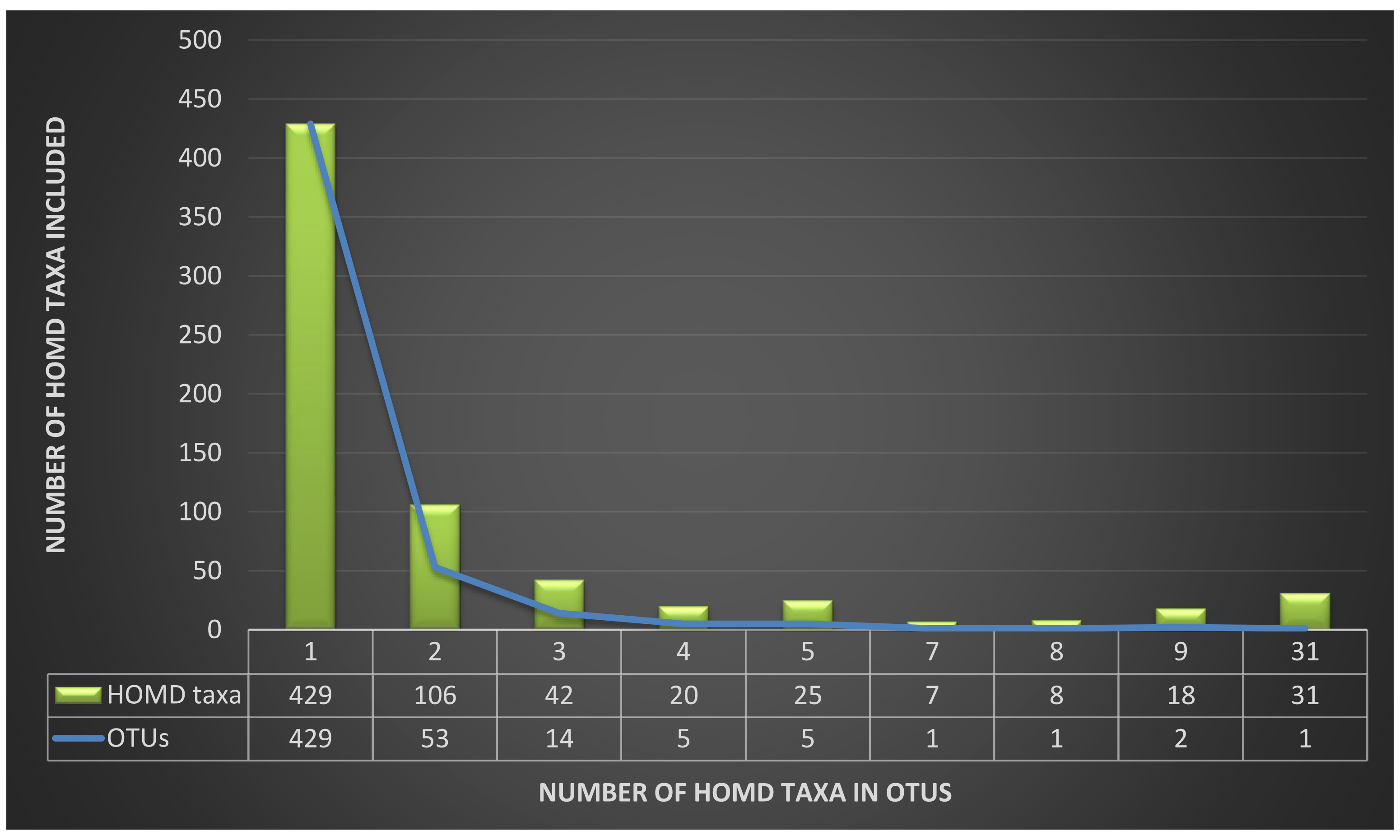


Fig2. Resolution of V3-V4 region from HOMD V13.2 reference sequences.



	1	2	3	4	5	7	8	9	31
HOMD taxa	429	106	42	20	25	7	8	18	31
OTUs	429	53	14	5	5	1	1	2	1

Fig3. OTU clustering of V3-V4 region from HOMD V13.2 reference sequences.

Table 1. Example of multiple species clustered into the same OTU. The genus *Streptococcus* clusters into multiple OTUs. Some species can be identified directly while others can only be identified as a sub-group of *Streptococcus* .

Streptococcus Identifiable species 9 total	*S. agalactiae; S. anginosus; S. intermedius; S. constellatus; S. downei; S. mutans; S. pyogenes; S. sobrinus; S. sp. oral taxon 487*
Streptococcus-sub-group1	*S. anginosus; S. intermedius*
Streptococcus-sub-group2	*S. salivarius; S. vestibularis*
Streptococcus-sub-group3 31 total	*S. australis; S. cristatus; S. gordonii; S. infantis; S. mitis; S. mitis bv 2; S. oligofermentans; S. oralis; S. parasanguinis; S. parasanguinis II; S. peroris; S. pneumoniae; S. sanguinis; S. sinensis; S. sp. oral taxon 055; S. sp. oral taxon 056; S. sp. oral taxon 057; S. sp. oral taxon 058; S. sp. oral taxon 061; S. sp. oral taxon 064; S. sp. oral taxon 065; S. sp. oral taxon 066; S. sp. oral taxon 067; S. sp. oral taxon 068; S. sp. oral taxon 069; S. sp. oral taxon 070; S. sp. oral taxon 071; S. sp. oral taxon 074; S. sp. oral taxon 423; S. sp. oral taxon 431; S. sp. oral taxon 486*

Link