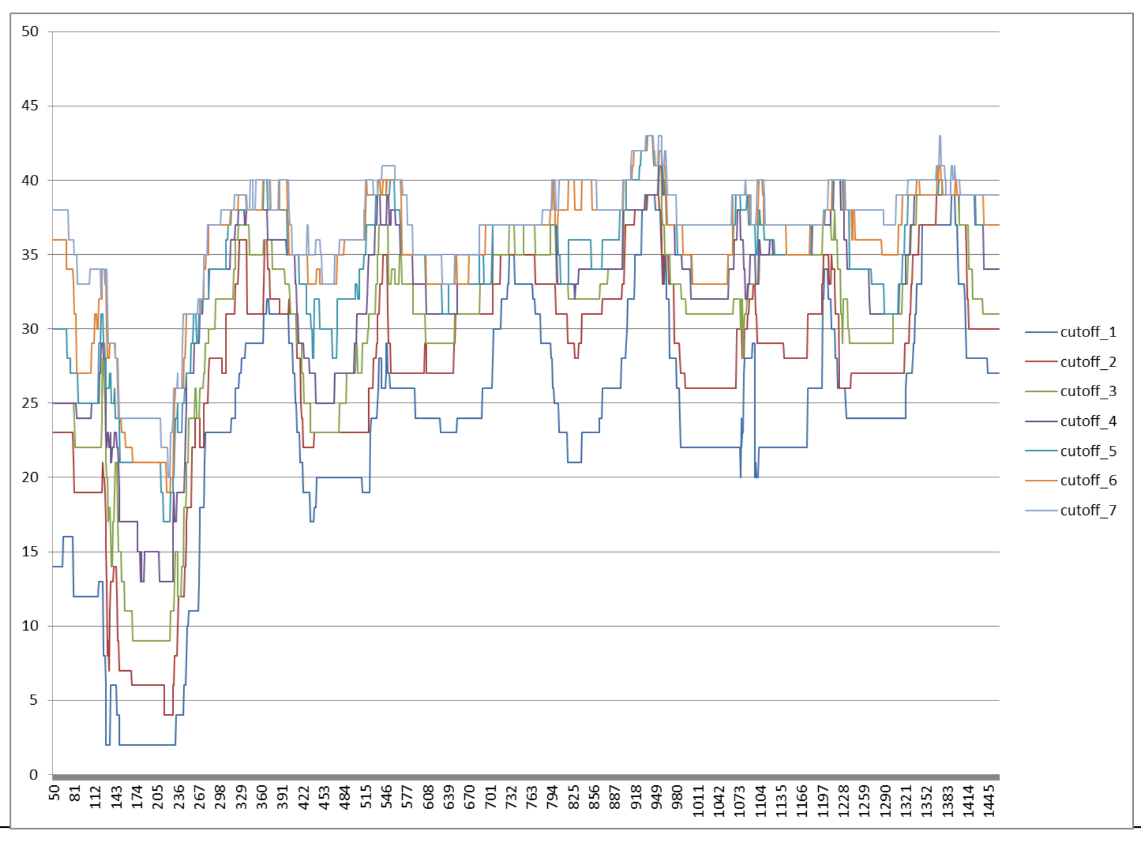
Wrong takeaway from the HMP mock community study

Floyd E. Dewhirst, Tsute Chen, and Lina L. Faller

Department of Microbiology, The Forsyth Institute
245 First St. Cambridge, MA 02142 USA

Introduction

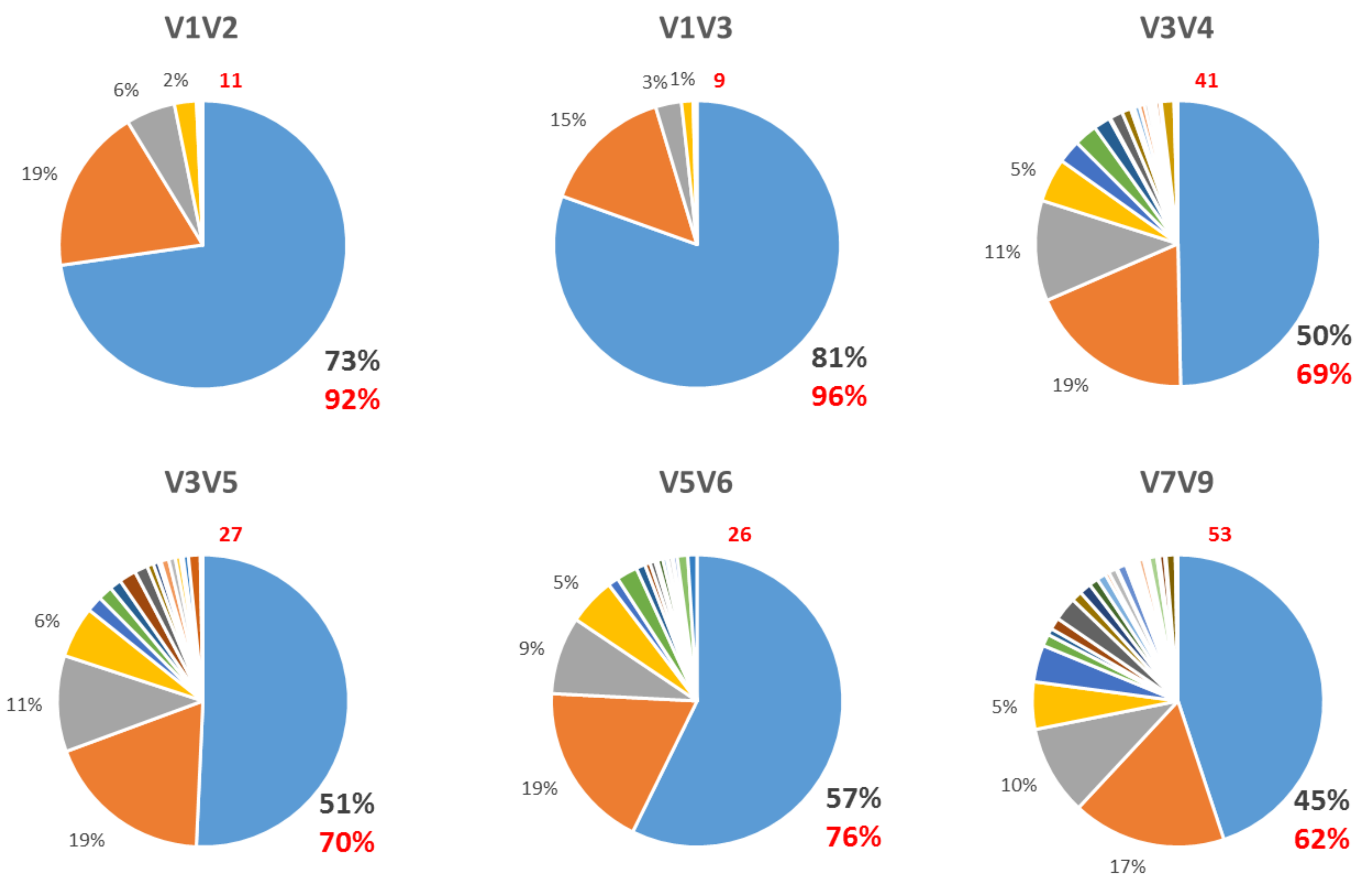
Understanding microbiomes, whether host associated, or in the natural environments, is critical to understanding the health or disease status or a host of the functioning of a natural environment. The ability to examine 16S rDNA sequences of prokaryotes in different environments has revealed a diversity of prokaryotes totally beyond what was known from older cultivation studies. From Sanger analysis of clone libraries to next gen sequencing with first 454 pyrosequencing and now Illumina sequencing, methods have evolved. While this progression in sequencing technologies has raised the number of reads by many orders of magnitude and similarly reduced the cost per read, these changes have come with an often overlooked cost. The phylogenetic resolution obtainable from sequencing any biological macromolecule is directly related to the length of the read. Full 1500 base 16S rDNA reads produces excellent phylogenies for prokaryotes and are supported by gene trees based on concatenated conserved proteins. Phylogenies based on 23S rDNA trees with 3000 bases, or 16S & 23S rDNA trees with 4500 bases are even more robust [Dewhirst 2005]. Identification or phylogeny based on shorter 500 base sequences is more problematic because of the limited information content. However, many excellent human microbiome studies have been published using 454 technology to sequence the 16S rDNA V1-V3 region with identifications to species level for most reads and taxa. For the human oral microbiome, there is a taxonomic framework with curated reference sequences for the nearly 700 most abundant oral taxa (www.homd.org) [Dewhirst 2010]. Many studies have been published mapping reads to human oral taxa (defined at the species level or 98.5% cutoff for phylotypes). For various technical reasons, few laboratories have been able to successfully sequence the V1-V3 region using Illumina sequencing, and have switched to sequencing the V3-V4 region. Analysis pipelines often group reads into taxonomically unanchored OTUs or to genus or higher taxonomic resolution. Two thirds of the taxa in the human oral microbiome database have full genomes and all of the named and many unnamed taxa are associated with an extensive published literature. When microbiome reads are not tied to carefully defined species level identifications, the link to the wealth of genomic, phenotypic, clinical and bibliographic information is lost. The figures shown in this poster present research in our laboratory documenting how one's ability to link NGS reads to defined species depends critically on the region of the 16S rDNA sequenced.



A sliding window of 100 bases was passed across aligned sequences for a mock community of 43 members (including 24 Streptococci). Plotted is the number of taxa that could not be differentiated from the other 42 taxa by the threshold number of bases.

A sliding window of 100 bases was passed across aligned sequences for the HOMD 688 taxa. Plotted is the number of taxa that could not be differentiated from the all other taxa by the threshold number of bases.

Utility of various V-regions



Aligned reference sequences for 688 taxa in Human Oral Microbiome Database were segmented into six V regions. Each sequence fragment was compared by BLASTN to the entire set of reference sequences. The number of Human Oral Taxa (HOTs) hit by each sequence at 98% similarity was recorded. Ideally each sequence should hit only one HOT. In the pie charts above, the blue represents the percentage of sequences which hit one taxon, orange two taxa, grey three taxa, yellow four, etc. The percentage in red below each chart is percentage of sequences that hit only one or two taxa.

Conclusions

- The region of 16S rDNA sequenced and sequencing length determine taxonomic resolution.
- Species level identification of reads (or better) is required for most sophisticated explanation of microbiome differences between sites, subjects or disease states.
- Most of the taxa in the HMP mock community study differed at the phylum, class, order or family level [Jumpstart 2012]. ***From this widely divergent set of taxa, many people wrongly concluded that V region doesn't matter for microbiome studies as the HMP mock community taxa were well differentiated in all V regions.*** Examination of the 700 species in the oral microbiome shows the weakness of using V3-V4 vs the V1-V3 region for species level identification. The V3-V4 is very poor for differentiation most members of the *Streptococcus* genus, the most abundant and important genus in the oral cavity. This is because phylogenetically significant variability occurs primarily in V1 and V2.
- For key body sites, full genomes are available for >2/3 of ~700 most abundant species/phylotypes. Microbiome results need to identify reads to species level taxa so that they can tap into the vast wealth of genomic, phenotypic, and bibliographic information linked to those species. OTU 123 is meaningless — *Staphylococcus aureus* is highly meaningful.
- There is a distressing tendency for investigations to report how summary statistics like alpha and beta diversity change rather than report how species or strains change over time.
- Despite published description of good primer design, many centers still use highly flawed, non degenerate, primers that miss significant groups of key organisms in many microbiomes [Frank 2008].

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.

Dewhirst FE, Shen Z, Scimeca MS, Stokes LN, Boumenna T, Chen T, Paster BJ, Fox JG. 2005. Discordant 16S rDNA and 23S rDNA phylogenies for the genus Helicobacter: Implications for phylogenetic inference and systematics. *J Bacteriol*. 187:6106-18.

Frank JA, Reich CI, Sharma S, Weisbaum JS, Wilson BA, Olsen GJ. 2008. Critical evaluation of two primers commonly used for amplification of Bacterial 16S rRNA Genes. *AEM* 74:2461-2470.

Jumpstart Consortium Human Microbiome Project Data Generation Working Group. 2012. Evaluation of 16S rDNA-Based Community Profiling for Human Microbiome Research. *PlosOne* 7:e39315