Supporting Information

Eren et al. 10.1073/pnas.1409644111

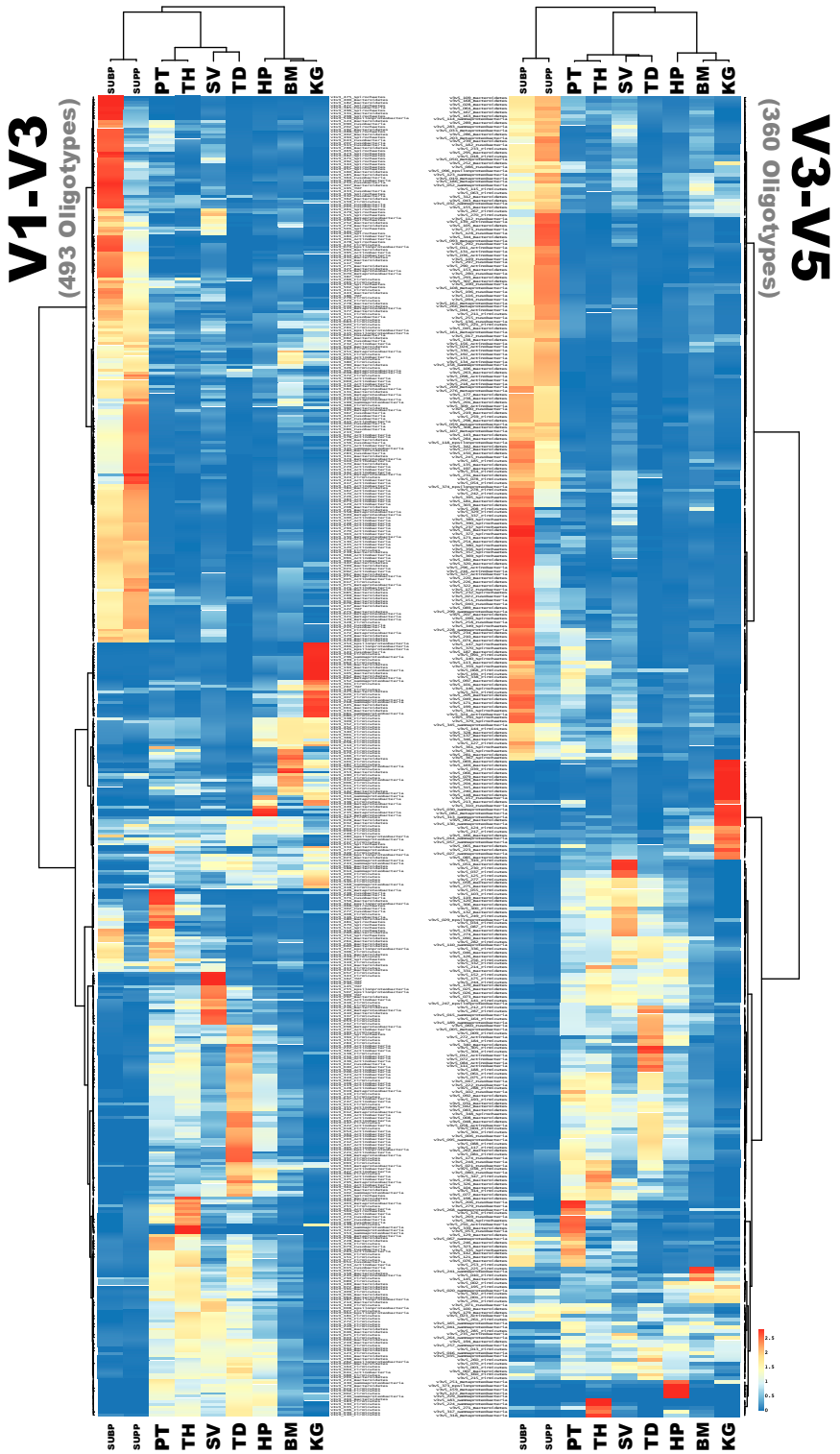

Fig. S1. Heat map analysis of oral oligotypes identified in the V1-V3 and V3-V5 datasets. Each oligotype is represented by a row, where the background color of each box represents the relative abundance of that oligotype in each of the oral sites. Blue and red indicate low and high relative abundance, respectively.

Legend continued on following page

Clustering of oral sites and oligotypes is defined by the Morisita-Horn dissimilarity index. Abbreviations for oral sites: SUBP, subgingival plaque; SUPP, supragingival plaque; PT, palatine tonsils; TH, throat; SV, saliva; TD, tongue dorsum; HP, hard palate; BM, buccal mucosa; KG, keratinized gingiva.

Dataset S1. Oligotyping results for V1-V3 data

Dataset S1

Tab 1 reports the overall number of reads analyzed and oligotypes detected for each phylum, including the average read length and the minimum substantive abundance criterion for each phylum. Tab 2 reports the number of reads classified into the analyzed and unanalyzed phyla for each sample. Tab 3 reports the number of unanalyzed reads in each phylum for each sample. Tab 4 shows each oligotype with its representative sequence; prevalence; mean abundance in samples in which it was detected; mean abundance over all oral samples; best match in HOMD along with the percent identity to that match and fraction of the sequence along which the percent identity applies; relative abundance of the oligotype at each sampling site; habitat preference; and comparison with LEfSe identification of biomarkers (1). The symbol "-" in the oligotype and in the representative sequence indicates a gap rather than a nucleotide at that position. Oligotypes are listed in order of their mean abundance across all oral sites. Tab 5 reports the observations (measured in counts, i.e. number of reads) of each oligotype in each sample from every subject.

1. Segata N, et al. (2011) Metagenomic biomarker discovery and explanation. *Genome Biol* 12(6):R60.

Dataset S2. Oligotyping results for V3-V5 data

Dataset S2

Tab explanations are the same as for Dataset S1.

Dataset S3. Mapping of oligotypes onto Human Oral Microbiome Database species

Dataset S3

V1-V3 oligotypes (tab 1) and V3-V5 oligotypes (tab 2) are grouped and tallied according to their mapping onto Human Oral Microbiome Database (HOMD) species.

Dataset S4. Oligotypes listed by habitat in which their relative abundance is most statistically significant

Dataset S4

For V1-V3 (tab 1) and V3-V5 (tab 2) oligotypes are listed and identified with their closest match in HOMD. The mean relative abundance for each oligotype across all subjects is shown for each sampling site (columns G–P). Column Q ("ABUNDANT_IN") shows the site(s) in which the mean relative abundance of that oligotype was most significantly greater than its abundance at other sites, as described in *Materials and Methods*. Columns R and S give the t-probability, Bonferroni corrected and raw, and column T gives the *t* statistic. W and the W-statistic are shown in columns U and V. Columns W-Z show the results of LEfSe analysis (1).

1. Segata N, et al. (2011) Metagenomic biomarker discovery and explanation. *Genome Biol* 12(6):R60.

Dataset S5. V3-V5 oligotypes detected preferentially in subgingival plaque and supragingival plaque shown in Fig. 6

Dataset S5

Oligotypes in lines 4–47 were identified as plaque-associated ($P < 0.01$) and at least threefold more abundant in subgingival plaque (SUBP) than in supragingival plaque (SUPP); oligotypes in lines 52–88 were identified as plaque-associated ($P < 0.01$) and 1.5-fold more abundant in SUPP than in SUBP. The oligotype name, prevalence in oral sites, mean abundance, and names, percent identity, and Human Oral Taxon (HOT) identification numbers of the corresponding HOMD taxa are listed in columns A through G, with the mean abundance at each oral site in columns H through P.

Dataset S6. All V3-V5 oral samples in which stool-characteristic oligotypes were detected at >1% relative abundance

Dataset S6

Column A lists the sample identifier (subject number followed by abbreviation for sampling site), and columns B through W report the relative abundance in these samples of the 22 oligotypes characteristic of stool and detected at >1% abundance in any oral sample. The stool samples from the same subjects, where available, are shown for comparison. Most stool oligotypes found at >1% abundance in oral samples are also found in the stool sample from the same subject; exceptions are shown in lines 42–50.