Supplemental Appendix 1

1. Supplemental Methods

1.1 Methods for genome comparisons of *Porphyromonas gingivalis* strains

*P. gingivalis* -- *16S rRNA sequence similarity*. 16S rRNA sequences for the 17 *P. gingivalis* strains in the NCBI database (GCF_001444325.1, GCF_000467995.1, GCF_000467795.1, GCF_001314265.1, GCF_000010505.1, GCF_000739415.1, GCF_001263815.1, GCF_000007585.1, GCF_001274615.1, GCF_001297745.1, GCF_000467975.1, GCF_000467955.1, GCF_000380305.1, GCF_000467835.1, GCF_000467815.1, GCF_000270225.1, GCF_000503975.1) were extracted and aligned using MAFFT (7). The percent identity matrix was generated using Clustal2.1.

*P. gingivalis* -- *MLST gene sequence similarity*. Ten MLST genes (8) representing 190 sequences/alleles (AJ555632 to AJ555821 in Nucleotide DB for genes *coa, dnaK, ef-tu, ftsQ, gdpxJ, hagB, mcmA, nah, pga, recA, pepO*) were extracted from the genome assemblies. Multilocal sequence analysis of *P. gingivalis* indicates frequent recombination. The MLST genes were aligned to the 17 full genomes using BLAST. The first gene hits for each genome were concatenated into single long sequence for each strain. Concatenated genes were aligned using MAFFT and a sequence identity matrix was generated with Clustal2.1.

1.2 Methods for Figure 1: Analysis of Evolutionary relationships of taxa within the bacterial families Enterobacteriaceae and Porphyromonaceae based on full-length16S rRNA gene sequences

All 16S rRNA sequences from the 99% clustered and aligned green genes 13.5 database for *Escherichia, Salmonella, Shigella, Klebsiella, Tannerella* and *Porphyromonas* were pulled and the 16S rRNA sequences of *Yersinia nurmii* and *Bacteroides nordii* were added as outgroups. Phylogenetic trees were built with QIIME version 1.9 using the *fasttree* tree building algorithm. The resulting tree files were loaded into FigTree and branches were colored by genus.

1.3 Methods for Figure 2: Reconstructed taxonomic profile for archaeological dental calculus using the QIIME, metaBIT, MIDAS, and MALT pipelines

Calculus sample C214 (19) was analyzed using MALT (NT database), MIDAS, METABIT, and QIIME. MALT, MIDAS, and METABIT were run with default parameters; QIIME closed-reference OTU picking parameters were modified for short reads by setting --max_accepts to 500 and --max_rejects to 500. Raw output files were used to calculate phylum proportion for the eleven top oral phyla with all others collapsed into "other".

1.4 Methods for Figure 3 and Supplemental Data 2: Analysis of Pathogens and their close relatives in the environment

16S rRNA gene sequences were obtained from Genbank for the following selected pathogens: *Bacillus anthracis* (NC_003997), *Bordetella pertussis* (NC_002929), *Clostridium tetani* (NC_004557), *Clostridium botulinum* (NC_009698), *Clostridioides difficile* (NC_009089),...
Mycobacterium tuberculosis (NC_000962), Mycobacterium leprae (NC_002677), Salmonella enterica (NC_003197), Shigella dysenteriae (NC_007606), Vibrio cholerae (NC_002505), Yersinia pestis (NC_003143). The sequences were concatenated and used as reference database for clustering. Unaligned 16S rRNA gene sequences for Bacteria and Archaea were retrieved from the RDP website to be used as queries. QIIME was used to cluster queries together with references at 97% identity. For each reference, the isolation sources of all clustering Bacteria and Archaea that are classified as uncultured were collected and reported as counts.

1.5 Methods for Figure 4: Lack of signal for highly abundant endogenous bacteria due to database bias

Five shotgun libraries generated from a medieval calculus sample (G12) (18) were obtained from the NCBI SRA database using fastq-dump (SRR957742-SRR957746) and concatenated. The merged fastq files were aligned using MALT v038 (4) to two different databases: 1) all NCBI RefSeq bacterial genomes from December 2015; 2) the same but without the genome for Propionibacterium propionicum F0230a (note proposed name change to Pseudopropionibacterium propionicum (15)). Reads with an identity score of <85% were removed. For LCA binning, the top percent parameter was set to 1. The alignments were visualised in MEGAN6 (5) with the tree collapsed to species-level, and the summed number of hits for parent and all subtree nodes were transformed by log scale.

1.6 Methods for Supplemental Table 1: Mismapping of environmental contaminants to M. tuberculosis and Y. pestis pathogen reference genomes

The representative genomes of environmental relatives of Yersinia pestis and Mycobacterium tuberculosis (NC_000962.3, NC_015848.1, NC_022663.1, NC_002944.2, NC_015576.1, NC_008596.1, NC_014814.1, NC_003143.1, NZ_CP008943.1, NC_008800.1, NZ_CP009801.1) were downloaded from the NCBI RefSeq Genome database, and converted to simulated sequencing reads of a length of 100 bp with 1bp tiling. The simulated reads were mapped to the pathogen reference genomes using BWA aln v0.7.12 (9) using default parameters. Mapping statistics were calculated using samtools flagstat (10) and qualimap v2.2 (12), with files for qualimap being prepared using samtools sort, index and Picard Tools (v1.140) cleanSam (http://broadinstitute.github.io/picard).

1.7 Methods for Supplemental Figure 1: Mismapping of environmental (soil and ocean) and host-associated (saliva) metagenomic sequences to selected obligate pathogens

Ocean (ERR315858), soil (ERR1041384) and saliva (SRR062396) shotgun metagenomes were obtained from the EBI-ENA and NCBI-SRA databases, and processed using the EAGER pipeline (v1.92.20; Peltzer et al. 2016). Forward and reverse reads were merged with the ClipAndMerge module and mapped to fifteen pathogen genomes using BWA (v0.7.12) aln default settings. Mapping statistics were then taken from the ReportTable module of EAGER.