SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Alpha diversity indexes across different spatial locations of aligner (**a-c**) and plaque (**f-h**) samples. P-values (p) are shown for Wilcoxon paired test.

Supplementary Figure 2. Alpha diversity indexes across different spatial locations of aligner (**a-b**) and plaque (**c-d**) samples. P-values (p) are shown for Wilcoxon paired test.

Supplementary Figure 3. A) Alpha diversity indexes of aligner samples over three consecutive nights. P-values of Wilcoxon test comparison are displayed. Abbreviations: N=number of samples. **B)** Uniform Manifold Approximation and Projection (UMAP) based on Euclidean distances of center-log ratio (CLR) counts on antimicrobial resistance genes. Each point represents a sample from the long-term study; only paired comparisons across timepoints are shown.

Supplementary Figure 4. A) Decontaminated sequencing information (Gigabasepairs) by specimen type and disease group. P-corrected values are shown for Wilcoxon paired tests. In each boxplot, the median value is displayed. **B)** Percentage of human contamination by specimen type and disease group. P-corrected values are shown for Wilcoxon paired tests. In each boxplot, the median value is displayed. **C)** Percentage of taxonomically assigned reads by specimen type and disease group. P-values are shown for Wilcoxon paired tests. In each boxplot, the median value is displayed. **D)** Heatmap displaying centered log-ratio (CLR) transformed raw counts for differentially abundant genera across specimen types (details on **Supplementary Table 5**). Genera were clustered into three groups based on k-means clustering. Oxygen tolerance and gram stain annotations were curated or predicted from the BacDive Metadatabase (see Methods).

Supplementary Figure 5. Heatmap displaying centered log-ratio (CLR) transformed raw counts for differentially abundant species across specimen types (details on **Supplementary Table 5**). Species were clustered into three groups based on k-means clustering. Oxygen tolerance and gram stain annotations were curated or predicted from the BacDive Metadatabase (see Methods).

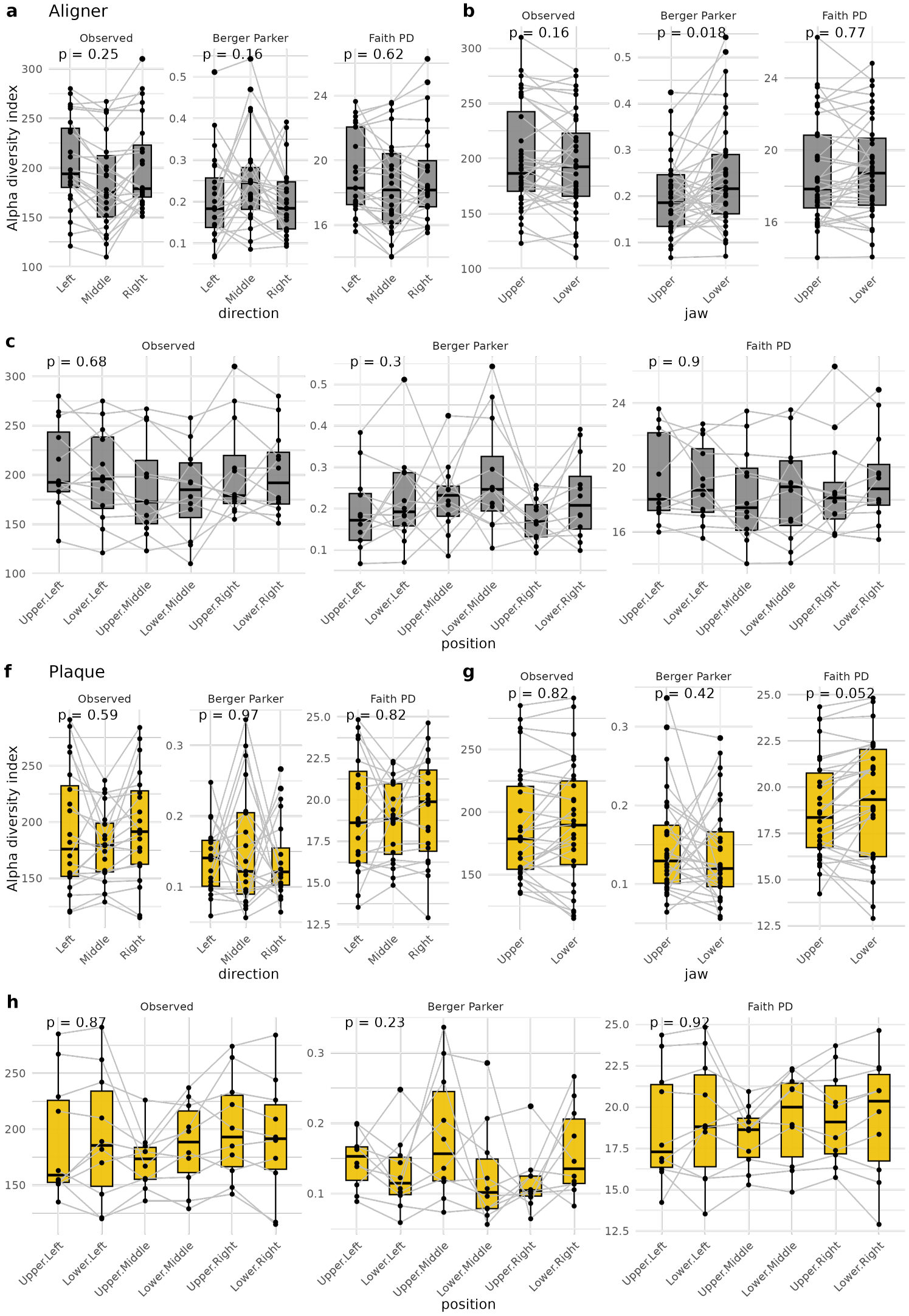
Supplementary Figure 6. Principal Coordinate Analysis (PCoA) using Bray-Curtis distances on raw counts for caries (**a**) and periodontitis (**b**) across specimen types. Each small dot represents a sample included in the cross-specimen cross-disease study.

Supplementary Figure 7. **A)** Distribution of the number of scaffolds in each sample at various length limits. On top of each boxplot, the total number of scaffolds surpassing the length threshold is displayed (Number of samples = 1240). **B)** Total number of representative species genome bins (SGBs) visualized by specimen of the initial sample and novelty. **C)** Distribution of predicted antimicrobial resistances (AMR) subclasses by putative host species. Only labels for the 20 most frequent AMR subclasses and species are displayed. Both AMR classes axis and species legend are ordered by decreasing frequency.

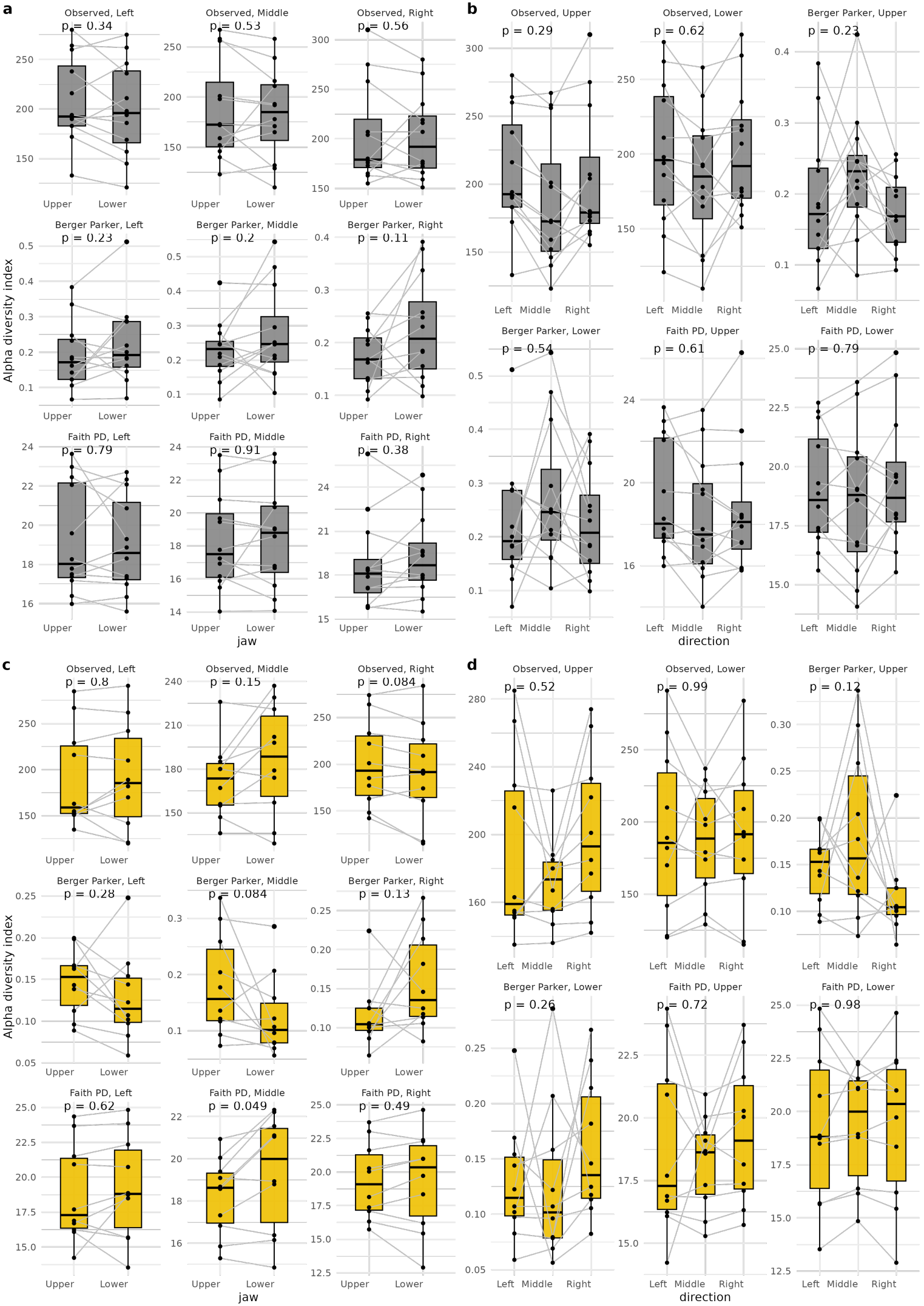
Supplementary Figure 8. Heatmap displaying centered log-ratio (CLR) transformed RPKM values for all 103 differentially abundant biosynthetic gene cluster families (GCFs) across oral status and specimen types (details on **Table 3**). GCFs were clustered into three groups based on k-means clustering.

Supplementary Figure 9. A) Network visualization of differentially abundant biosynthetic gene cluster families (GCFs) and their putative bacterial hosts in the oral microbiome (details on **Table 3**). All GCFs with predicted hosts are displayed. Rectangular nodes represent GCFs, while diamond nodes represent bacterial species (putative hosts). GCF nodes are colored according to the specimen type in which they were found differentially abundant. Connections between GCFs and bacterial species represent predicted host-BGC relationships. Edge color indicates log fold-change (logFC), whereas edge width corresponds to statistical significance (adjusted p-value), where thicker edges indicate lower p-values. GCFs connected by multiple edges to the same species indicate that the GCF was identified as differentially abundant in multiple contexts (e.g. different diseases or in different specimen types). Candidates selected for wet-lab validation are marked with stars. **B)** Number of predicted biosynthetic gene clusters (BGCs) per representative species genome bin (SGB). Only SGBs with more than 14 predicted BGCs are shown. Each SGB is labeled with the corresponding assigned species name and its unique identifier (y-axis). If an SGB could not be assigned to a known species, it is labeled with the prefix 'bin_' followed by its unique identifier. BGC types are annotated for the 21 most abundant categories, representing 90% of the total predicted BGCs.

Supplementary Figure 1

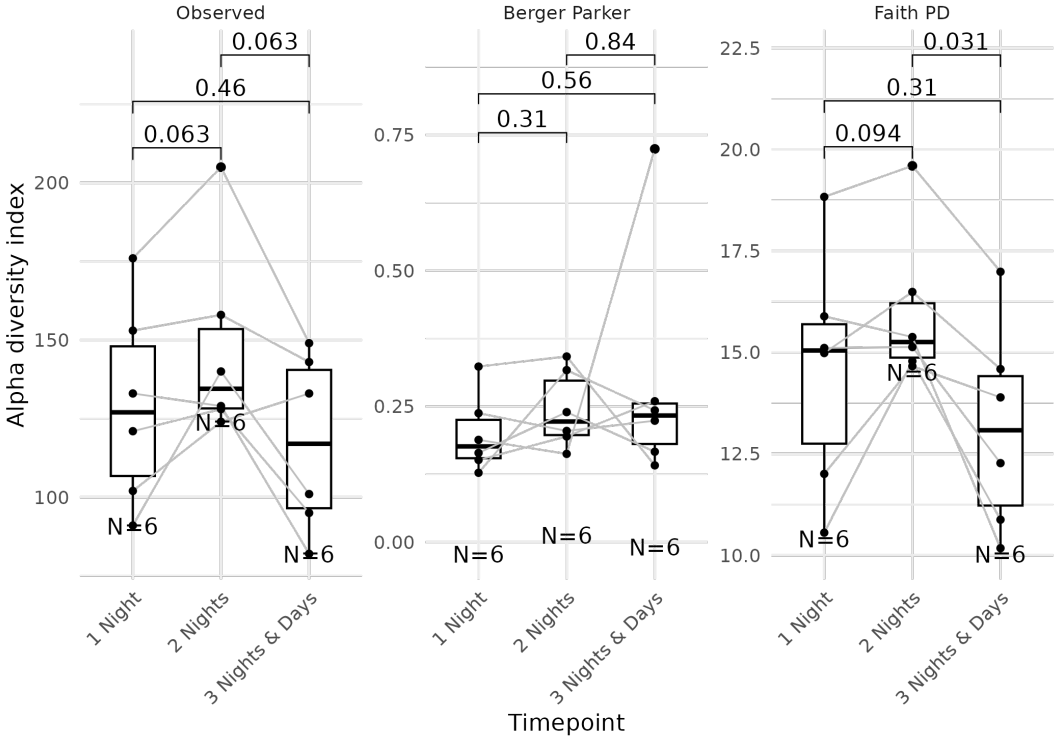


Supplementary Figure 2

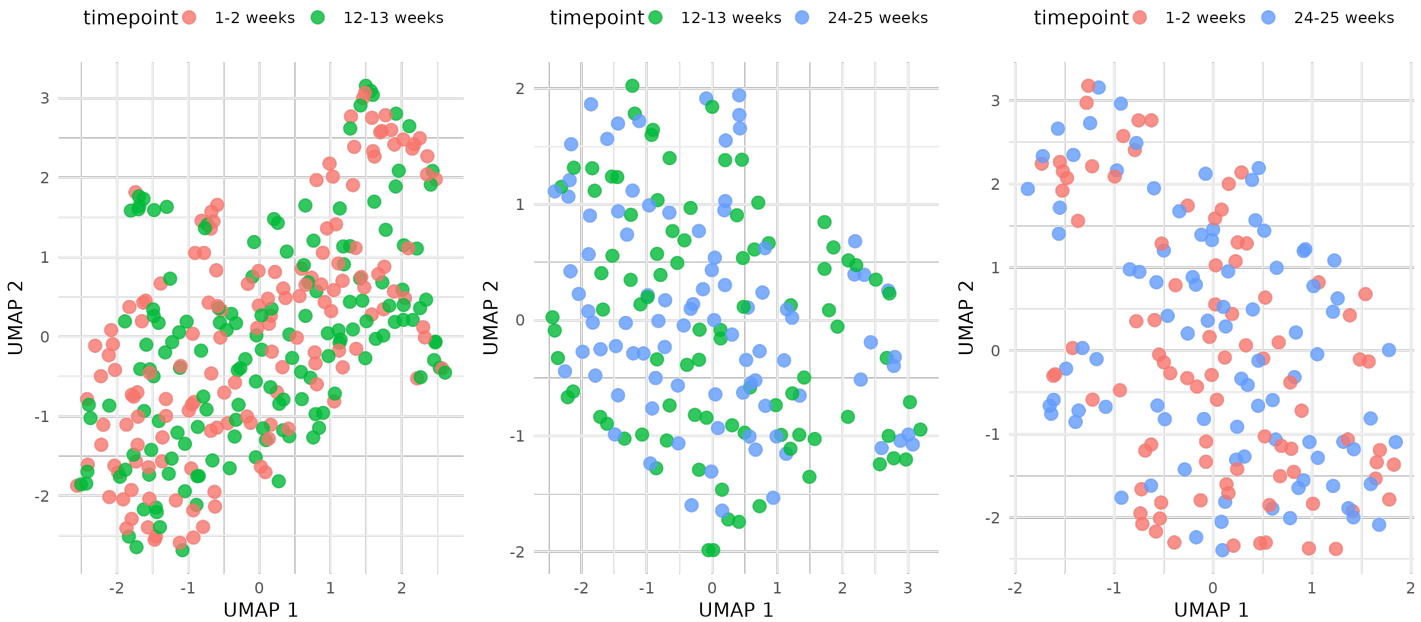


Supplementary Figure 3

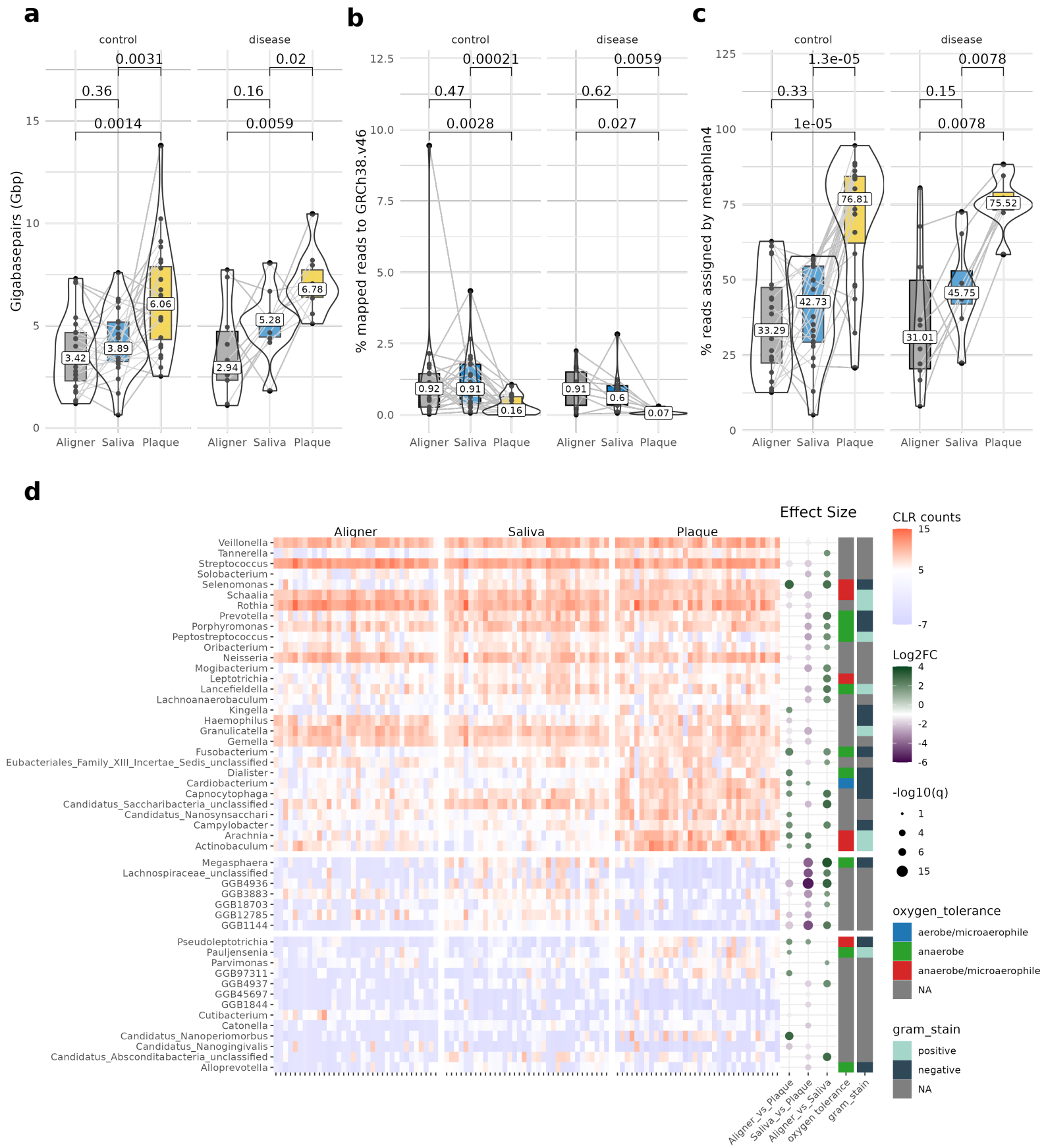
a



b

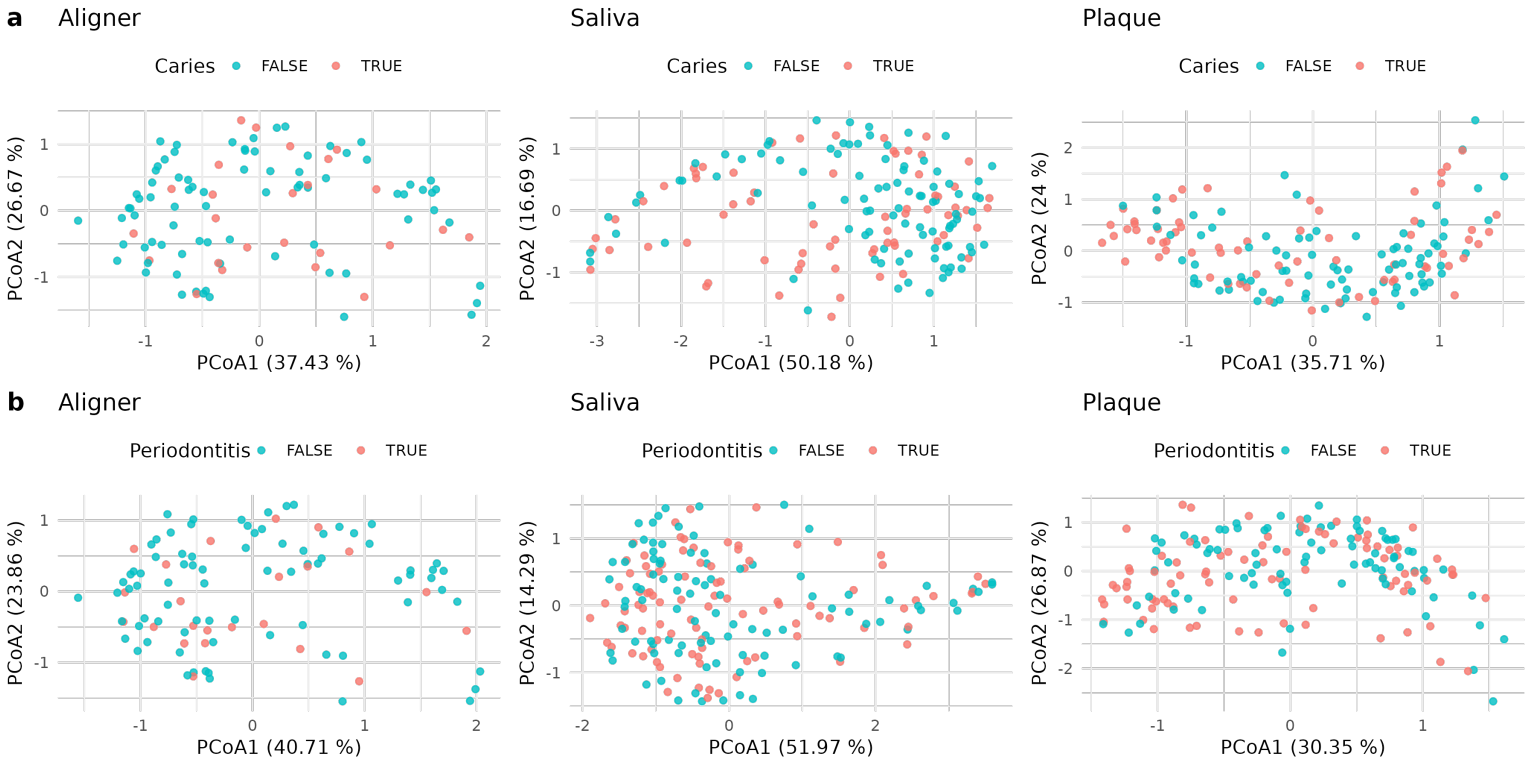


Supplementary Figure 4

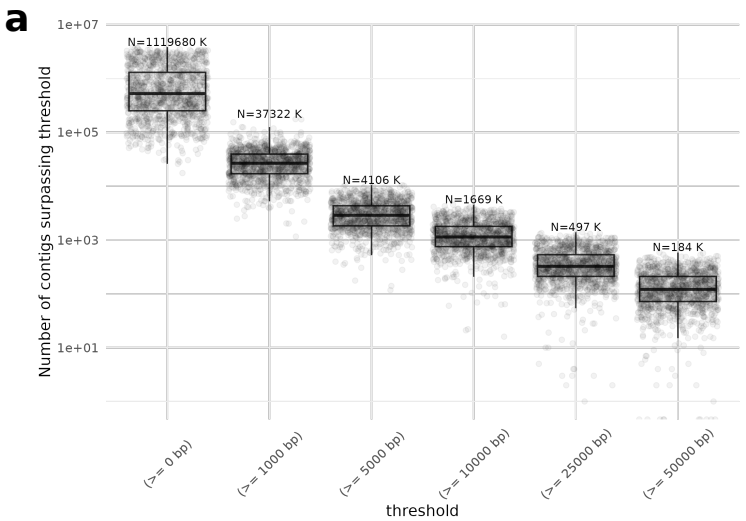


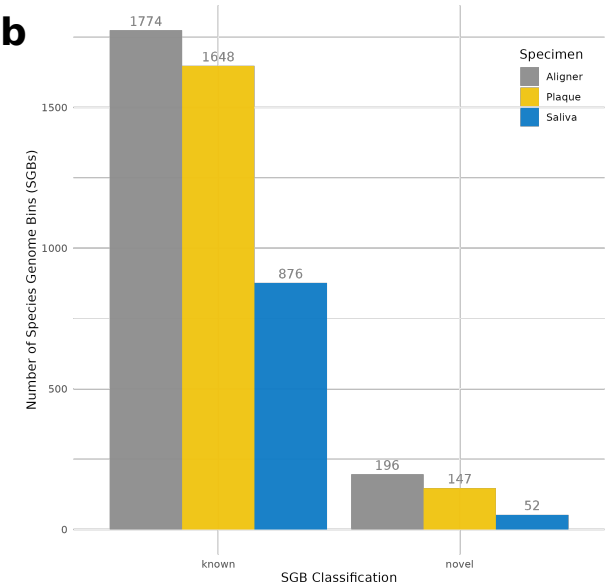
Supplementary Figure 5

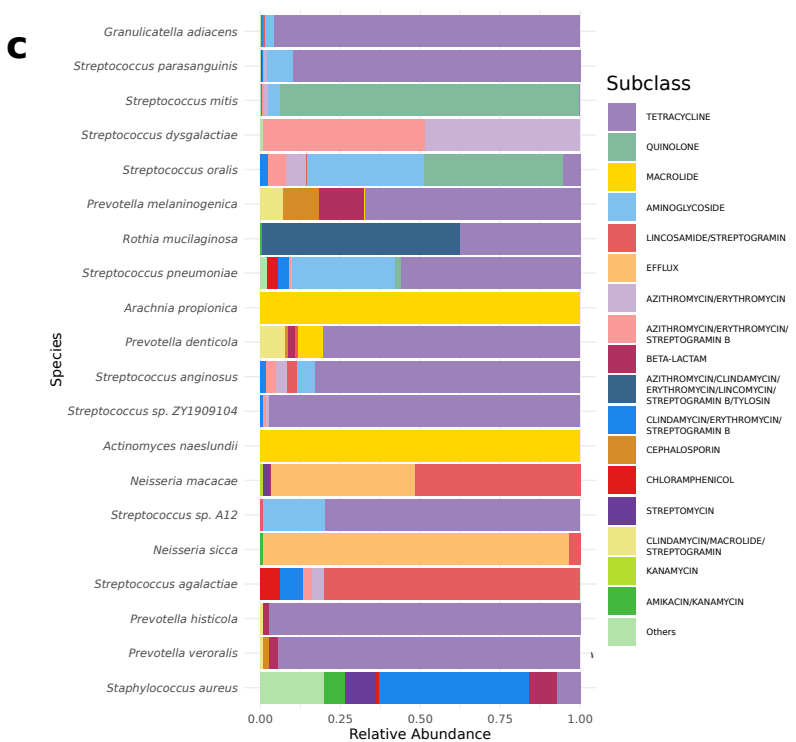
Supplementary Figure 6



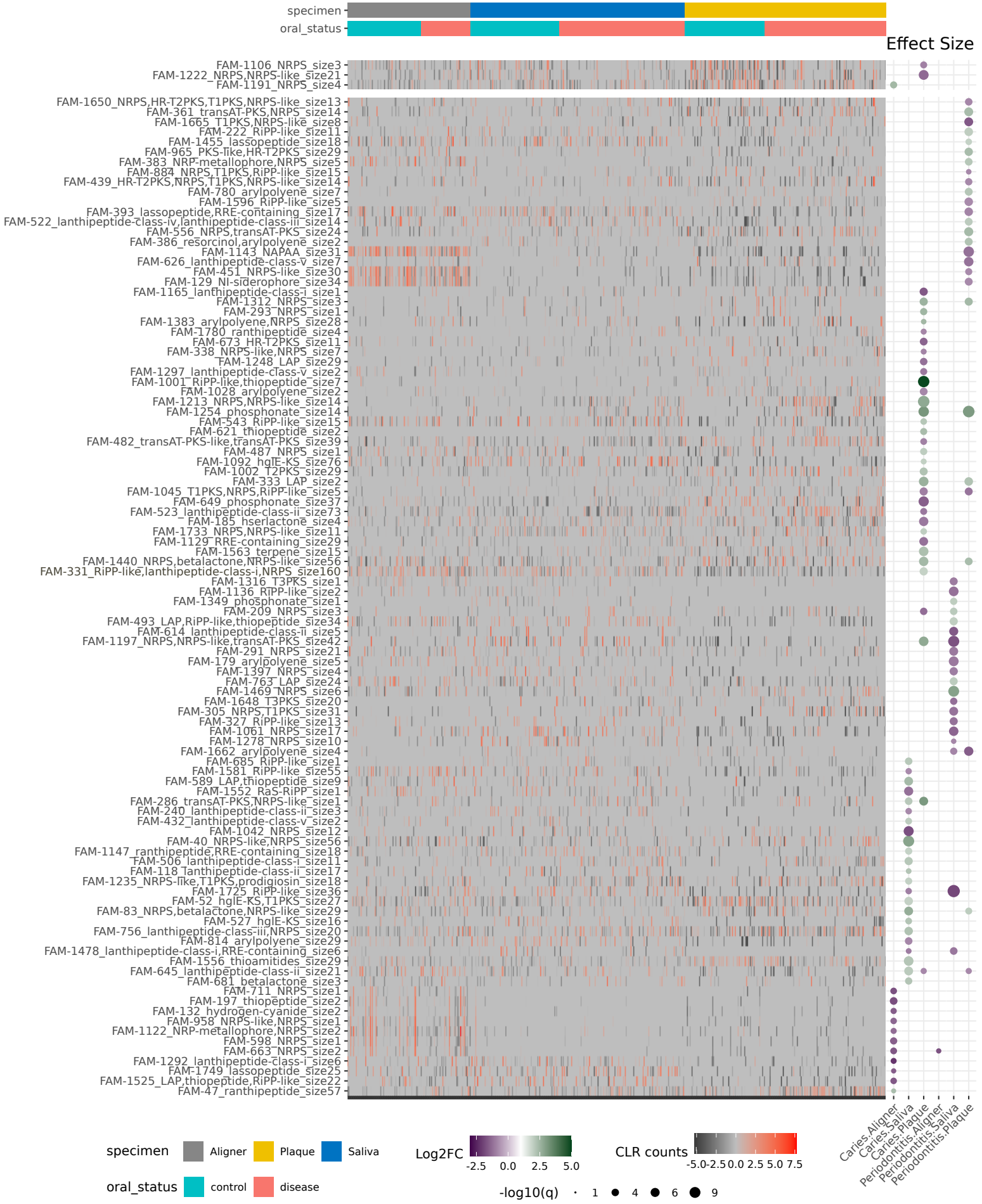
Supplementary Figure 7





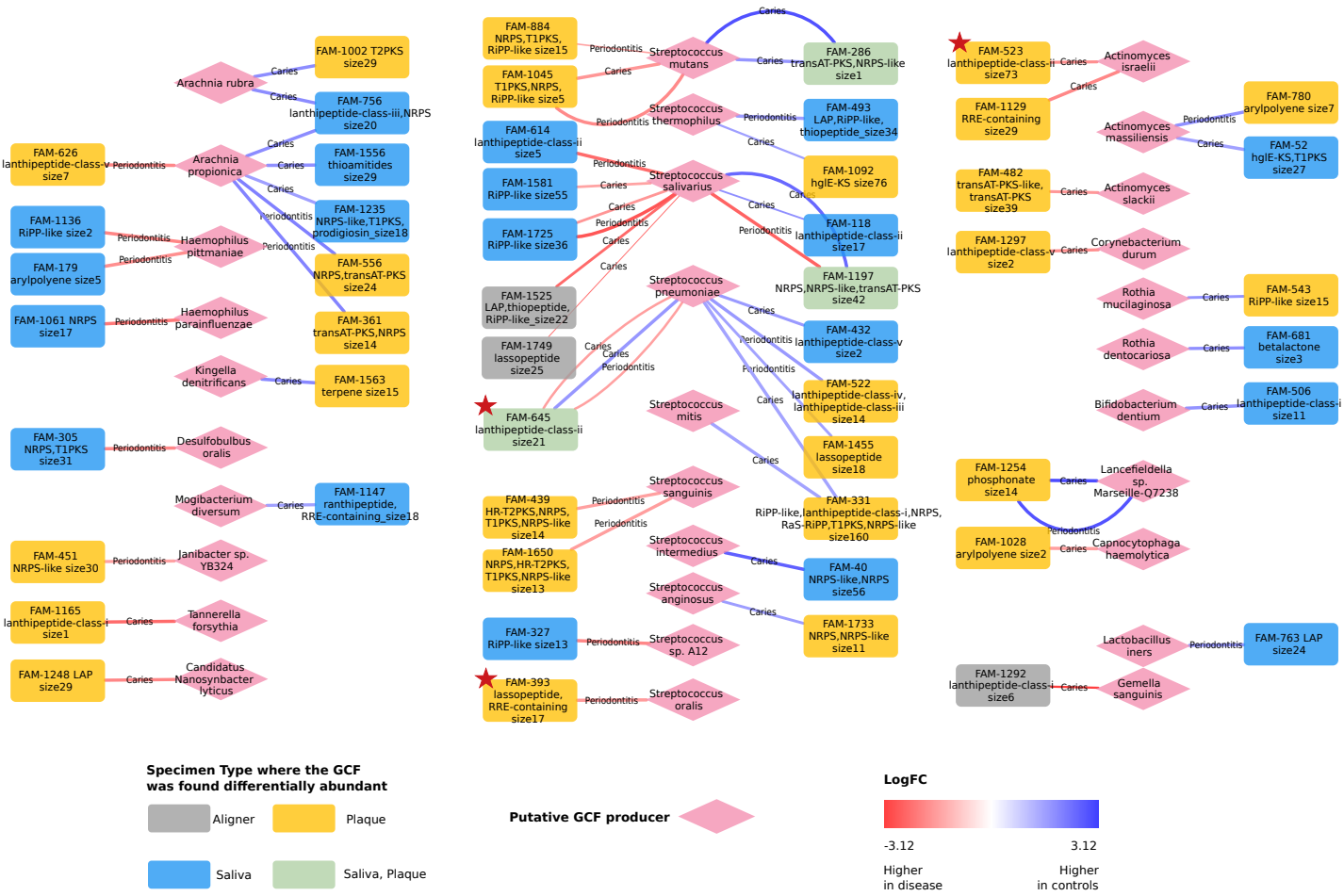


Supplementary Figure 8



Supplementary Figure 9

a



b