Supplementary Information

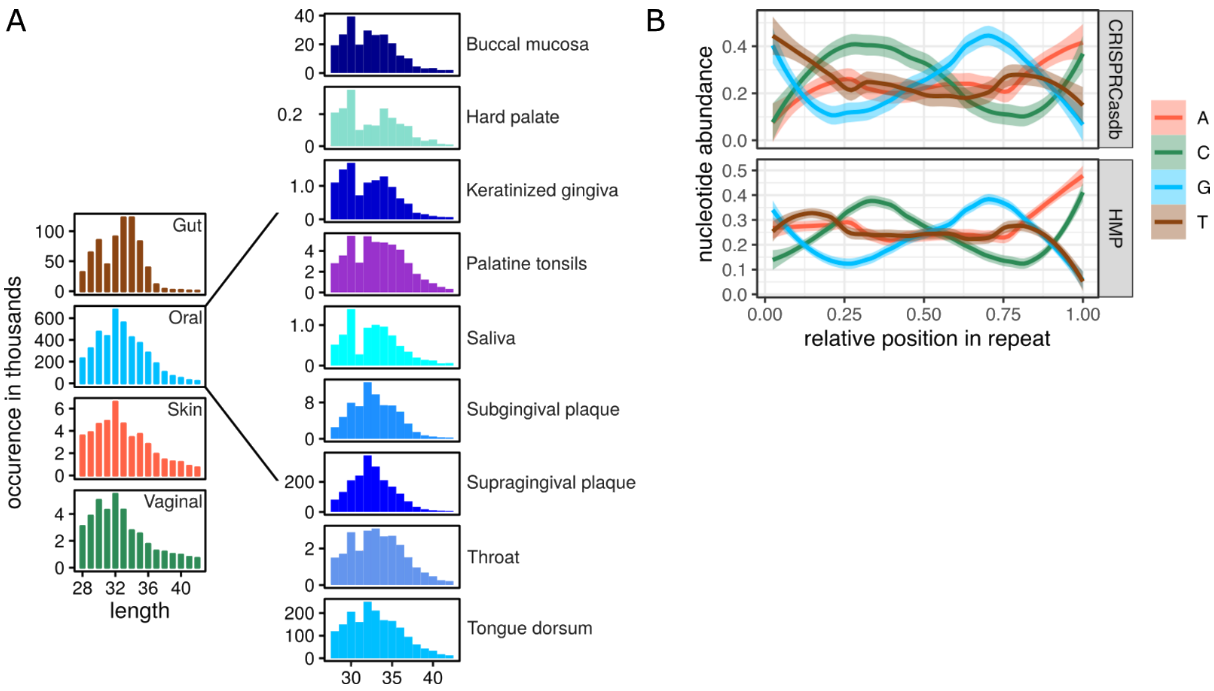

Figure S1: Spacer length distribution and nucleotide abundance; Related to Figure 1 and Figure 2. A) HMP1-II spacer length distribution for oral body sites. Length of spacer sequences by body area (left) and for body sites of the oral area (right). B) Nucleotide distribution of repeat sequences from HMP and CRISPRCasdb exhibit a high similarity. Repeats with length from 23 to 39nt are shown.

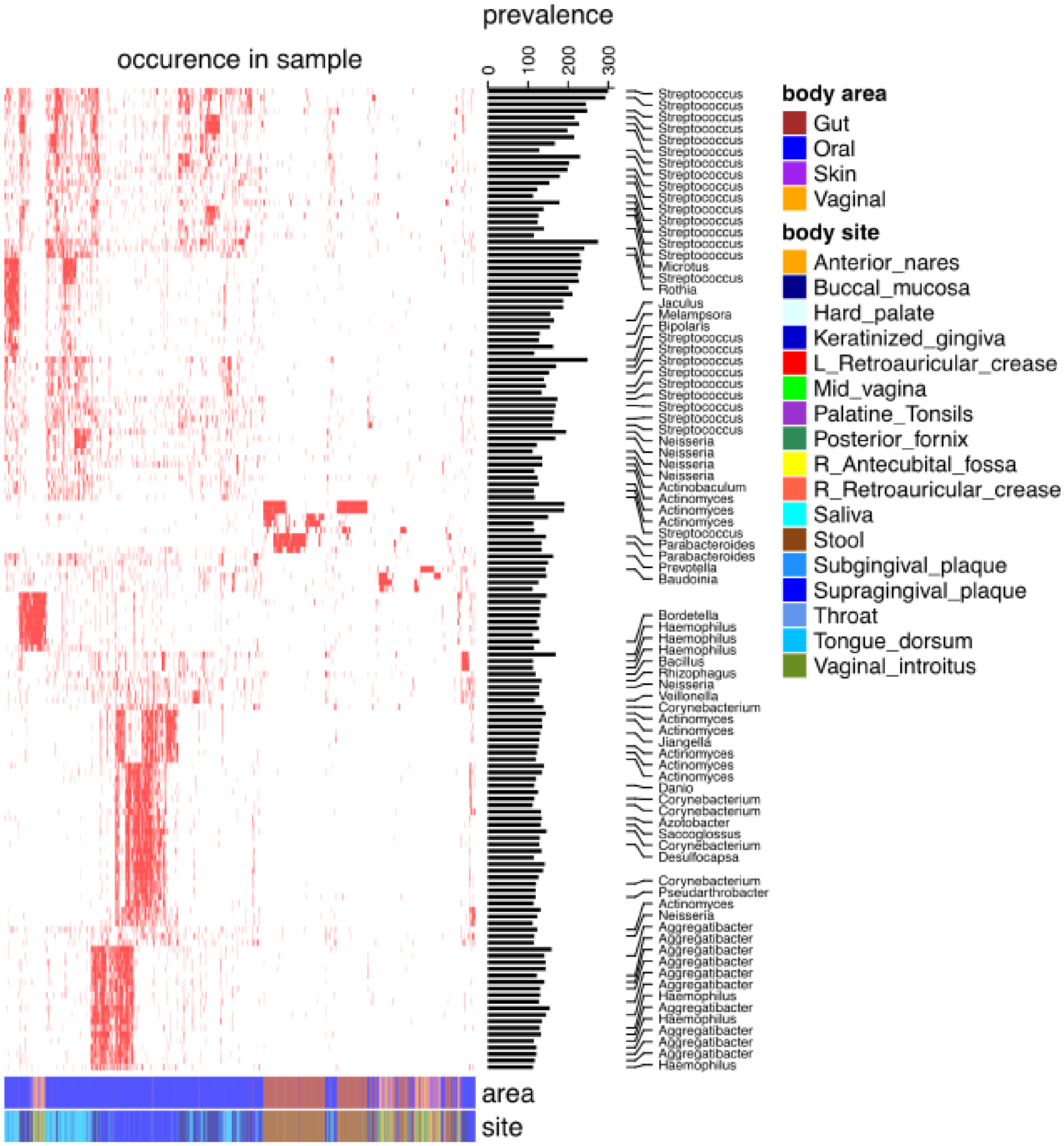

Figure S2: Occurrence of the 150 most prevalent spacers found in HMP1-II; Related to Figure 2. red: present in sample, white: absent in sample. Columns and rows are sorted by single-linkage clustering. Taxonomic annotation is based on combined mapping to UniRef90 groups and to the sample's assembly.

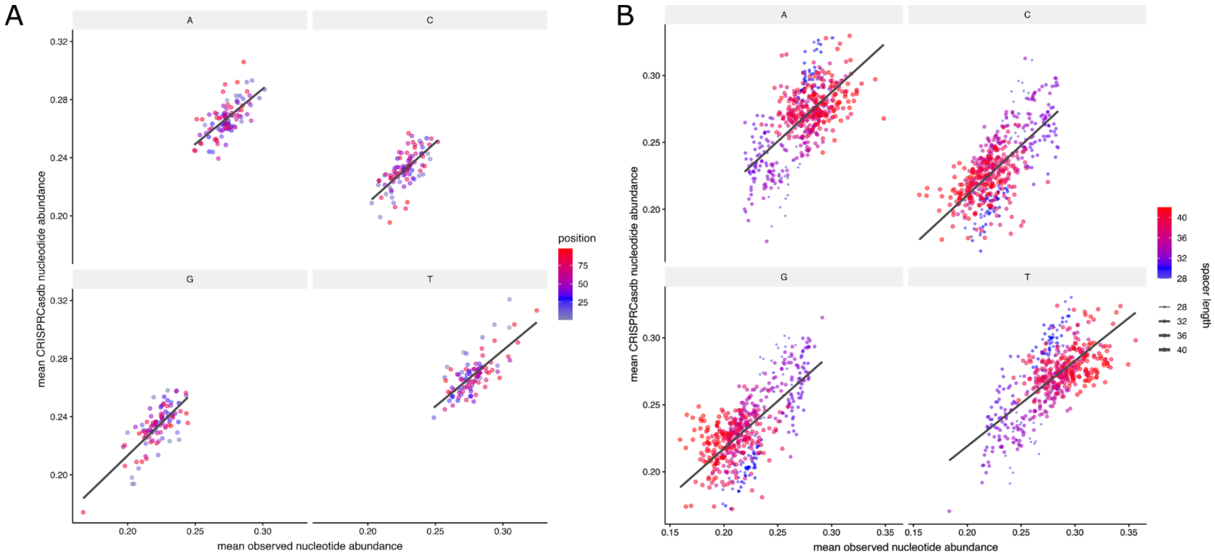

Figure S3: Agreement of relative abundances of observed spacers and CRISPRCasdb spacers; Related to Figure 1. A) Binned relative abundance per relative position of CRISPRCasdb and HMP1-II spacer show high correlation. B) Mean relative nucleotide abundances of spacers found in HMP1-II and CRISPRCasdb. Color and point size indicate spacer length. Line is the linear fit.

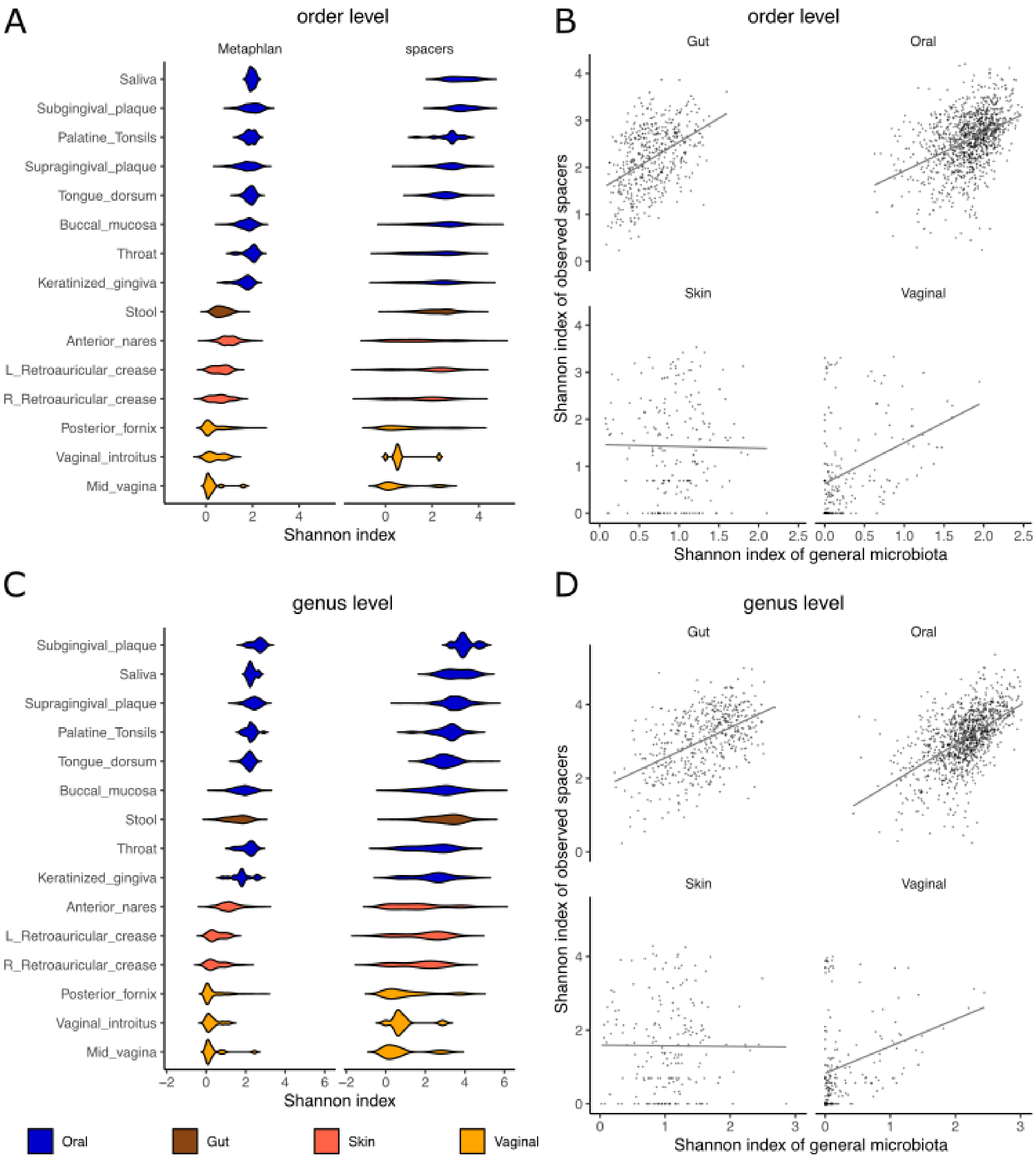

Figure S4: **Alpha diversity of general microbiota and spacer sequences; Related to Figure 5.** A,C) Alpha diversity by body site on order and genus level, respectively; B,D) correlation between alpha diversity of general microbiota for order and genus level, respectively.

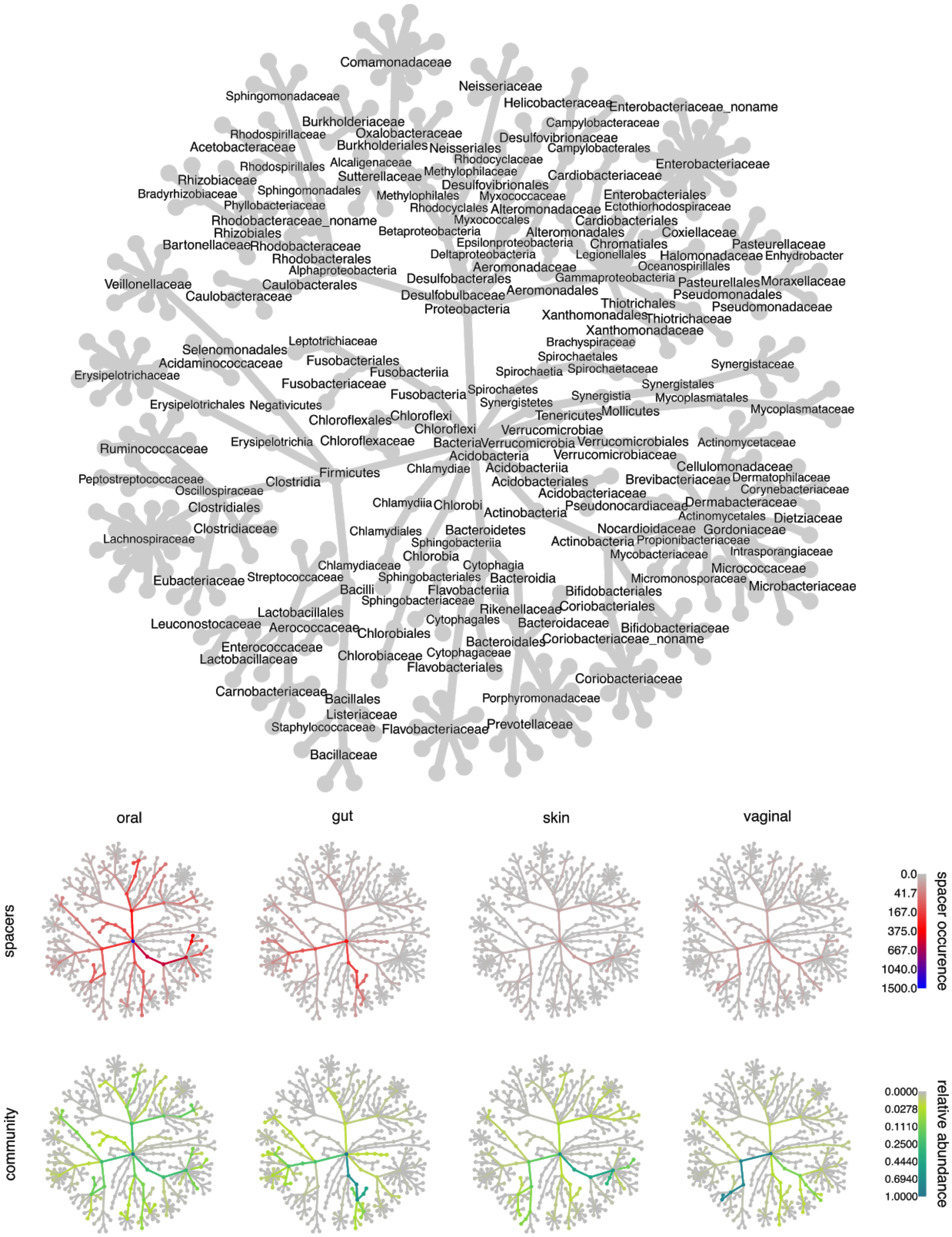

Figure S5: Phylogenetic trees colored by the abundance of members of the general microbiota (MetaPhlAn) and taxonomic assignments of HMP1-II spacers (spacer abundance); Related to Figure 3. Left panel provides a legend for individual phylogenetic trees by body area.

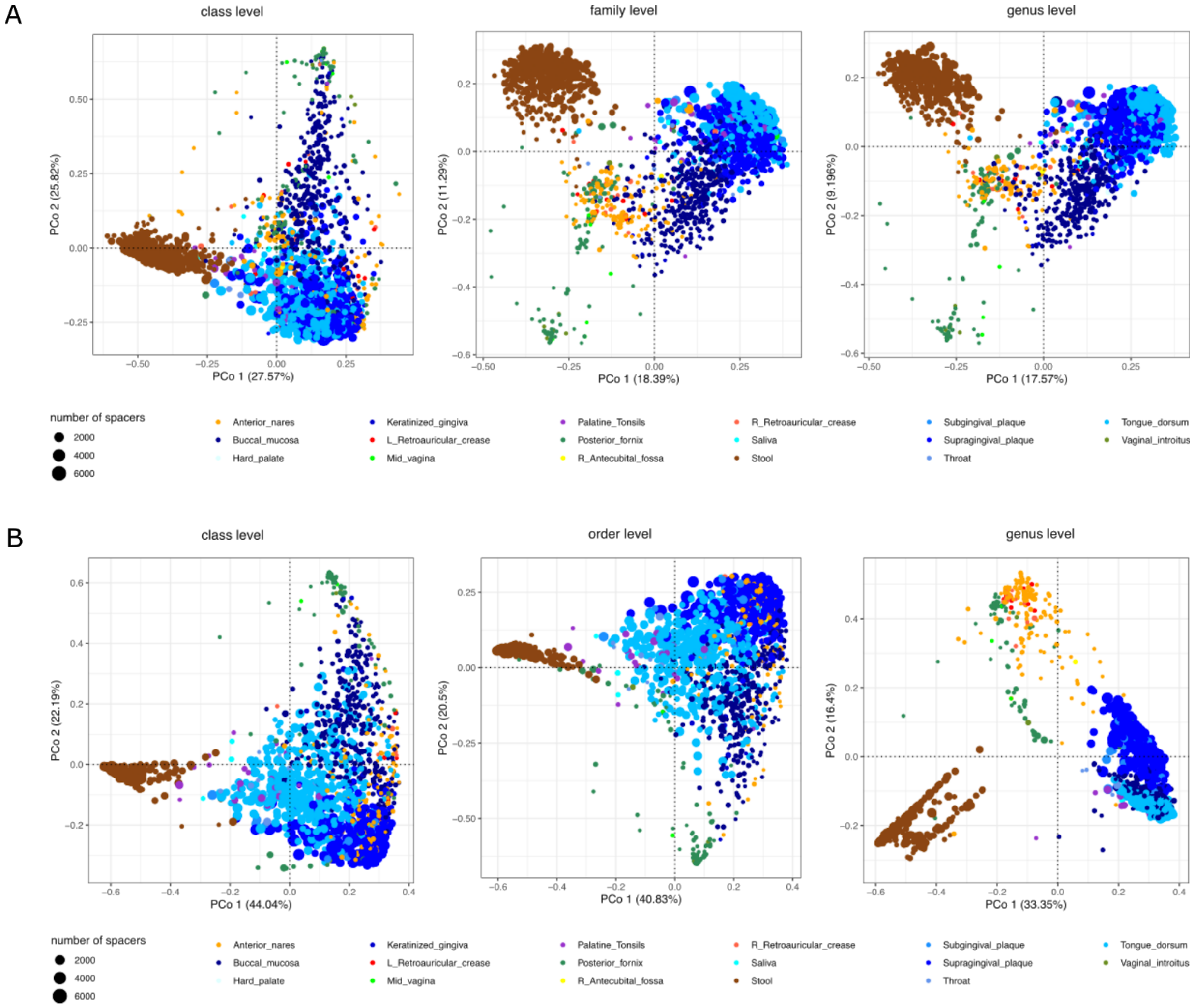

Figure S6: PCoA of spacer taxonomy (Bray Curtis) on different taxonomic levels; Related to Figure 3. Size of points indicate the number of spacer (cluster representatives at 80% identity). B) PCoA of Bray–Curtis dissimilarity of the general microbiota (MetaPhlAn). Size of points indicate the number of spacers (cluster representatives at 80% identity).

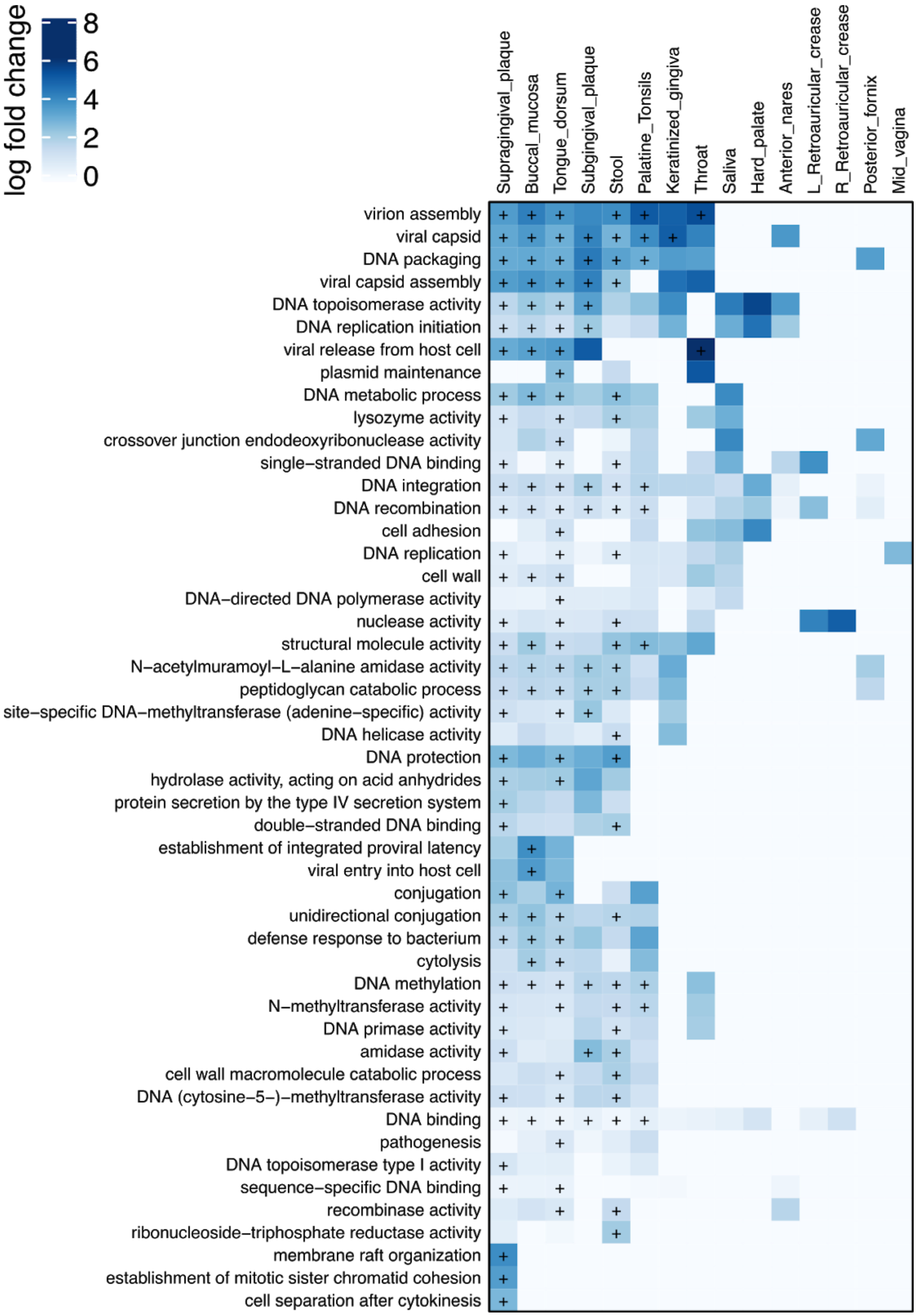

Figure S7: Functional enrichments within predicted spacer targets for mapping against the UniRef90 database; Related to Figure 4.

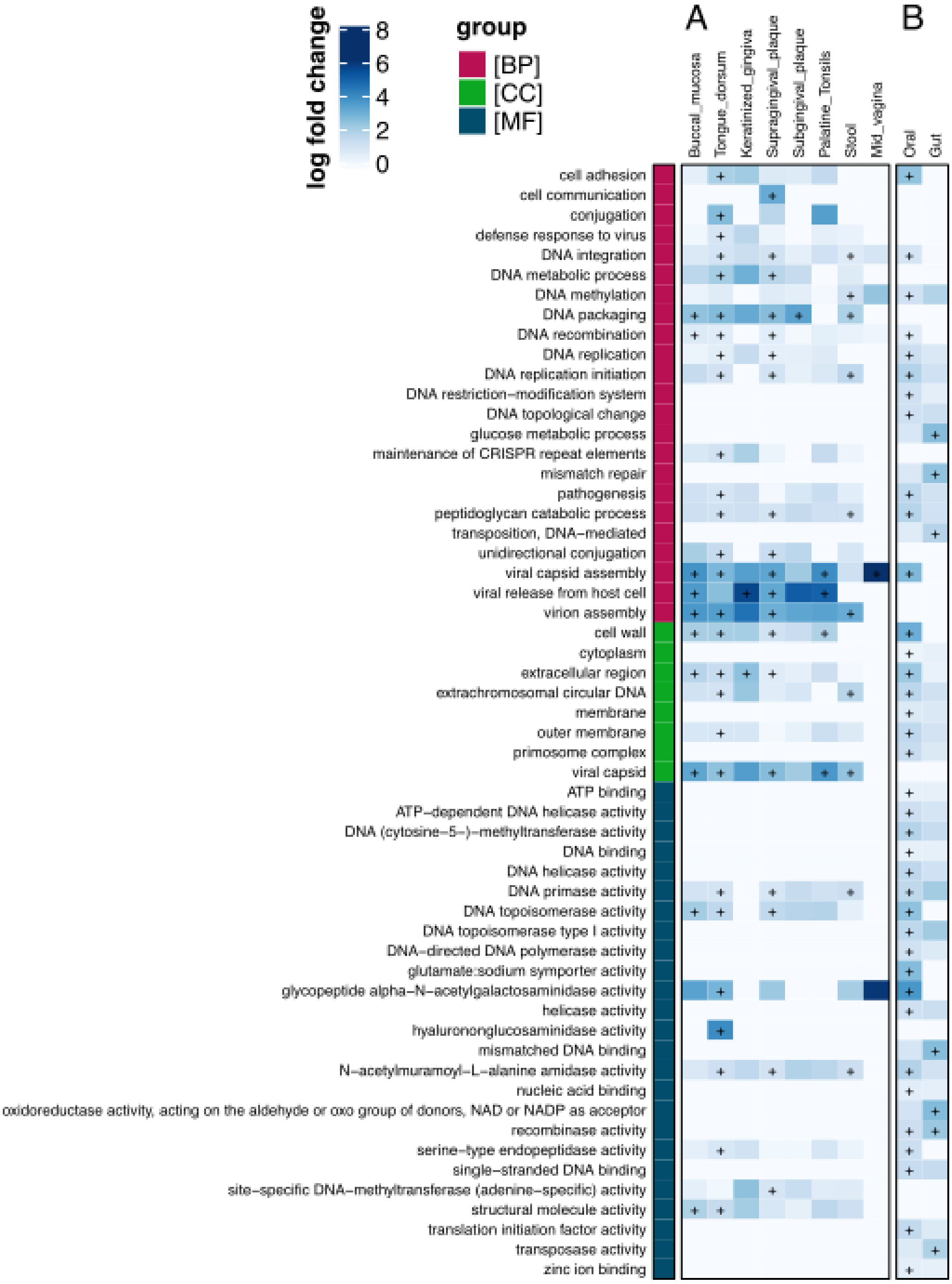

Figure S8: Functional enrichments within predicted spacer targets; Related to Figure 4. Log fold enrichment of Gene Ontology (GO) terms for all spacer targets within sample assemblies per body site. Terms shown here achieved at least one FDR corrected *q* value < 0.05 based on a Fisher test of enriched UniRef90 terms with respect to the overall contig annotation of the site.

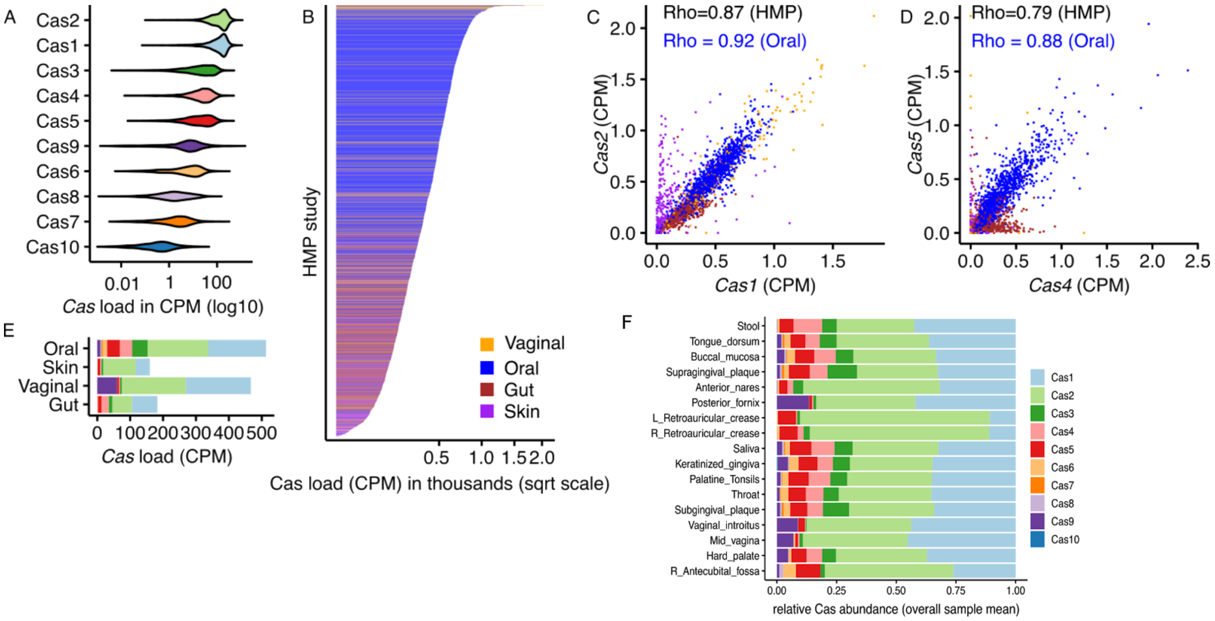

Figure S9: Visualization of cas gene abundance; Related to Figure 5. A) Violin plots of *cas* load (log10 scale). B) barplot of individual samples sorted by their mean cas abundance. C) *cas1* and *cas2* is highly correlated, especially for samples taken from the oral site. D) Correlation of *cas4* and *cas5* is high on samples taken from the oral site. E) *cas* abundance stratified by body area. F) Overview of relative *cas* gene abundance among body sites.

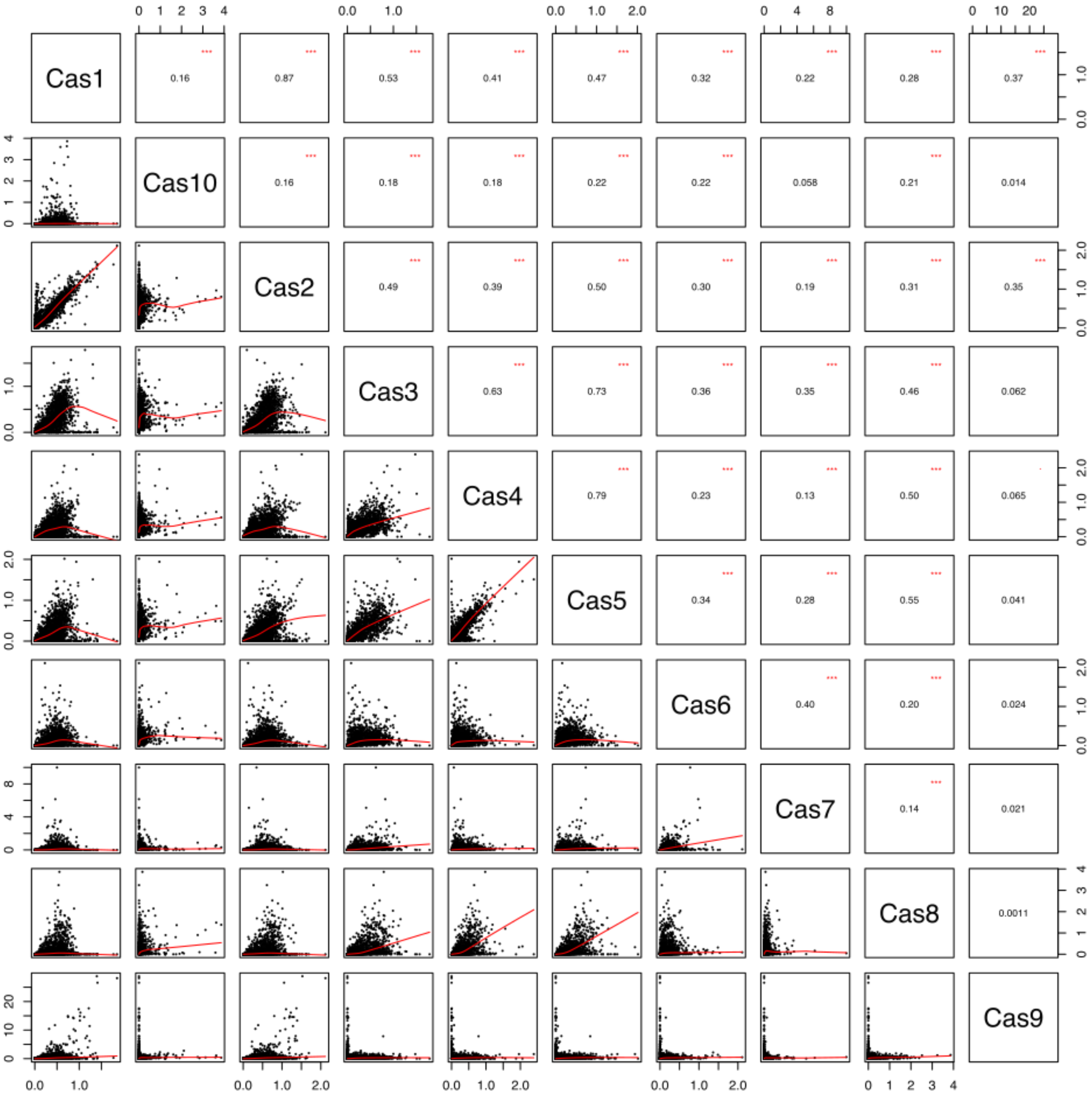

Figure S10: Pairwise correlations of *Cas* gene abundance; Related to Figure 5

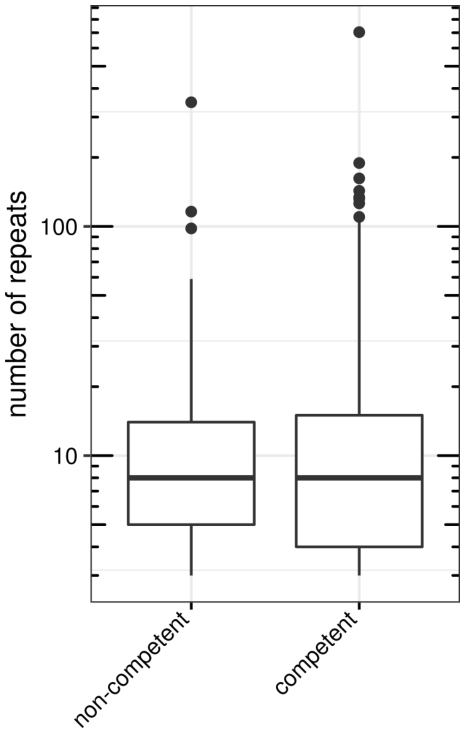

Figure S11: Presence of natural competence proteins; Related to STAR methods. No associations between the presence of natural competence function (measured by the presence of a competence-associated protein family) and the number of CRISPR repeats within a genome was found.

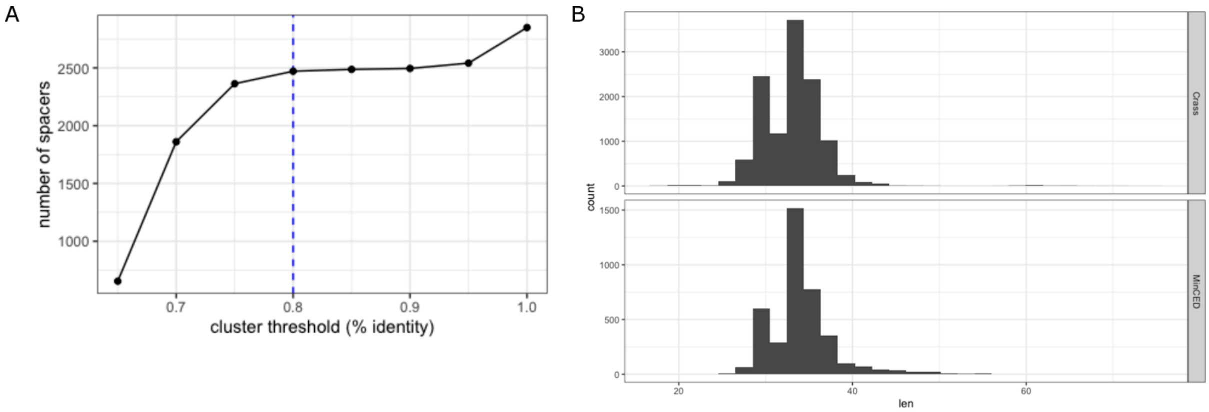

Figure S12: Sensitivity and uniqueness of spacer retrieval at increasing sequence similarity thresholds; Related to STAR methods. A) Number of clusters found using the Crass with arguments used in our study for increasing similarity cutoffs. Possible cutoff values which are located on the plateau of the graph would produce a similar number of clusters and we have chosen the widest value indicated in blue. B) Recovered spacers using a assembly-based method (MinCED) show similar length-distributions as we have reported in the manuscript using the Crass dataset, especially, the absence of 31-nt long spacer sequences. The non-normality of spacer length distributions cannot be explained by a Crass specific length bias.