Cell Host & Microbe, Volume *31*

Supplemental information

Extension of the *Segatella copri* complex to 13 species with distinct large extrachromosomal elements and associations with host conditions

Aitor Blanco-Míguez, Eric J.C. Gálvez, Edoardo Pasolli, Francesca De Filippis, Lena Amend, Kun D. Huang, Paolo Manghi, Till-Robin Lesker, Thomas Riedel, Linda Cova, Michal Punčochář, Andrew Maltez Thomas, Mireia Valles-Colomer, Isabel Schober, Thomas C.A. Hitch, Thomas Clavel, Sarah E. Berry, Richard Davies, Jonathan Wolf, Tim D. Spector, Jörg Overmann, Adrian Tett, Danilo Ercolini, Nicola Segata, and Till Strowig

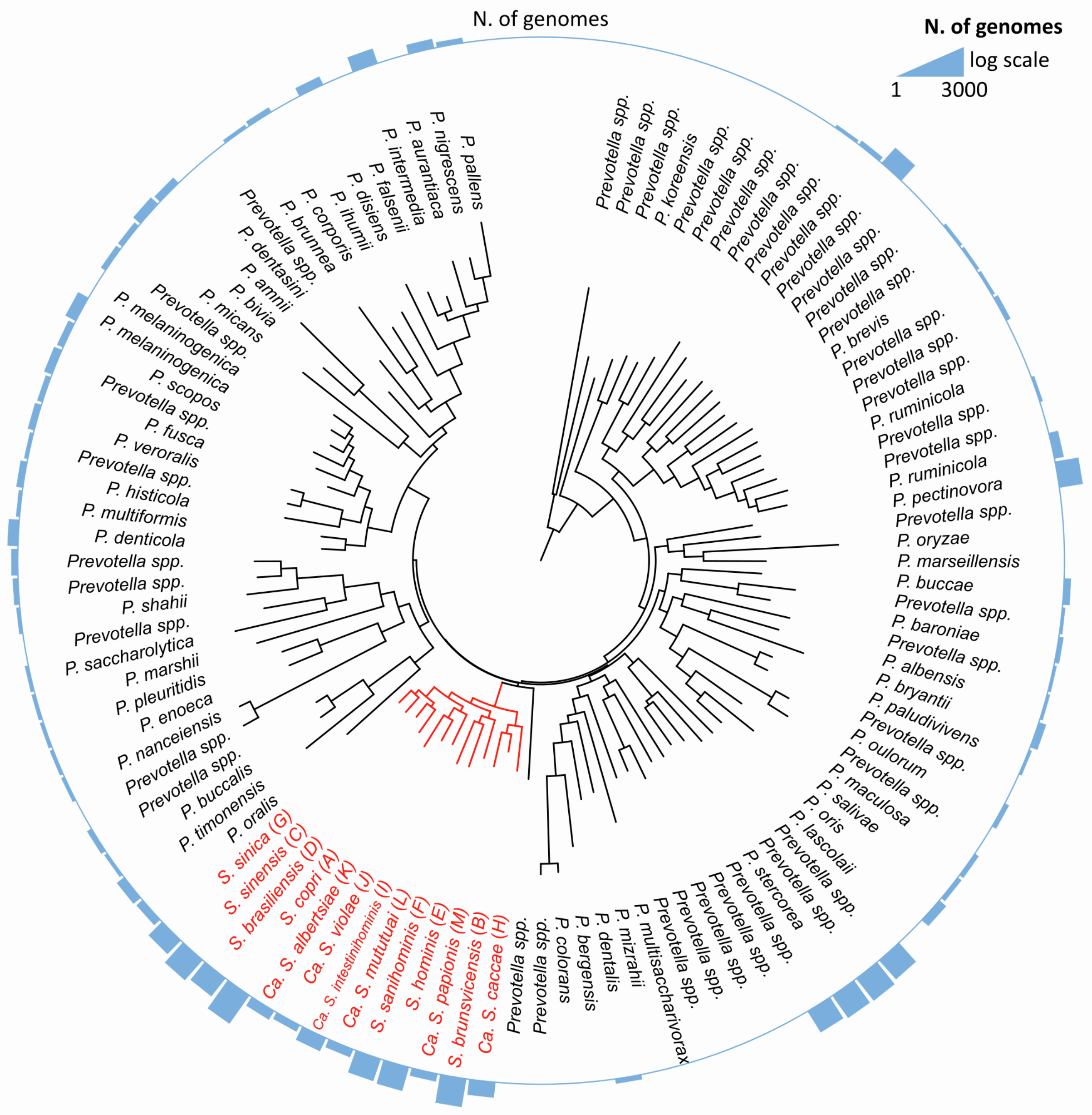

Figure S1: Phylogenetic tree of genomes of strains of *Prevotella* species. The tree was built using a maximum likelihood approach based on marker genes. Only one representative per species is considered. *Prevotella* spp. are reported as assigned in the NCBI RefSeq database as of June 2021 (previous to its reclassification in 7 different genera). ScC species are highlighted in red. Related to **Figure 1**.

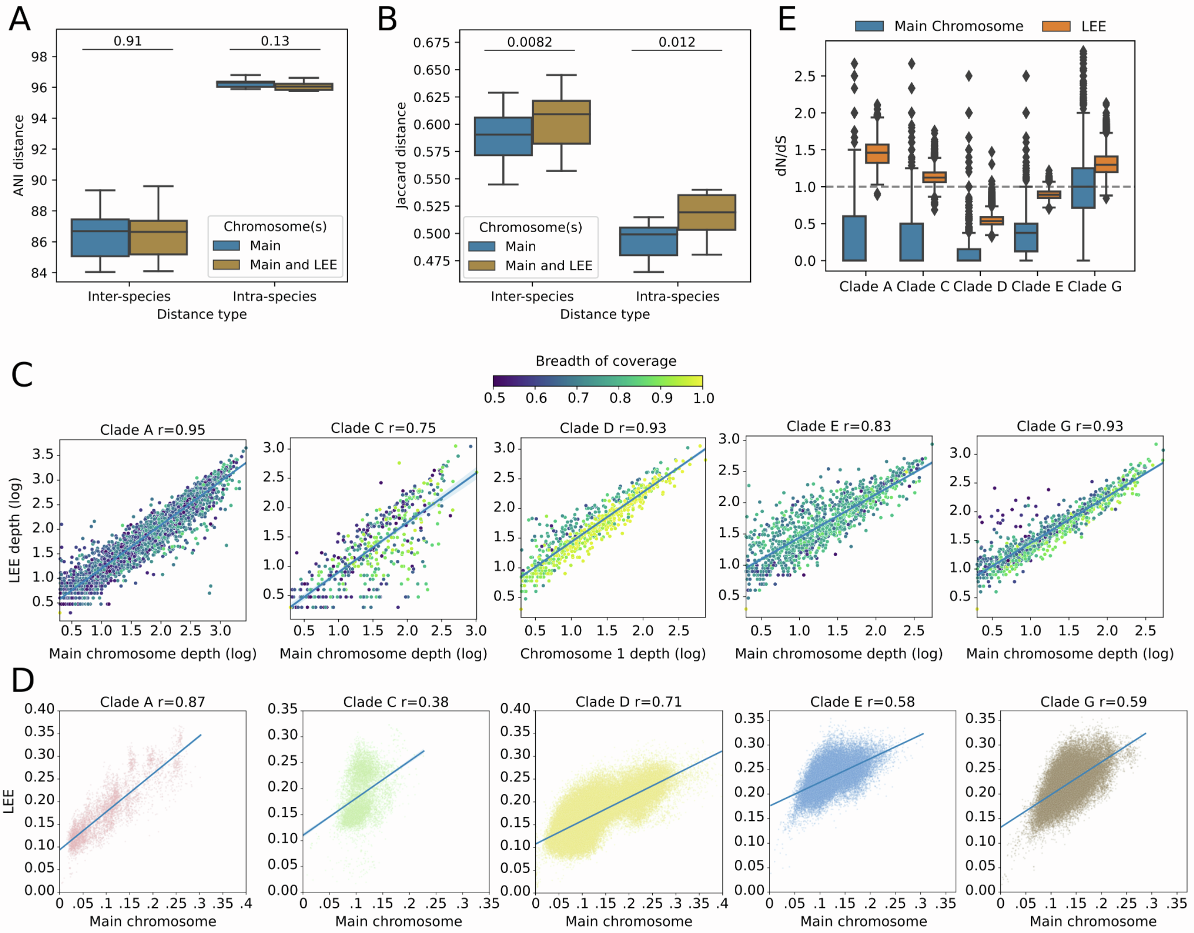

Figure S2: Analysis of ScC large extrachromosomal elements (LEE). (A) Spearman correlation between the depth of coverage of the main chromosome (x-axis) and the LEE (y-axis) when the depth of coverage of the main chromosome is above 2x and the breadth of coverage of the LEE is above 50%. The color gradient represents the breadth of coverage of the LEE. **(B)** Spearman correlation of the single-nucleotide polymorphisms (SNP) rates between the main (x-axis) and the secondary (y-axis) chromosomes when the breadth of coverage of the secondary chromosome is above 80%. SNP rates of the main chromosome were calculated using the multiple-sequence alignment (MSA) of the StrainPhlAn marker genes and those from the secondary chromosome 2 using the MSA of the full chromosome alignment. (**C**) Pairwise dN/dS rates difference between the main chromosome and the LEE (Wilcoxon signed-rank test = 0.0). Main chromosome dN/dS rates were calculated using the StrainPhlan marker genes while LEE dN/dS rates were calculated using the predicted ORFs of the different LEE variants. (**D**) ANI distances between the closed genomes when accounting only the main chromosome or when accounting for the main chromosome and the LEE together. Significance was assessed using Mann-Whitney U tests. (**E**) Differences based on Jaccard distances from presence/absence of gene families (clustered at 50% identity) between the closed genomes when accounting only the main chromosome or when accounting for the main chromosome and the LEE together. Significance was assessed using Mann-Whitney U tests. Box plots in **C**, **D** and **E** show the median (center), 25th/75th percentile (lower/upper hinges), 1.5× interquartile range (whiskers) and outliers (points). Related to **Figure 3**.

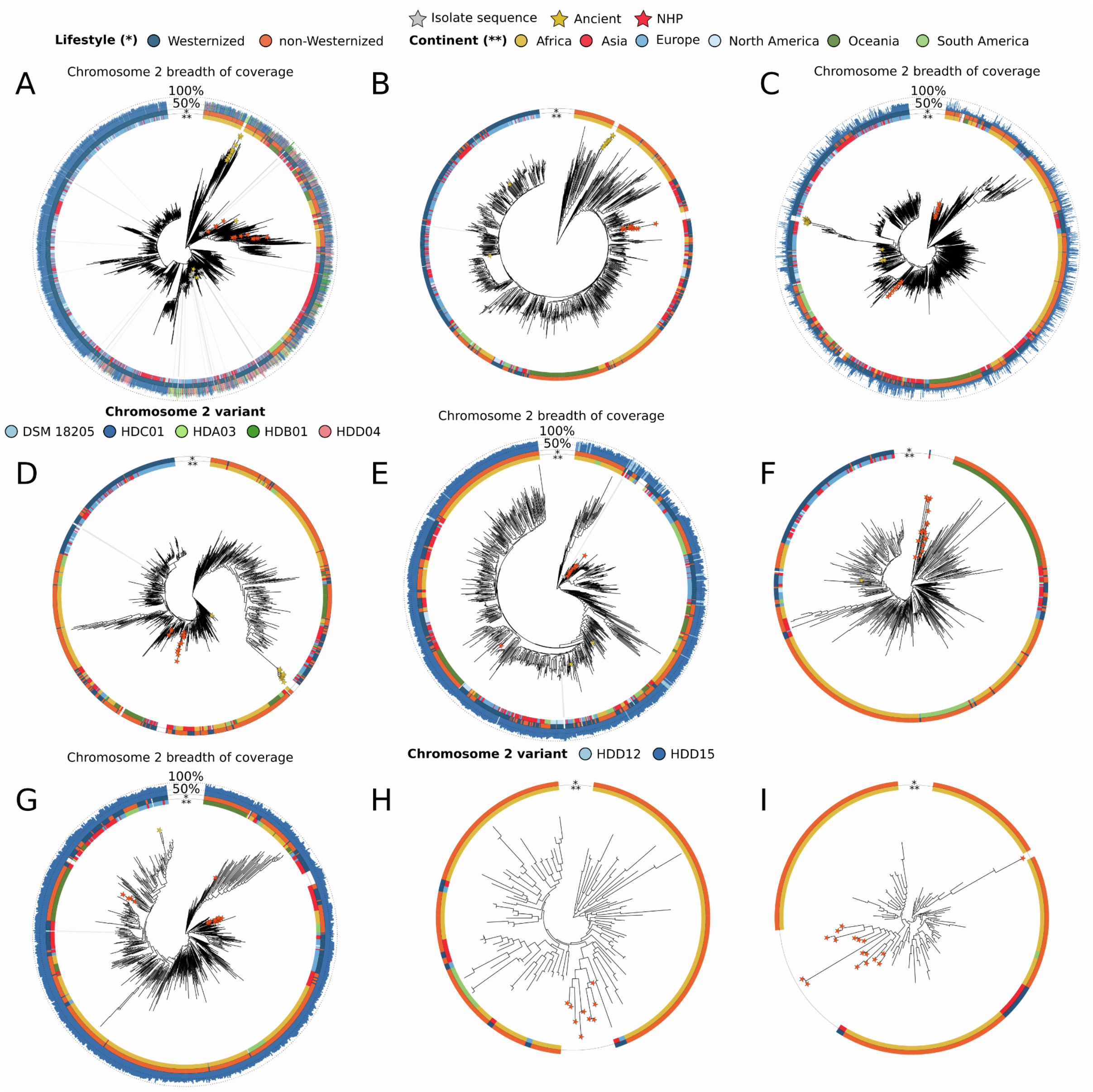

Figure S3: Strain-level phylogenetic analysis of the human ScC species. (A-I) Clades A to I. The inner ring represents the continent of origin, the center ring represents the lifestyle of the host and the outer ring (colored by variant in the cases more than 1 variant is available) represents the breadth of coverage of the second chromosome. Red stars represent captive non-human primates (NHP), yellow stars ancient samples and gray stars isolate sequences. * = Lifestyle, ** = Continent. Related to **Figures 3** and **4**.

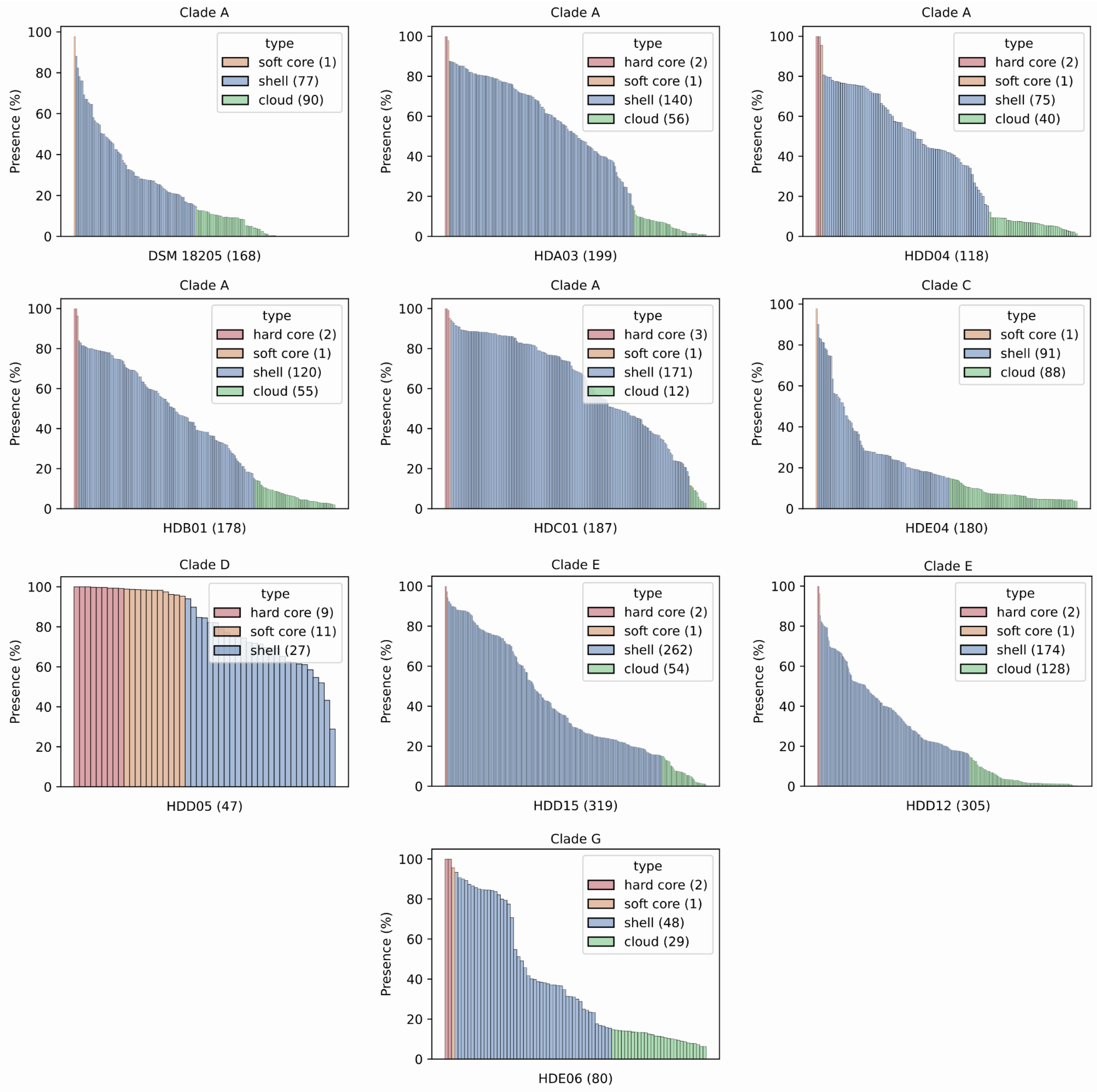

Figure S4: Pangenome analysis of the LEE variants using the metagenomic-reconstructed sequences. Each bar of the histogram represents an individual gene of the pangenome. Only samples with a depth of coverage > 10x of the corresponding variant were used. A gene was defined present when the breadth of coverage was above 50%. Numbers between parentheses represent the number of genes of each category. Related to **Figure 3**.

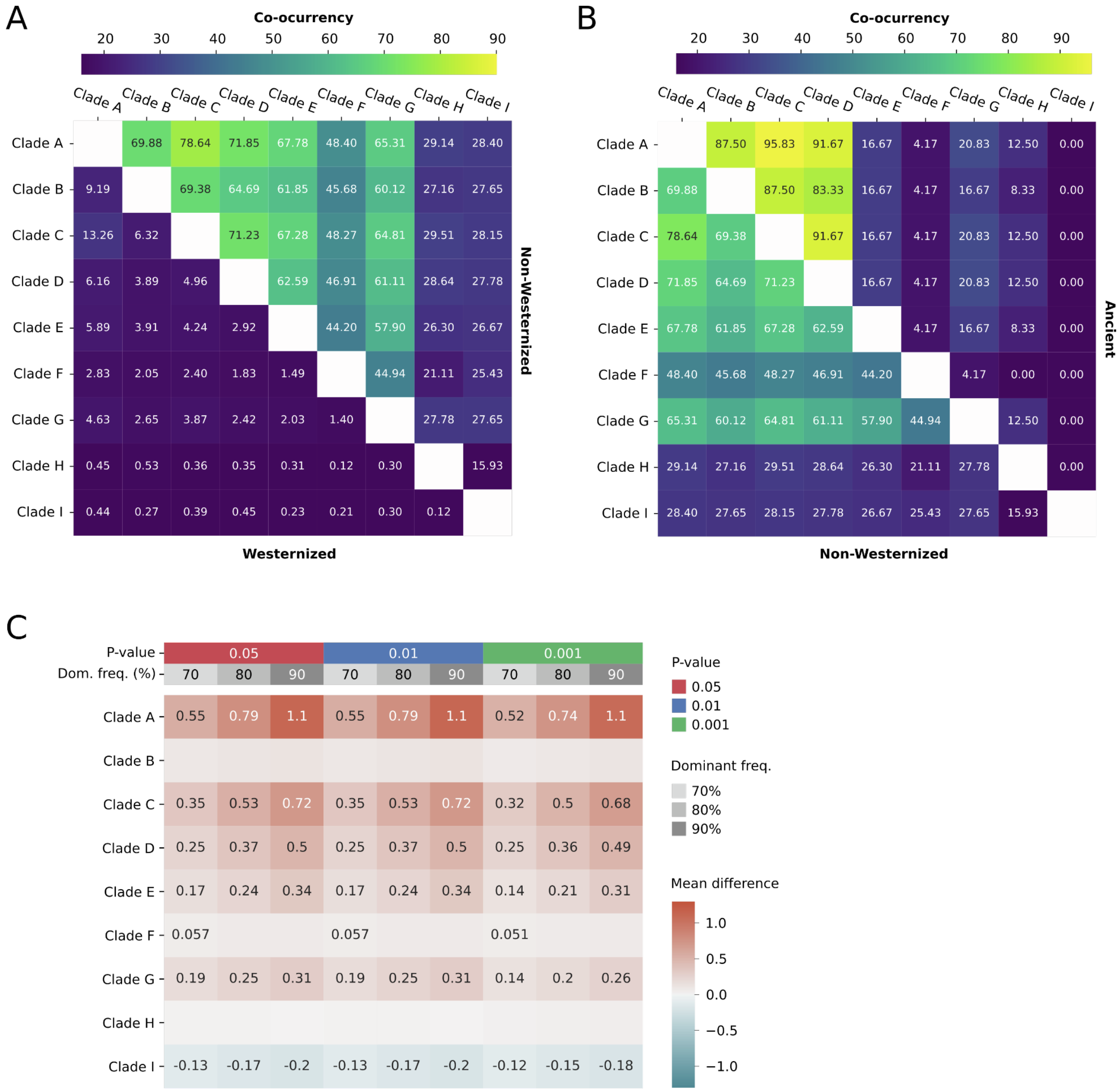

Figure S5: Characterization of ScC co-occurrence. (**A**) Co-occurrence of the human ScC species in Westernized (bottom-left triangle) vs non-Westernized (top-right triangle) individuals. (**B**) Co-occurrence of the human ScC species in non-Westernized individuals (bottom-left triangle) vs ancient (top-right triangle) samples. (**C**) Average differences between the intra-individual polymorphic rates of Westernized and non-Westernized individuals across different p-values and dominant allele frequencies thresholds. Average differences were calculated using a randomly selected subset of 100 samples per each lifestyle. Numbers in the heatmap represent the parameters showing statistically significant differences (Mann-Whitney U test < 0.05). Related to **Figure 3**.

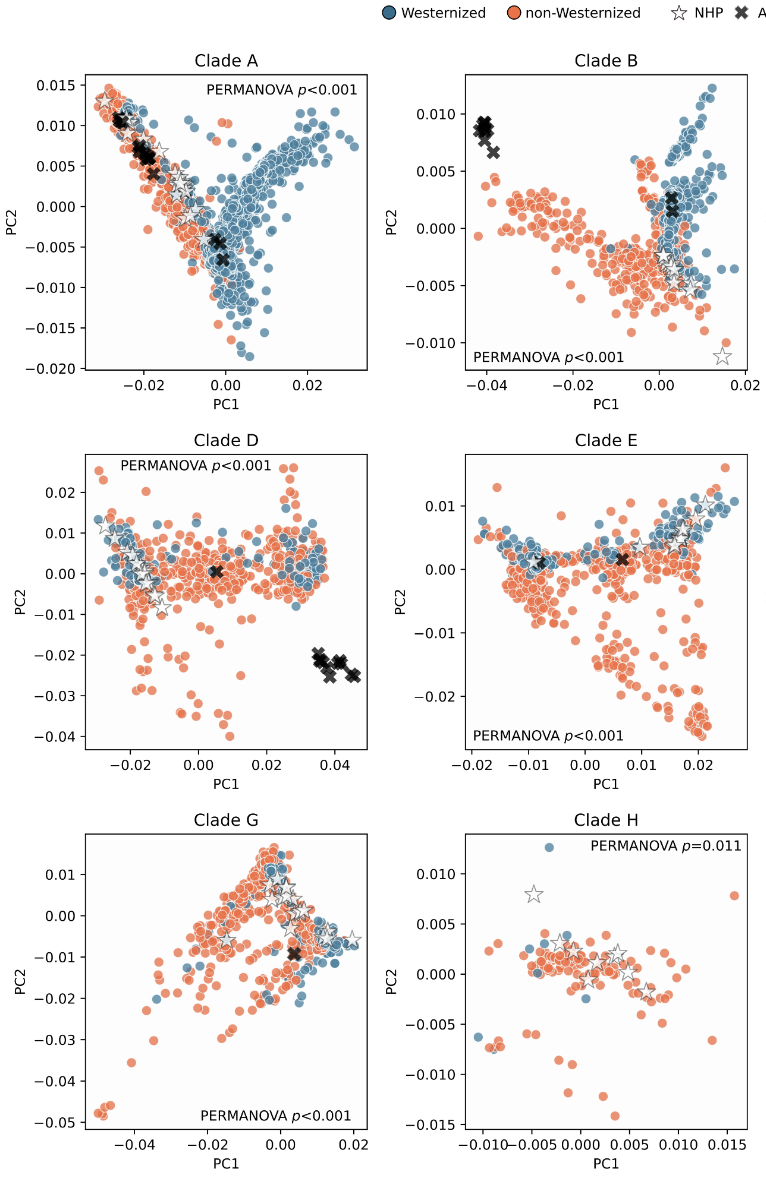

Figure S6: Analysis of genetic diversification in the ScC. Multidimensional scaling (MDS) based on the pairwise SNP rates on the StrainPhlAn SGB-specific marker genes annotated by lifestyle for each species of the ScC. NHP = Non-human primates. Related to **Figure 4**.

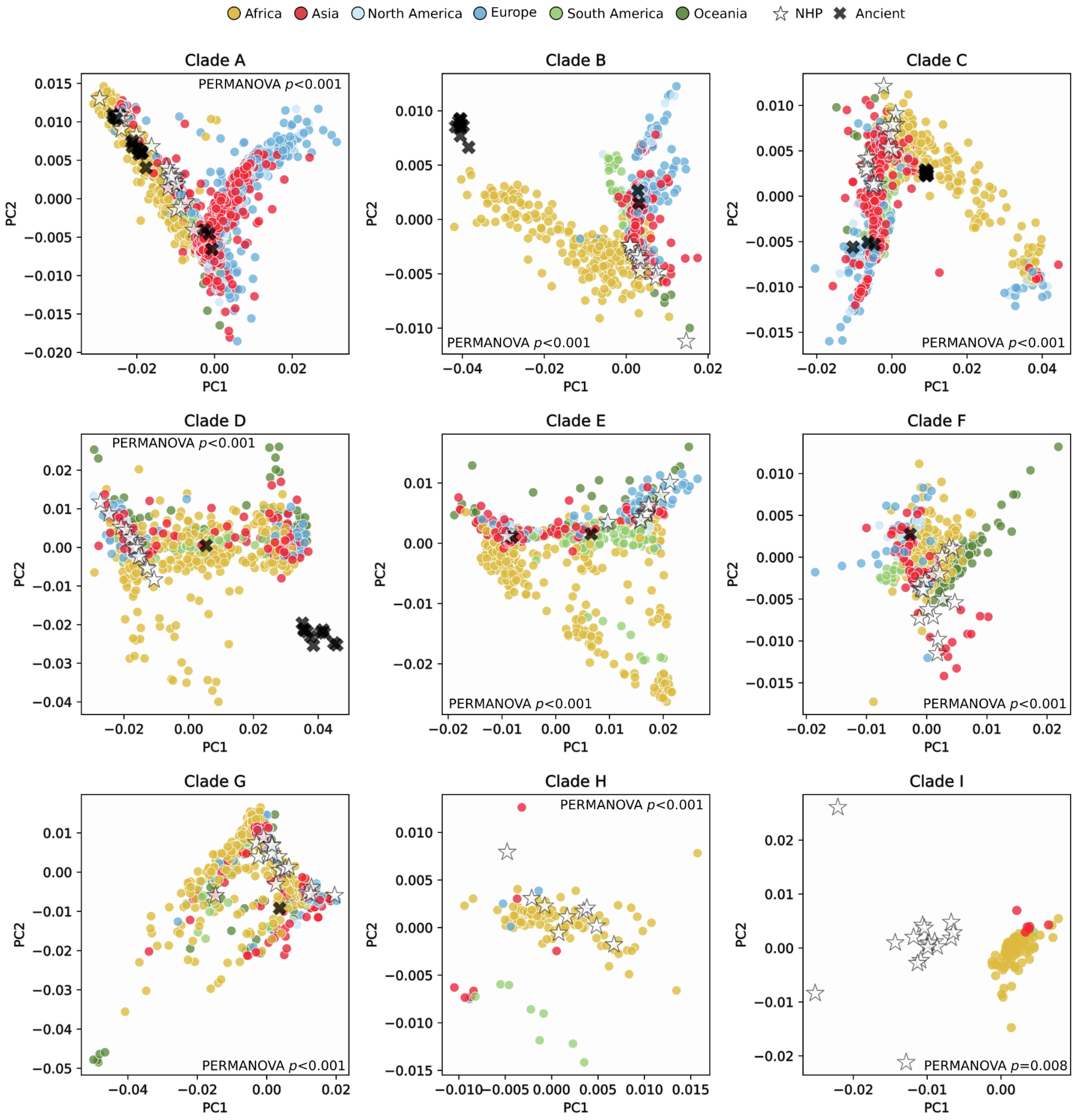

Figure S7: Analysis of genetic diversification in the ScC depending on geographical origin. Multidimensional scaling (MDS) based on the pairwise SNP rates on the StrainPhlAn SGB-specific marker genes annotated by continent for each species of the ScC. NHP = Non-human primates. Related to **Figure 4**.

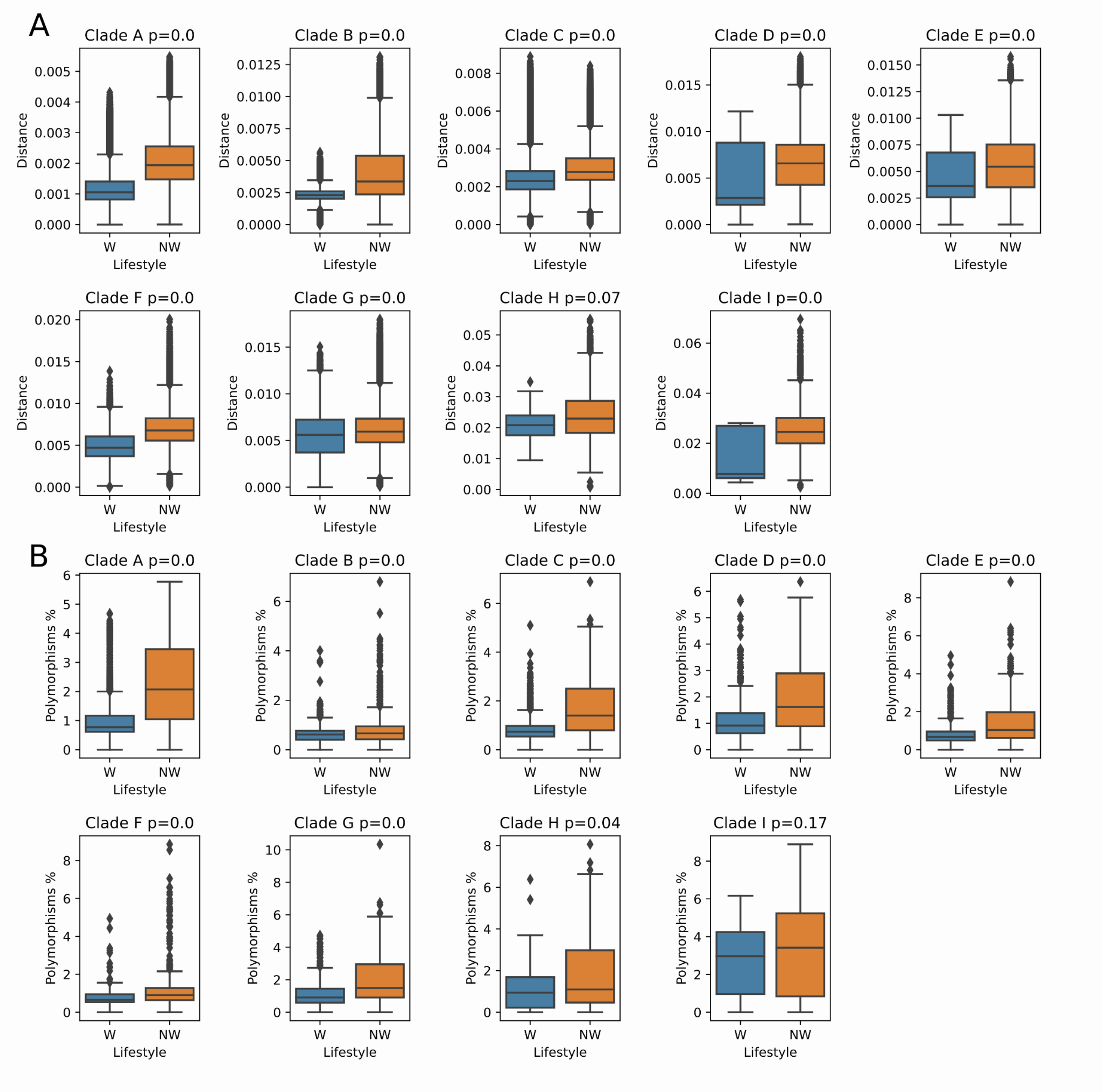

Figure S8: Phylogenetic analysis of the ScC. Differences in the (**A**) intra-lifestyle phylogenetic distances comparison and in the (**B**) polymorphisms found between Westernized and non-Westernized samples for the ScC species (Mann-Whitney U test). Pairwise phylogenetic distances were calculated using the StrainPhlAn tree branch lengths normalized by the total branch length. Polymorphisms were calculated using the StrainPhlAn consensus marker genes and were defined as positions in the reconstructed markers with a dominant allele frequency below 80%. Significance of the comparisons was assessed using linear mixed effects models accounting for sequence depth as confounding variable and dataset as fixed effect. Box plots show the median (center), 25th/75th percentile (lower/upper hinges), 1.5× interquartile range (whiskers) and outliers (points). Related to **Figure 4**.

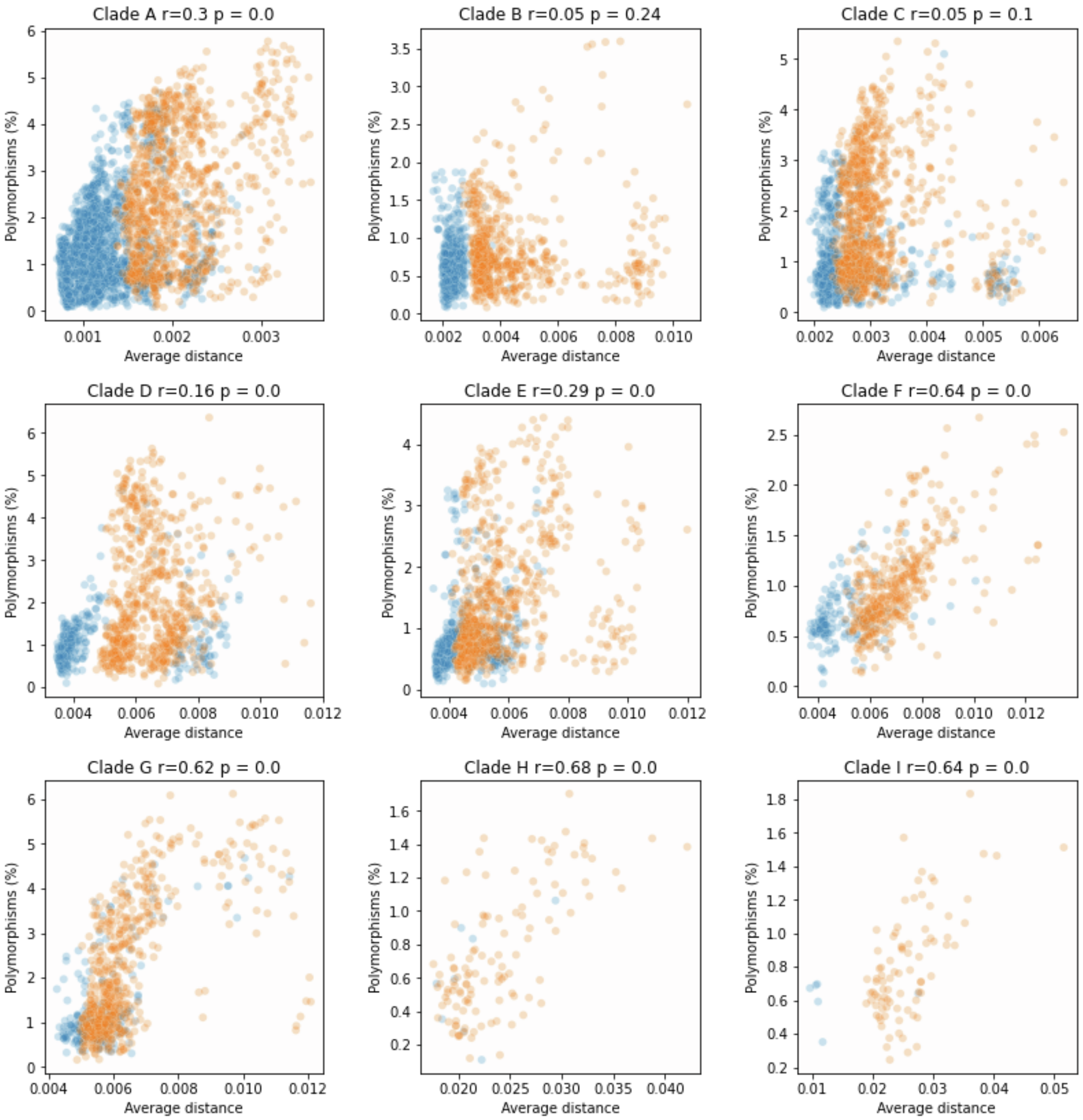

Figure S9: Pearson's correlation between average phylogenetic distances and polymorphic rates between strains from populations following different lifestyles. Blue dots represent Westernized strains while orange dots are non-Westernized ones. Pairwise phylogenetic distances were calculated using the StrainPhlAn tree branch lengths normalized by the total branch length. Polymorphisms were calculated using the StrainPhlAn consensus marker genes and were defined as positions in the reconstructed markers with a dominant allele frequency below 80%. Related to **Figure 4**.

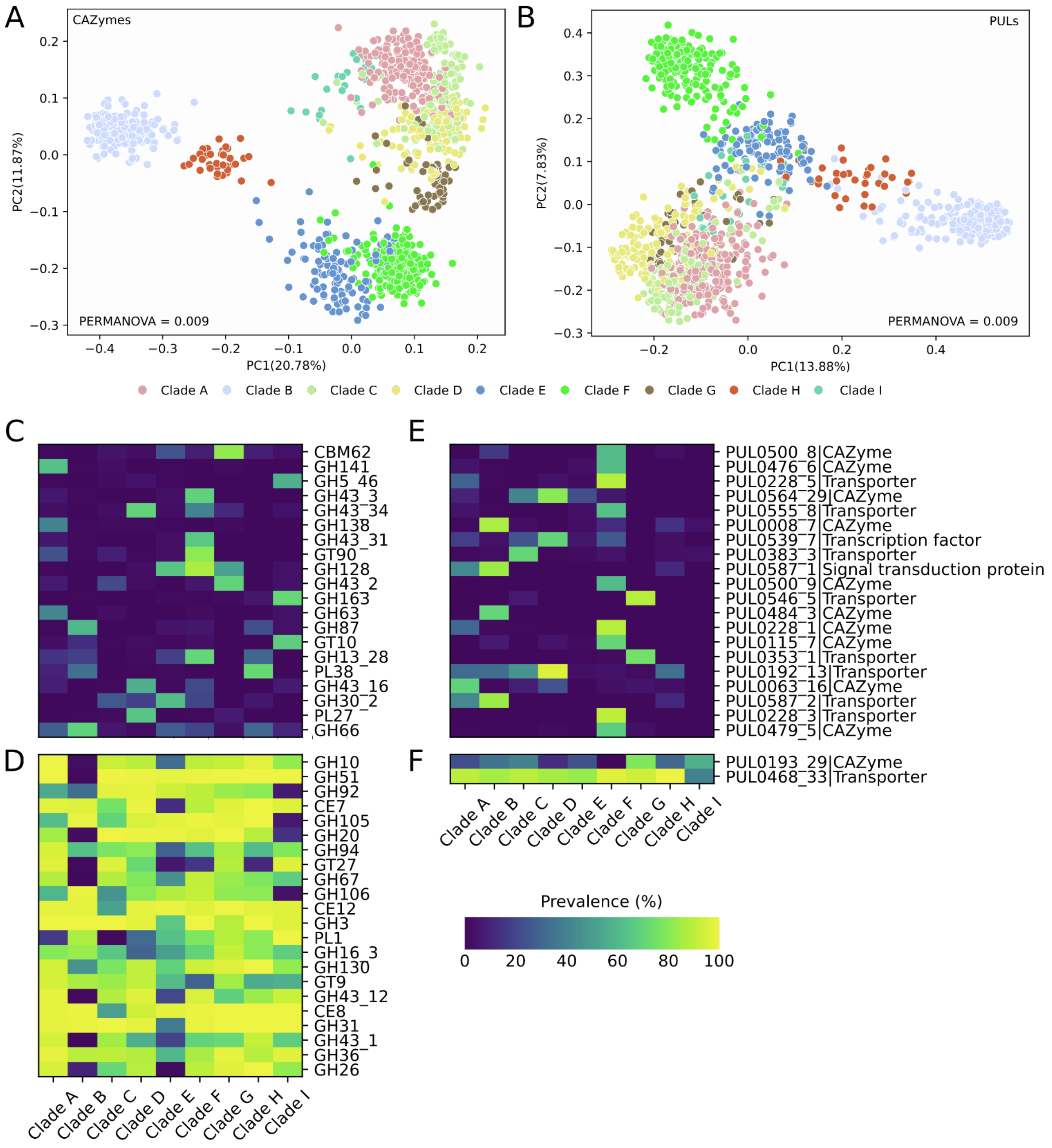

Figure S10: Carbohydrate utilization potential of the ScC. (A) PCoA based on the Jaccard distances of the CAZymes (PERMANOVA p-value = 0.009). **(B)** PCoA based on the Jaccard distances of the PULs (PERMANOVA p-value = 0.009). **(C-D)** Prevalence heatmap for CAZymes significantly enriched **(C)** and depleted **(D)** in at least one species in comparison to each of the other eight (Fisher's exact test FDR < 0.05). Prevalence denotes the percentage of genomes in that species for which they possess at least one gene from the given PUL. (**E-F**) Prevalence heatmap for PULs significantly enriched **(E)** and depleted **(F)** in at least one species in comparison to each of the other eight (Fisher's exact test FDR < 0.05). Prevalence denotes the percentage of genomes in that species for which they possess at least one gene from the given PUL. Related to **Figure 5**.

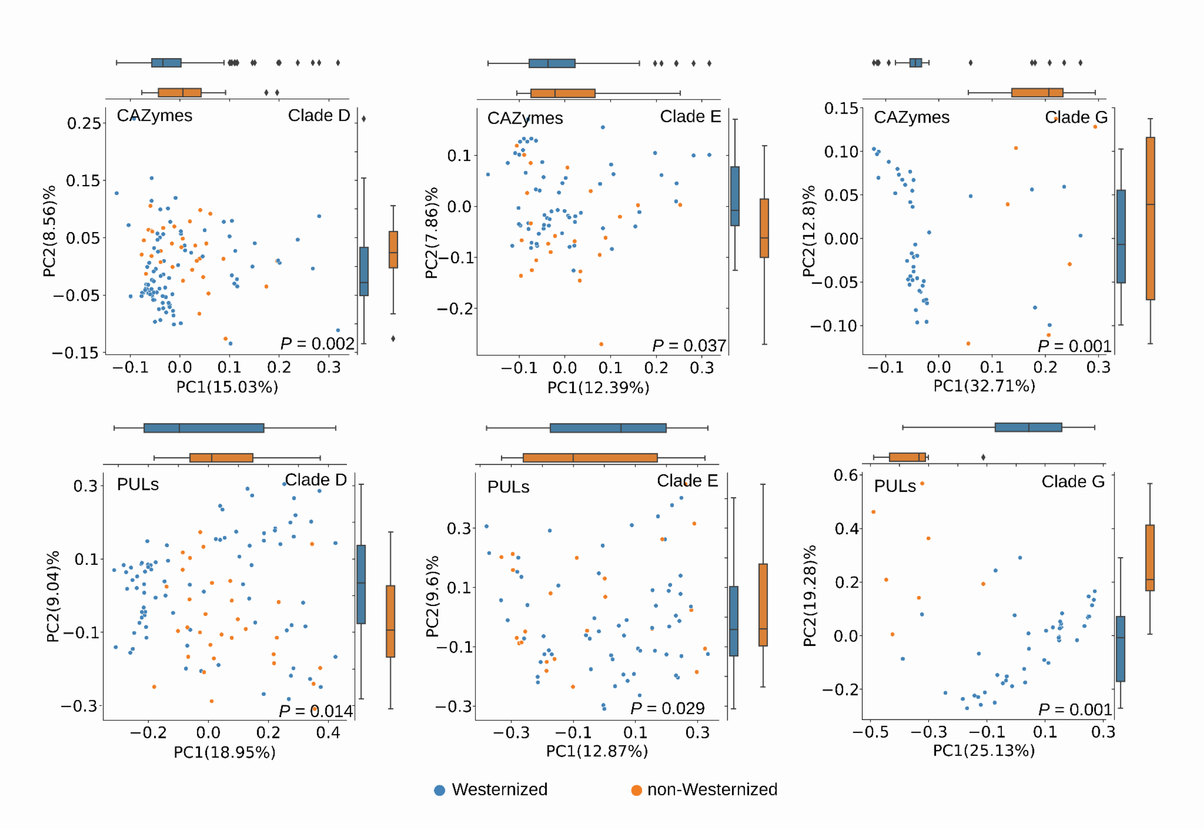

Figure S11: Carbohydrate utilization potential of ScC members from Westernized and non-Westernized countries. PCoA based on the Jaccard distances of the predicted CAZymes (top) and PULs (bottom) between *S. brasiliensis* (clade D), *S. hominis* (clade E) and *S. sinica* (clade G) using MAGs reconstructed from Westernized or non-Westernized individuals (PERMANOVA p-values are shown for each PCoA plot). Related to **Figure 5**.