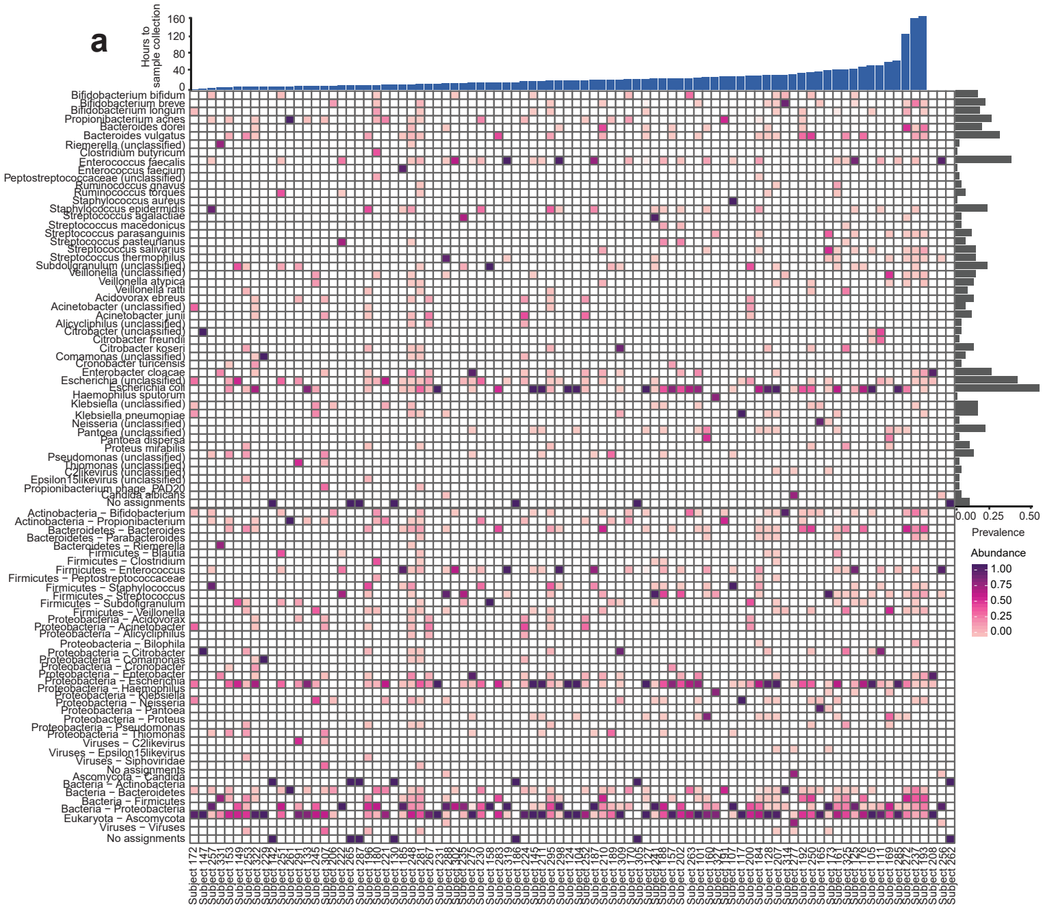
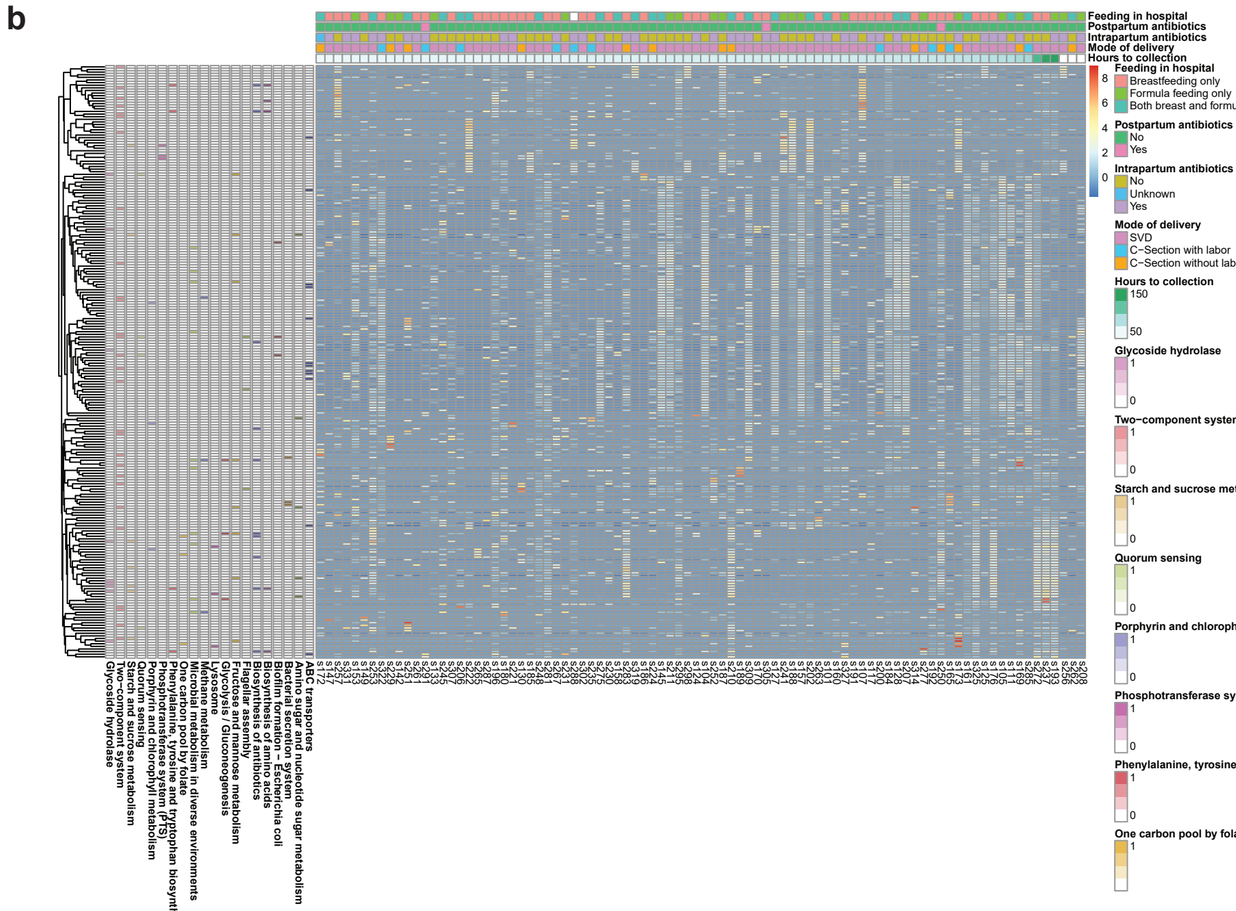
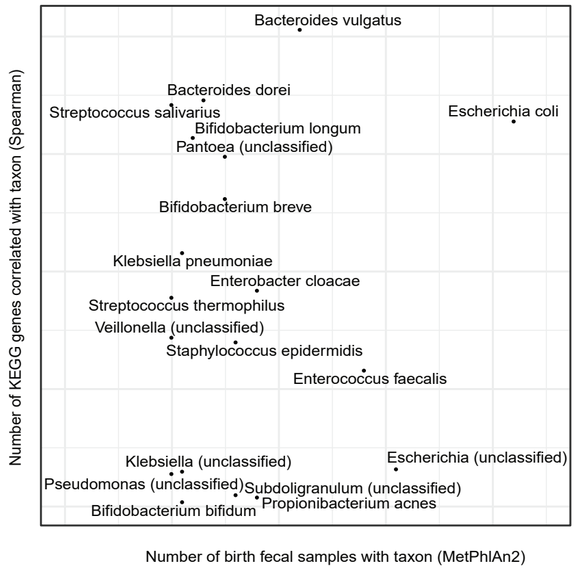
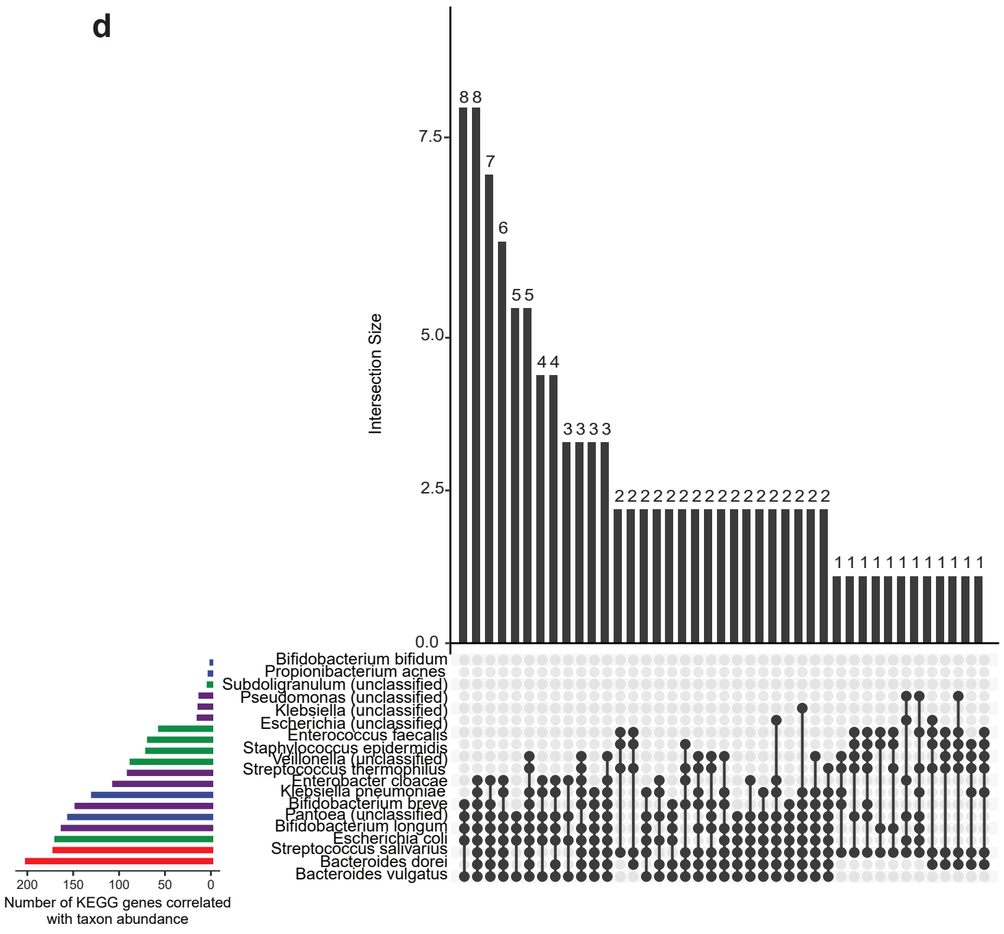
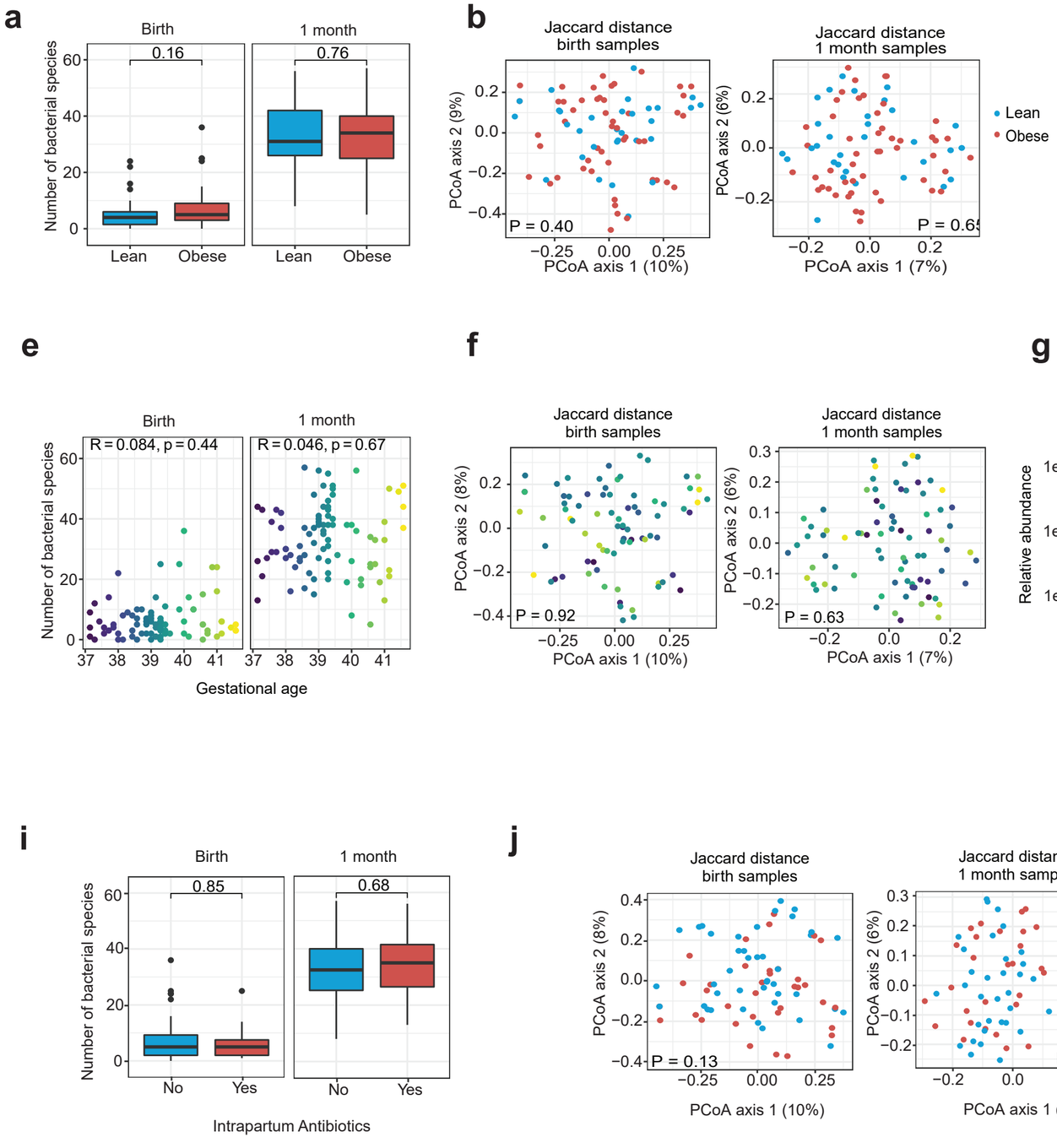




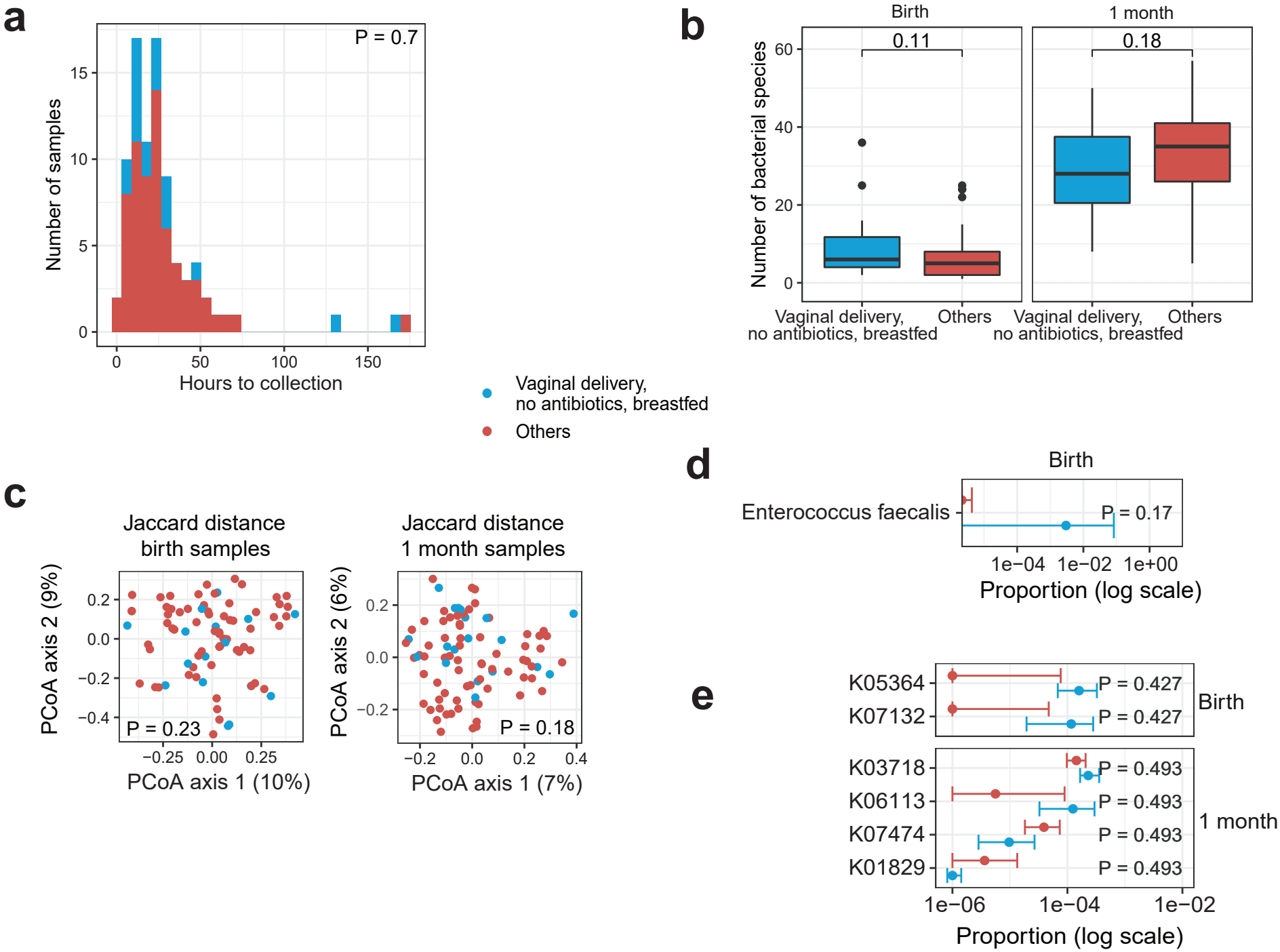




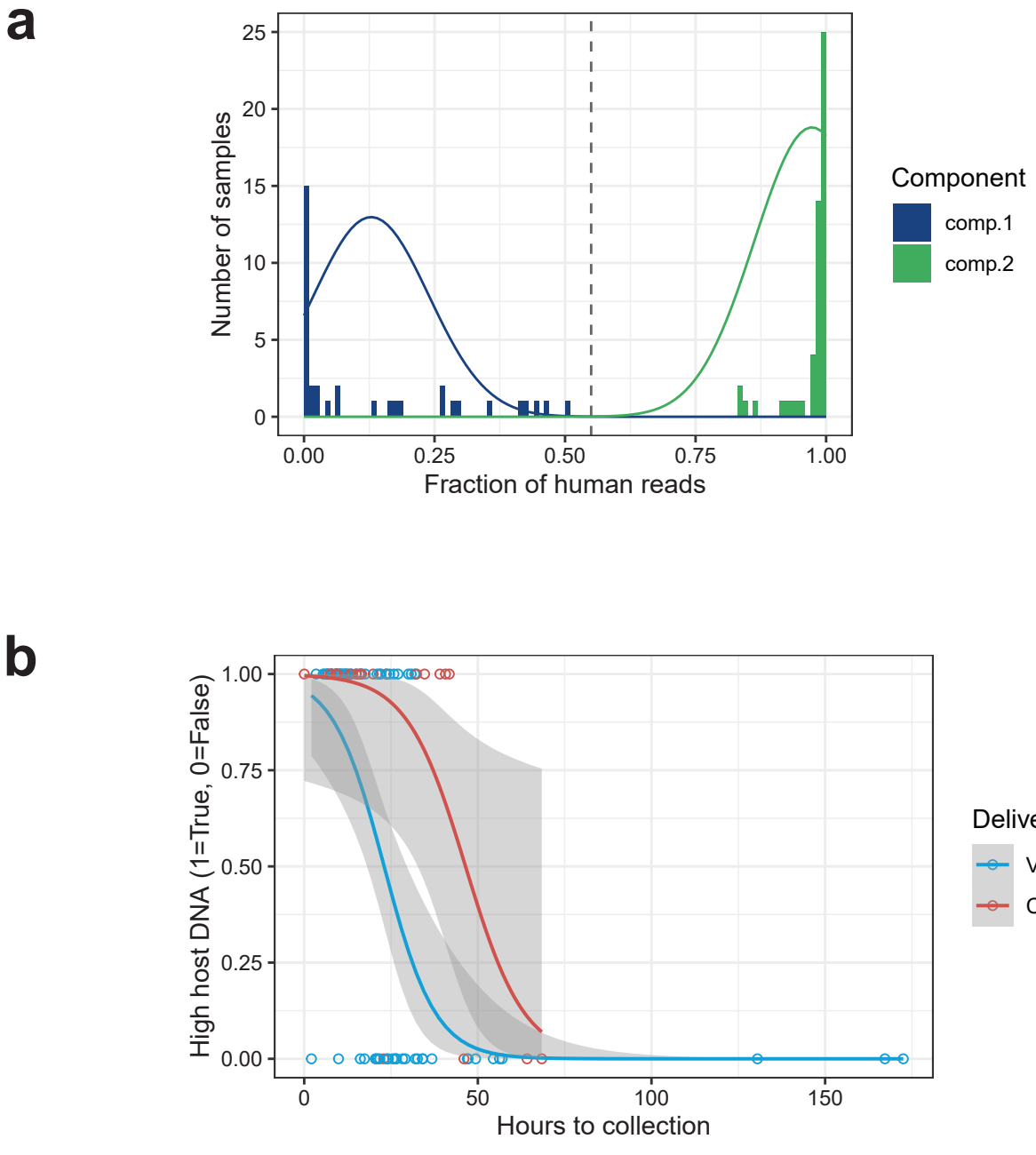
Supplementary Figure 1: **Metagenomic sequencing of meconium samples.** (a) Bacteria, viruses, and fungi detected in taxonomic analysis of meconium samples. Taxa were included if the relative abundance was greater than 10% in any sample. The bar chart above the heatmap shows the number of hours to sample collection. The bar chart to the right of the heatmap shows prevalence of each taxon. N.d. = no data. (b) KEGG gene orthologs positively correlated with taxon richness (one-sided test of Spearman correlation, n=88 samples, 4,725 orthologs tested). Genes are shown if the p-value was less than 0.001 after correction for false discovery rate (2,642 orthologs were significant at the P<0.05 level after correction). Broad functional categories are marked by colored squares to the left of the heatmap. The dendrogram on the far left indicates clustering of genes according to abundance. (c) Number of KEGG gene orthologs correlated with taxa identified. Fewer than 20 genes were associated with detection of Propionibacterium acnes, unclassified Klebsiella, unclassified Pseudomonas, unclassified Subdoligranulum, and unclassified Escherichia coli. (d) Upset plot showing overlap in gene orthologs associated with taxa identified.



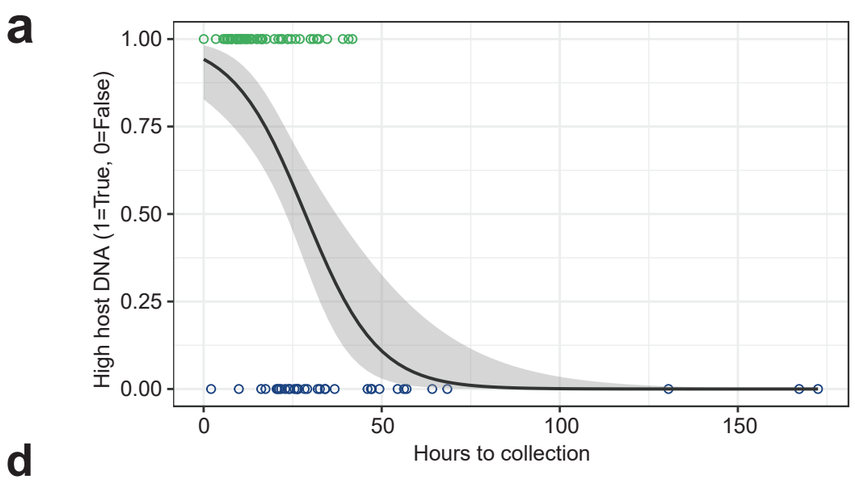
Supplementary Figure 2: **Effect of maternal obesity, gestational age, and intrapartum antibiotics on microbiota.** (a-d) Association of maternal obesity with (a) number of bacterial species (two-sided Mann-Whitney test), (b) Jaccard distance (PERMANOVA test), (c) taxon relative abundance (two-sided Mann-Whitney test), and (d) KEGG ortholog relative abundance (two-sided Mann-Whitney test). The sample size for maternal obesity was $n_1 = 35$ lean, $n_2 = 53$ obese. (e-h) Association of gestational age with (e) number of bacterial species (two-sided test of Spearman correlation), (f) Jaccard distance (PERMANOVA test), (g) taxon relative abundance (two-sided test of Spearman correlation), and (h) KEGG ortholog relative abundance (two-sided test of Spearman correlation). The sample size for gestational age was n = 88. (i-l) Association of intrapartum antibiotics with (i) number of bacterial species (two-sided Mann-Whitney test), (j) Jaccard distance (PERMANOVA test), (k) taxon relative abundance (two-sided Mann-Whitney test), and (l) KEGG ortholog relative abundance (two-sided Mann-Whitney test). The sample size for intrapartum antibiotics was $n_1 = 35$ with exposure and $n_2 = 48$ without. Taxa and genes with largest effect size are shown, though we observed no statistically significant effects. Color coding for maternal obesity, gestational age, and intrapartum antibiotics is indicated on the right. Boxes in (a) and (i) indicate the median and interquartile distance, whiskers indicate maximum and minimum data points within 1.5 times the interquartile range, points represent values outside this range. Points in (c), (d), (k), and (l) indicate the median; horizontal error bars indicate the first and third quantile of data values. Linear regression estimates in (g) and (h) are indicated with blue lines; 95% confidence intervals are indicated by grey areas.

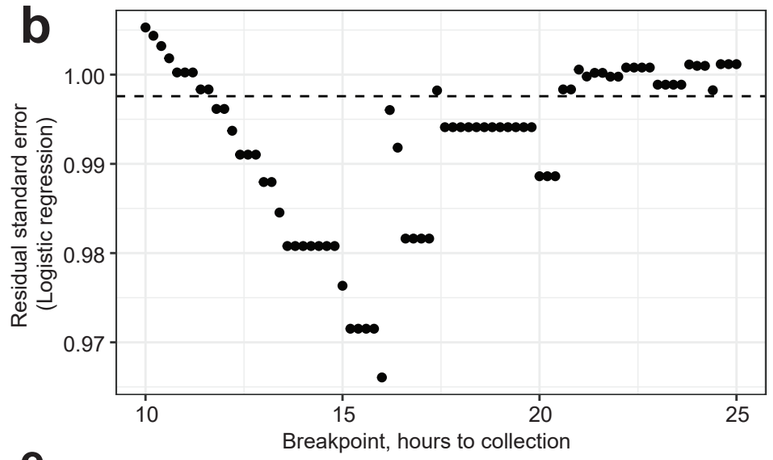


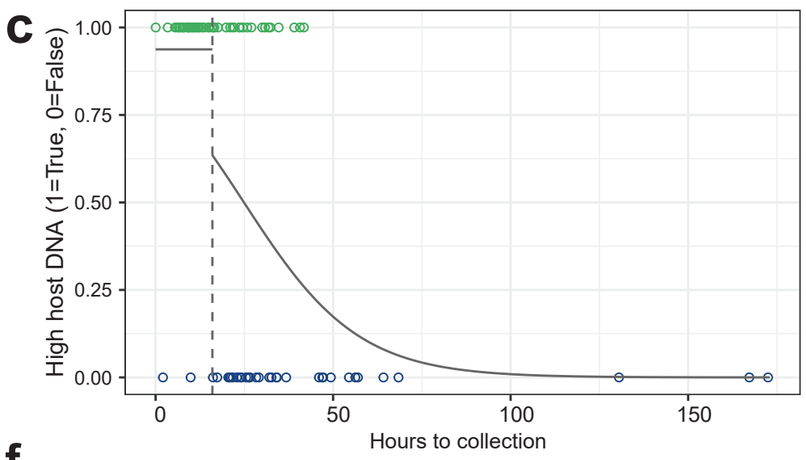
Supplementary Figure 3: **Subgroup of vaginally born infants not exposed to intra- or postpartum antibiotics and exclusively breastfed.** (a) Time from birth in subgroup was not different from other samples in study (two-sided Mann-Whitney test). (b) Alpha diversity, (c) beta diversity, (d) taxon abundance, and (e) KEGG gene ortholog abundance did not differ between the subgroup and other samples. Taxa and genes with largest effect size are shown, though we observed no statistically significant effects (PERMANOVA test for beta diversity, two-sided Mann-Whitney test otherwise). Boxes in (b) indicate the median and interquartile distance, whiskers indicate maximum and minimum data points within 1.5 times the interquartile range, points represent values outside this range. Points in (d) and (e) indicate the median; horizontal error bars indicate the first and third quantile of data values. The sample size for all comparisons was $n_1 = 19$ in the subgroup and $n_2 = 69$ others.

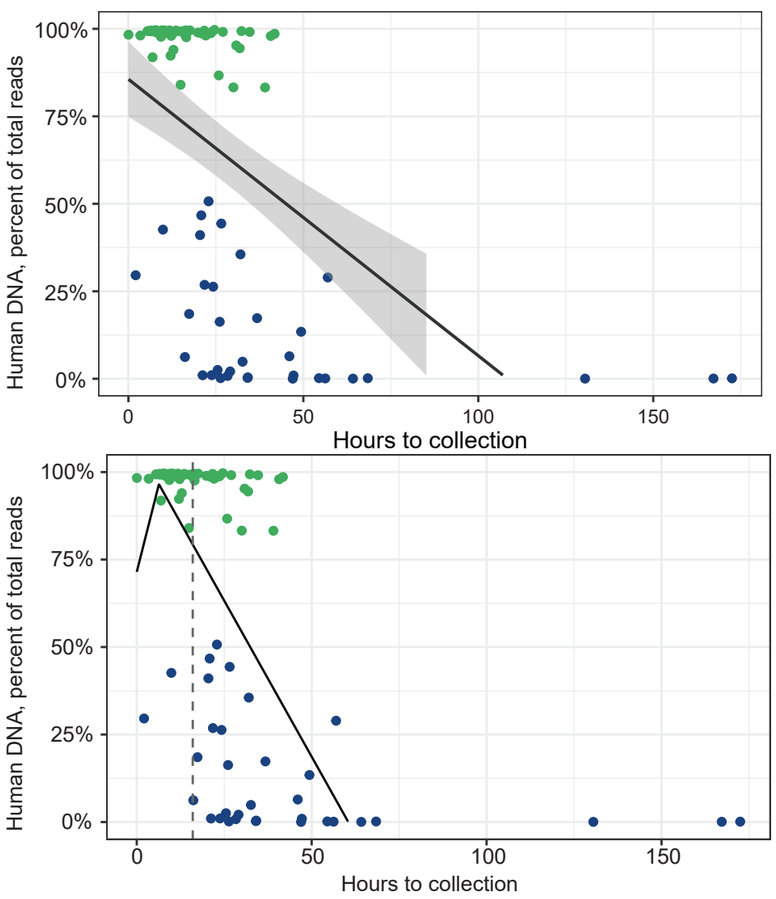


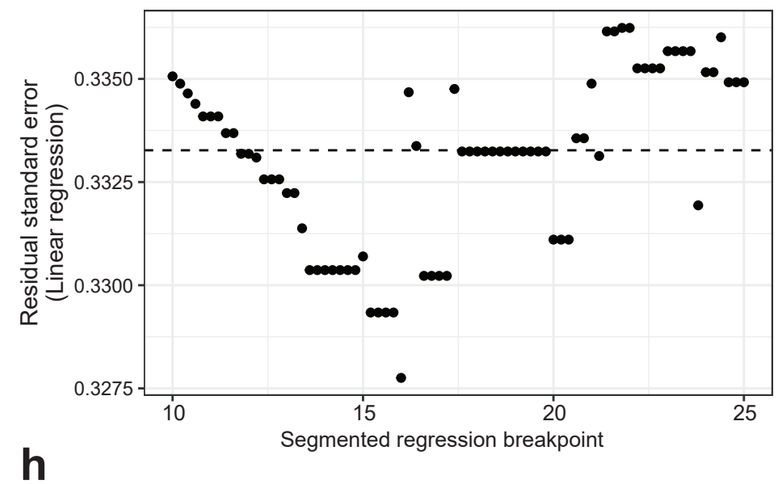
Supplementary Figure 4: **Association between high human DNA and delivery mode.** (a) A Gaussian mixture model with equal variances divides meconium samples into two components of high- and low-human DNA (P = 0.001, 999 bootstrapping simulations, n = 88). The dashed vertical line represents the crossover point between the two components, at a host proportion of 0.55. (b) Vaginal delivery was associated with a decreased odds ratio for high vs. low human DNA over time (P = 0.002, logistic regression, n_1 = 64 vaginal delivery, n_2 = 24 c-section). Solid lines represent the regression estimates, grey areas indicate the 95% confidence interval.

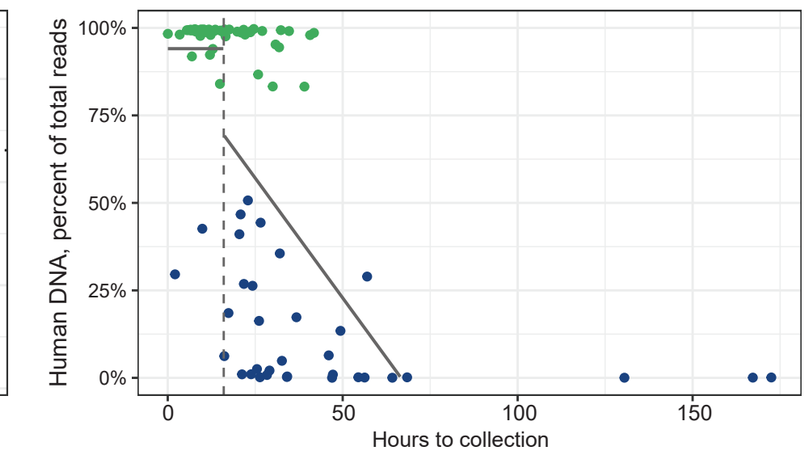


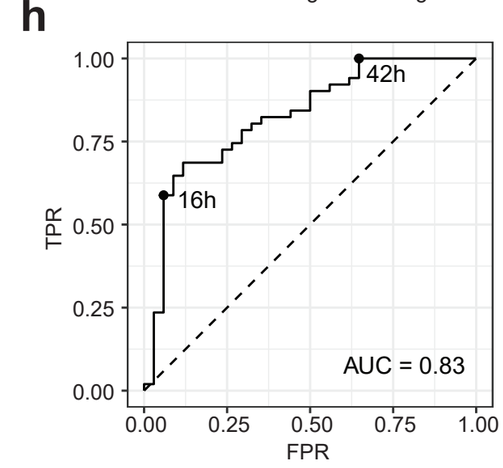




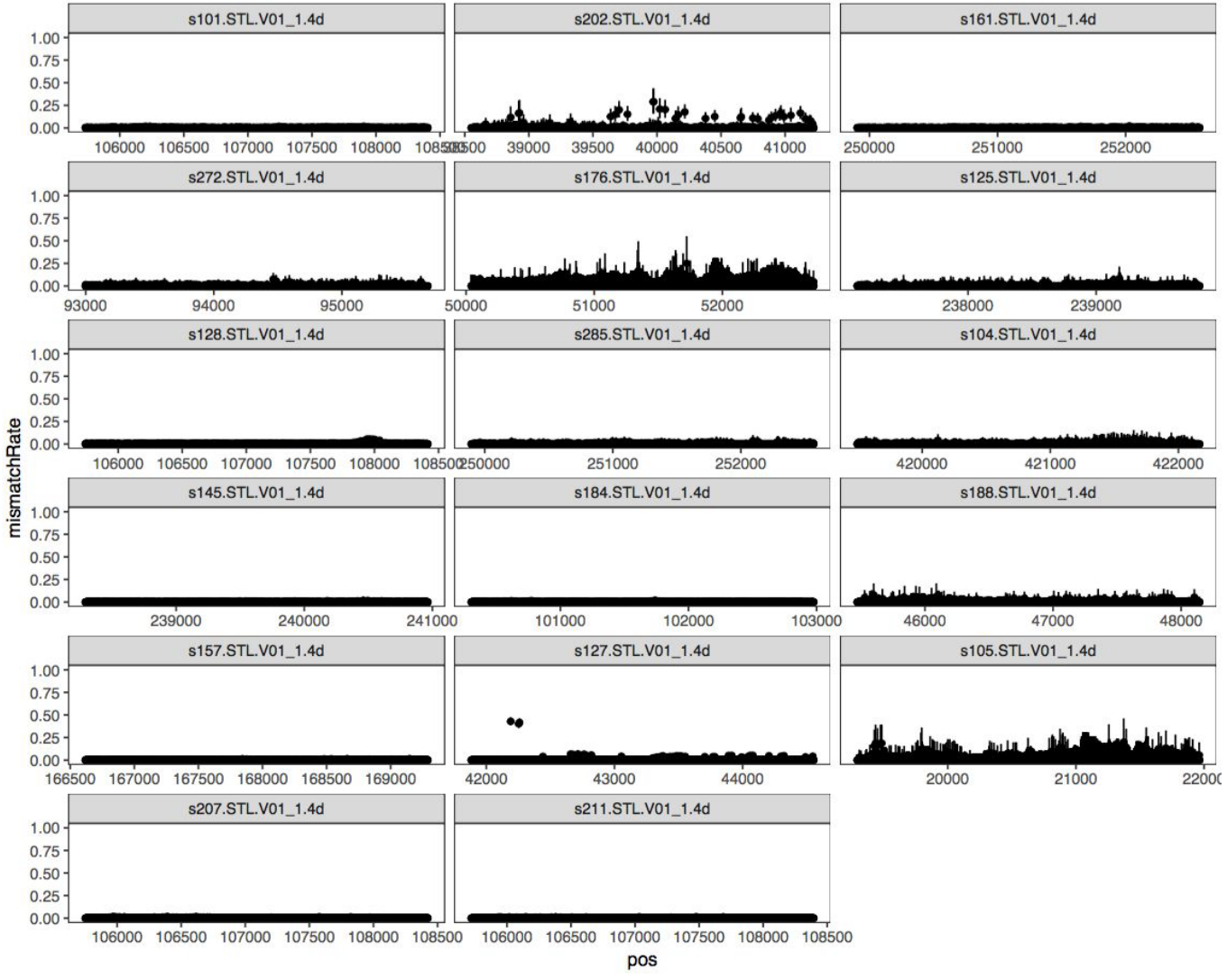




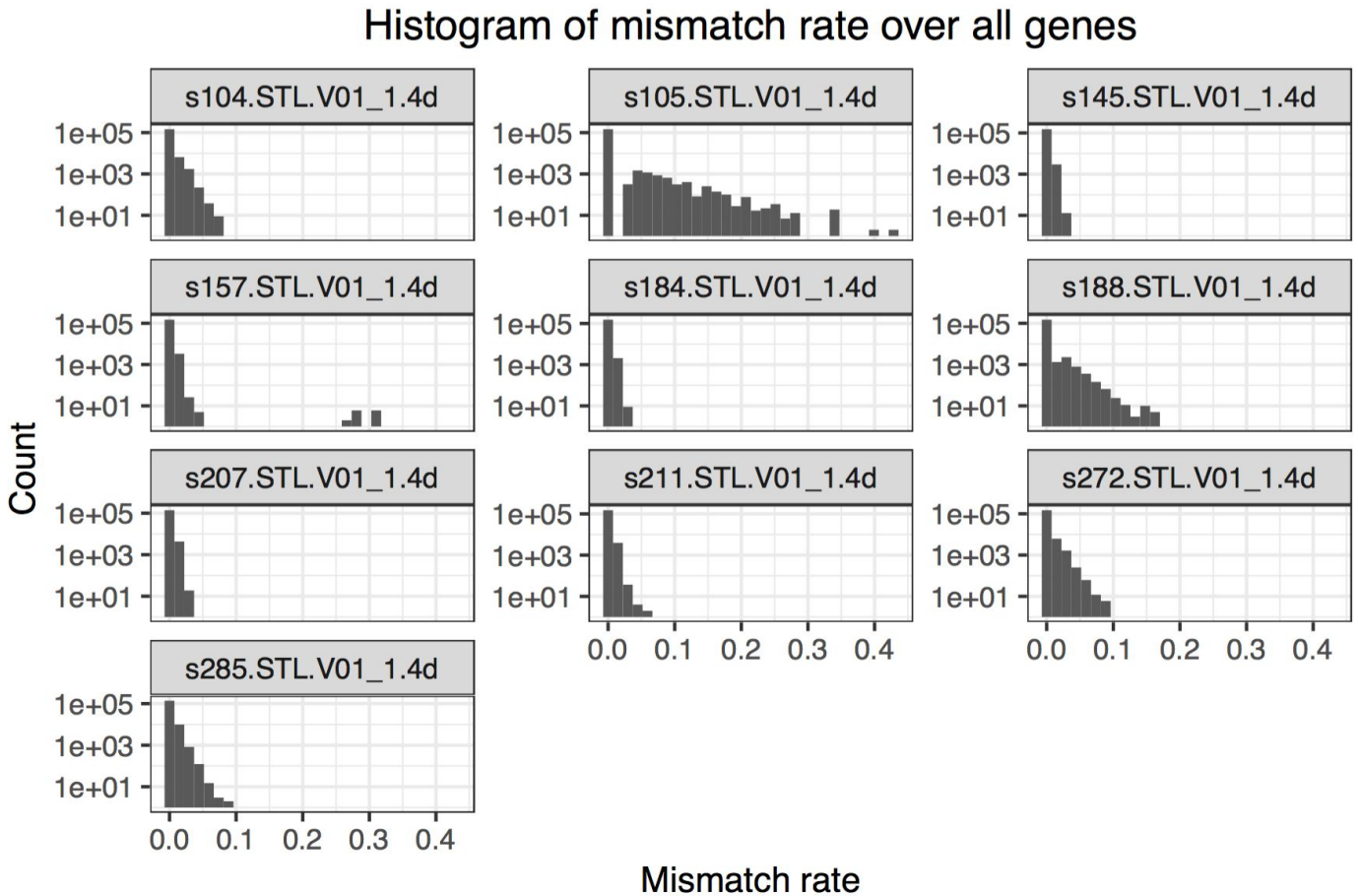




Supplementary Figure 5: **Break point in level of human DNA over time.** (a-c) A break point of 16.0 hours minimizes the sum of squares in logistic regression of high vs. low human DNA. (a) Best fit with no break point, (b) sum of squared residuals for various break point values, (c) best fit with selected break point. (d-f) A break point of 16.0 hours also minimizes the sum of squares in linear regression of human DNA fraction. (d) Best fit with no break point, (e) sum of squared residuals for various break point values, (f) best fit with selected break point. (g) A segmented regression, where the prediction must be continuous, yields a break point of 6 hours, earlier than models with a discontinuous break point. (h) Receiver operating characteristic (ROC) curve for a simple model of high human DNA before the breakpoint, and low human DNA after the breakpoint. The curve illustrates that 16 hours represents a point of inflection with high predictive value for high human DNA. Solid lines represent logistic or linear estimates; 95% confidence intervals are indicated by grey areas. Dashed horizontal lines in (b) and (e) indicate the sum of squared residuals with no break point. The sample size was n = 85 for all regressions (3 samples excluded due to no time of collection data).

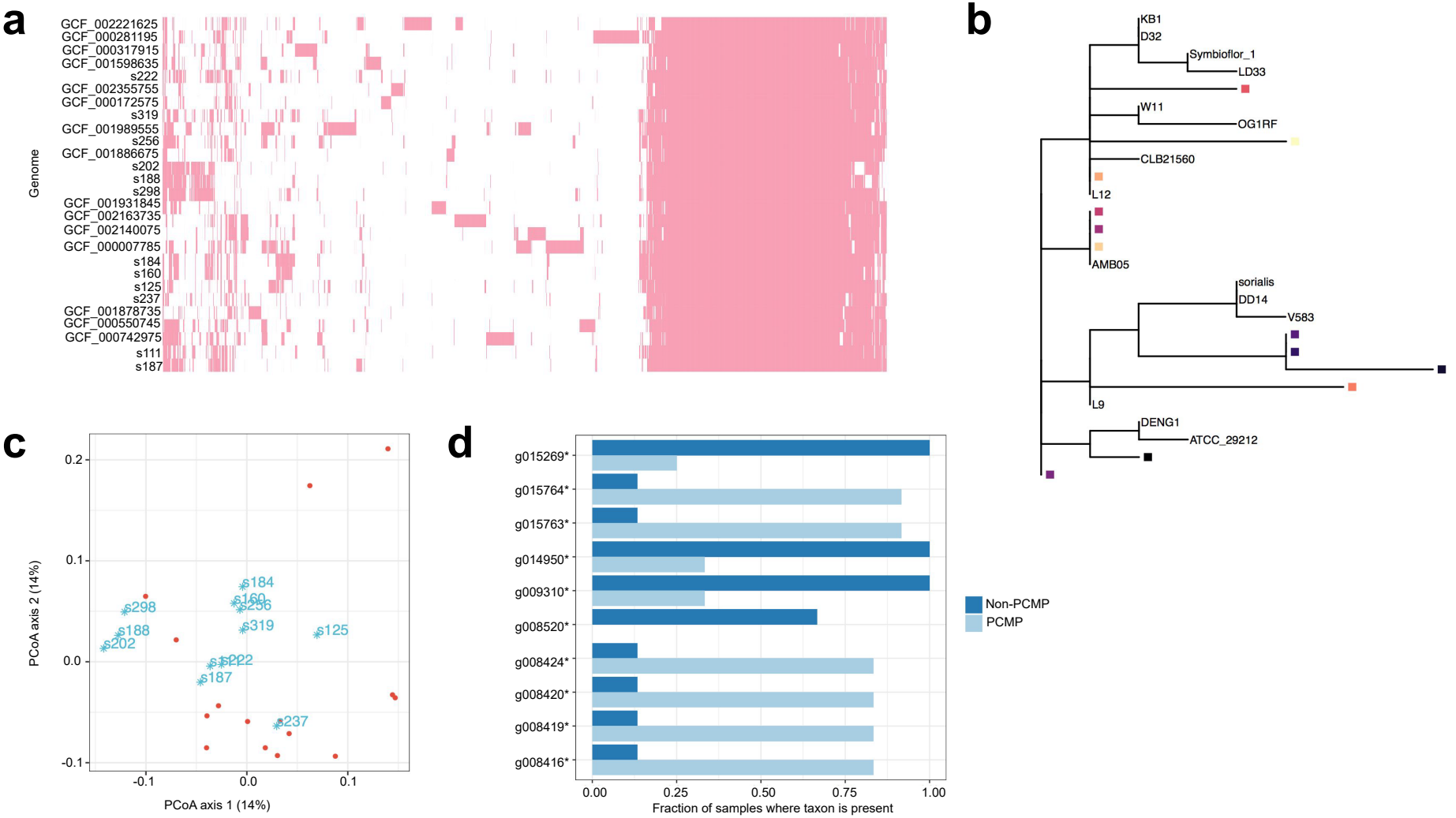


Supplementary Figure 6: **Consensus among core gene sequences in *E. coli* metagenomes.** Fraction of reads not matching to the consensus sequence in *E. coli* single-copy core genes. If a single strain were present, mismatches would be introduced by sequencing error alone, at a rate of approximately 1 in 106 base pairs.

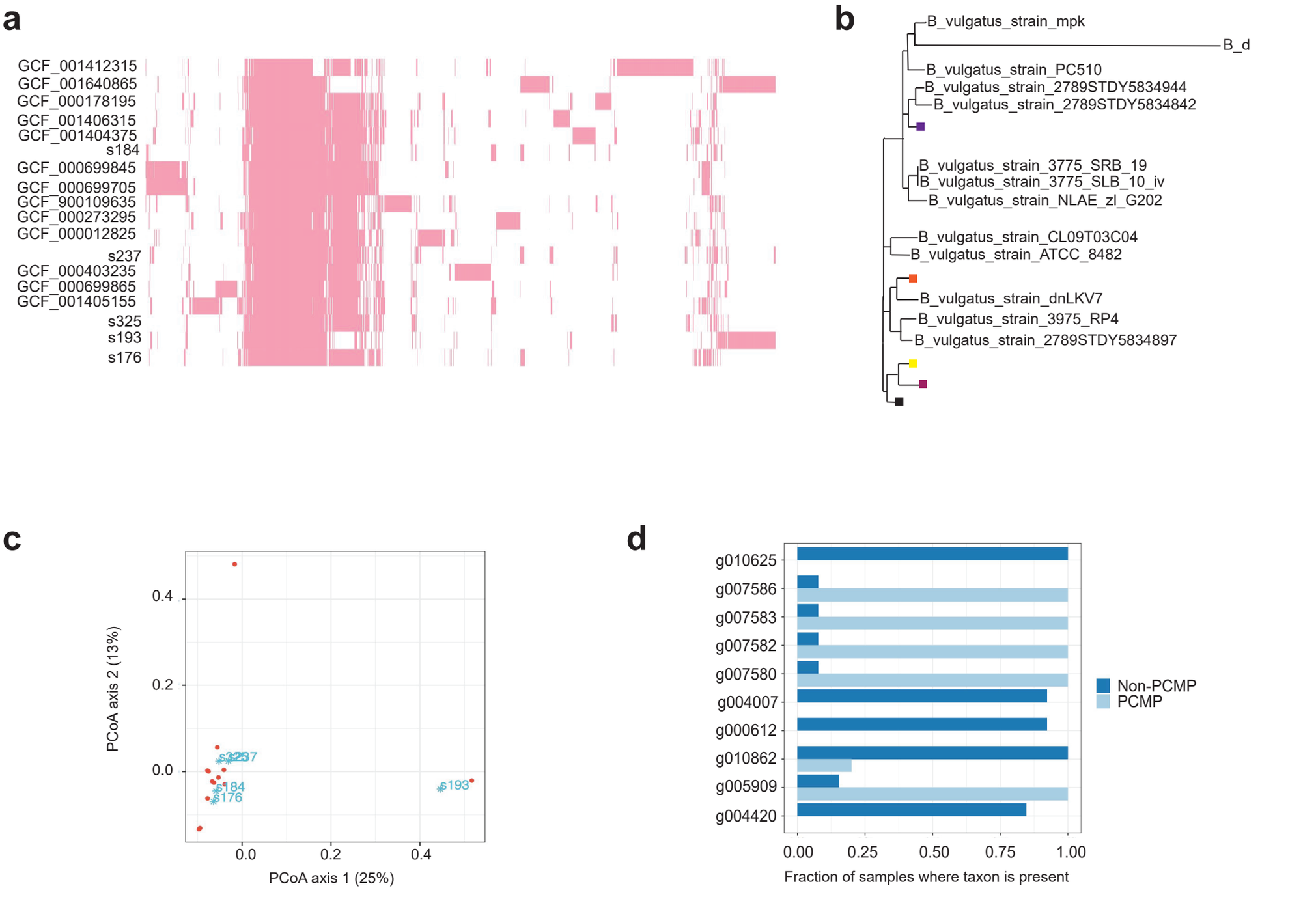


Supplementary Figure 7: **Strain number estimation in *E. coli* metagenomes.** Histogram of mismatch rate across all single-copy core genes in *E. coli* metagenome assemblies. The distributions were tested for consistency with the presence of two strains. Under this model, the gene sequence of the less abundant strain would introduce mismatches to the consensus sequence at a constant rate, wherever the genomes differed.

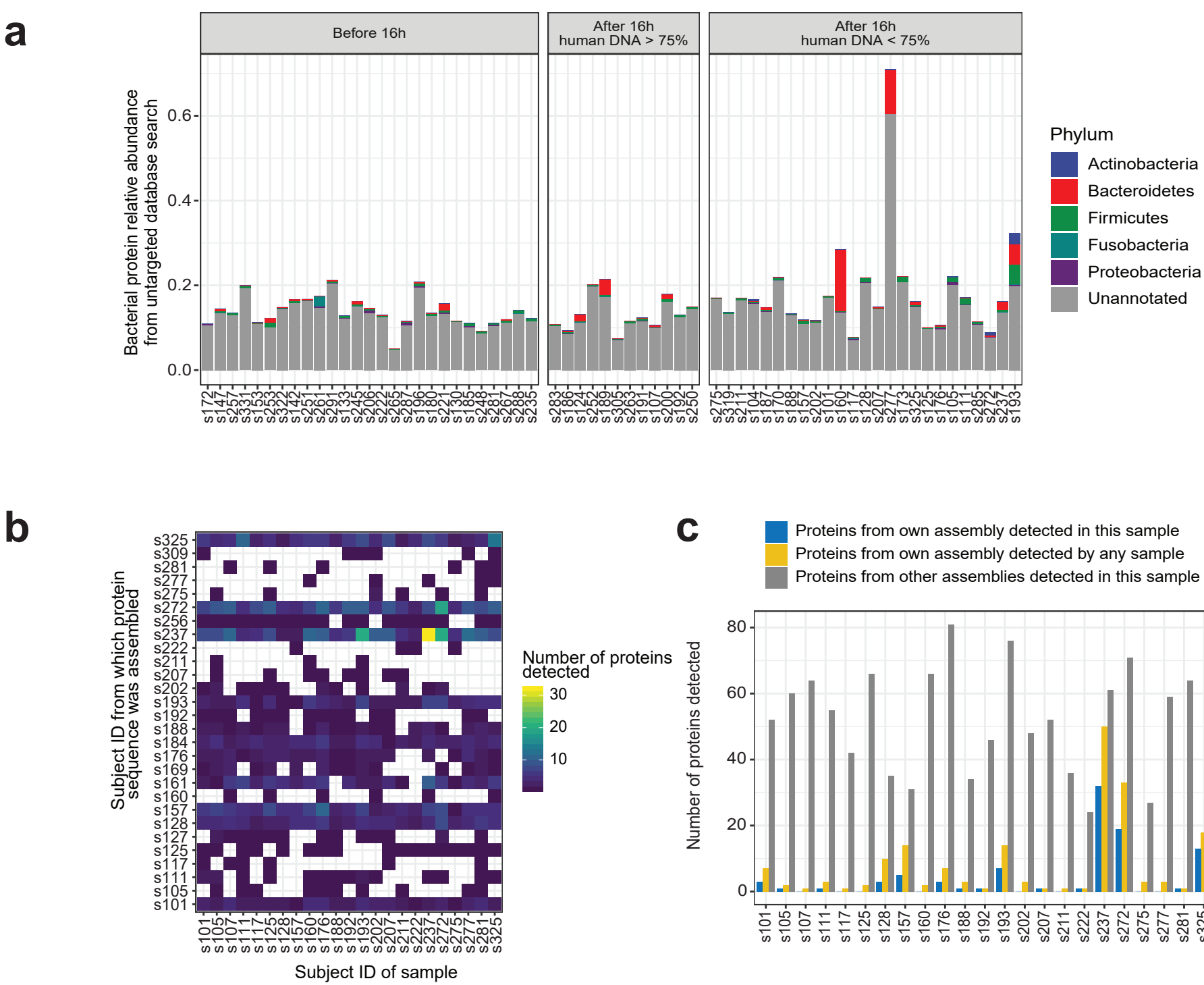
Supplementary Figure 8: **Expanded phylogenetic tree of *E. coli* metagenomes.** *E. coli* metagenomes assembled from meconium samples were combined with a set of 269 reference genomes, and the set of core gene were subjected to phylogenetic analysis. Colors represent *E. coli* phylogenetic groups. Genomes from this study are labeled with the subject identifier and colored magenta.



Supplementary Figure 9: **Analysis of *E. faecalis* metagenomes.** (a) Heatmap of genes present in the pan-genome of 12 *E. faecalis* metagenomes assembled from meconium samples, and 15 reference genomes. Genes present in each genome are shown in pink. Genomes from this study are labeled with the subject identifier; reference genomes are labeled with the NCBI genome identifier. (b) Phylogenetic tree estimated from core gene alignments. Genomes from this study are indicated by colored boxes. (c) Principal coordinates analysis of pan-genome content, quantified by Jaccard distance. Genomes from this study are shown in blue and labeled with the subject identifier; reference genomes are shown in red (n_1=12 genomes from this study, n_2=15 reference genomes). (d) Genes differentially present or absent in meconium samples, relative to the reference genome collection. Gene frequency from this study ("PCMP") is shown in light blue; reference genomes are shown in dark blue.

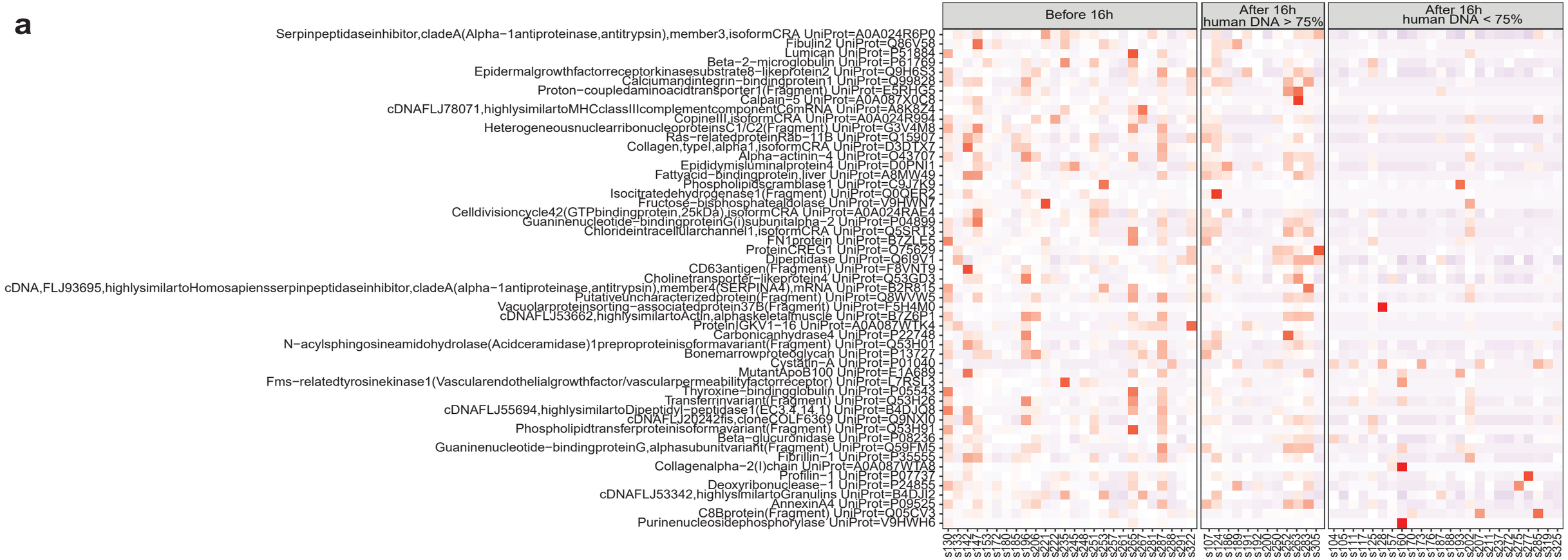


Supplementary Figure 10: **Analysis of *B. vulgatus* metagenomes.** (a) Heatmap of genes present in the pan-genome of 5 *B. vulgatus* metagenomes assembled from meconium samples, and 13 reference genomes. Genes present in each genome are shown in pink. Genomes from this study are labeled with the subject identifier; reference genomes are labeled with the NCBI genome identifier. (b) Phylogenetic tree estimated from core gene alignments. Genomes from this study are indicated by colored boxes. (c) Principal coordinates analysis of pan-genome content, quantified by Jaccard distance. Genomes from this study are shown in blue and labeled with the subject identifier; reference genomes are shown in red (n_1=5 genomes from this study, n_2=13 reference genomes). (d) Genes differentially present or absent in meconium samples, relative to the reference genome collection. Gene frequency from this study ("PCMP") is shown in light blue; reference genomes are shown in dark blue.

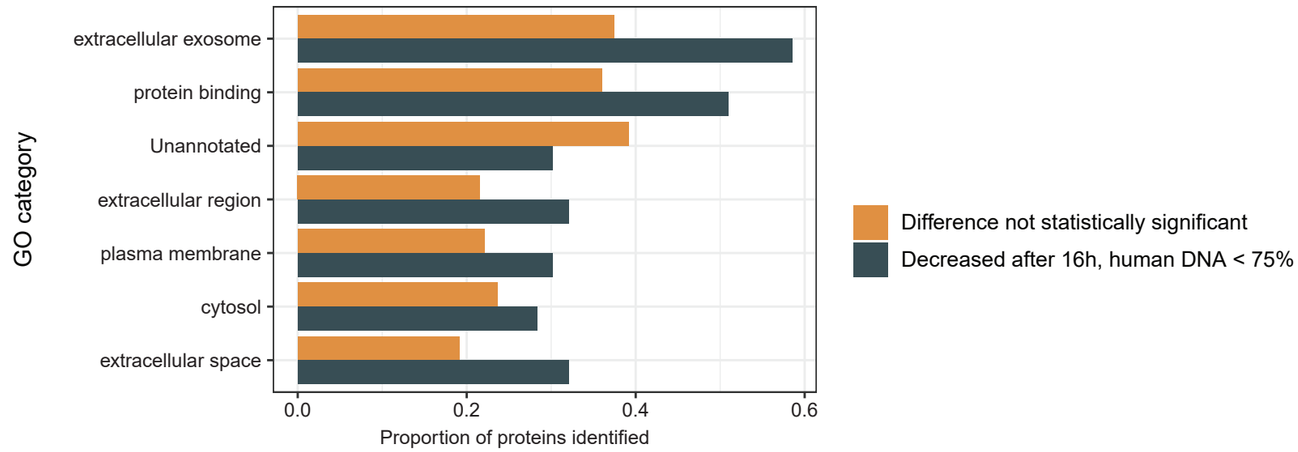


Supplementary Figure 11: **Analysis of bacterial proteins in proteomics results.** (a) When a large, untargeted database was used, only a small fraction of proteins were attributed to a bacterial phylum for analysis. (b) When a custom database of protein sequences assembled from our shotgun metagenomic data was used, the small number of bacterial proteins identified did not correspond to the sample from which the protein was assembled. (c) For each sample in the custom database, more proteins were detected from other sample assemblies than from the corresponding sample assembly.

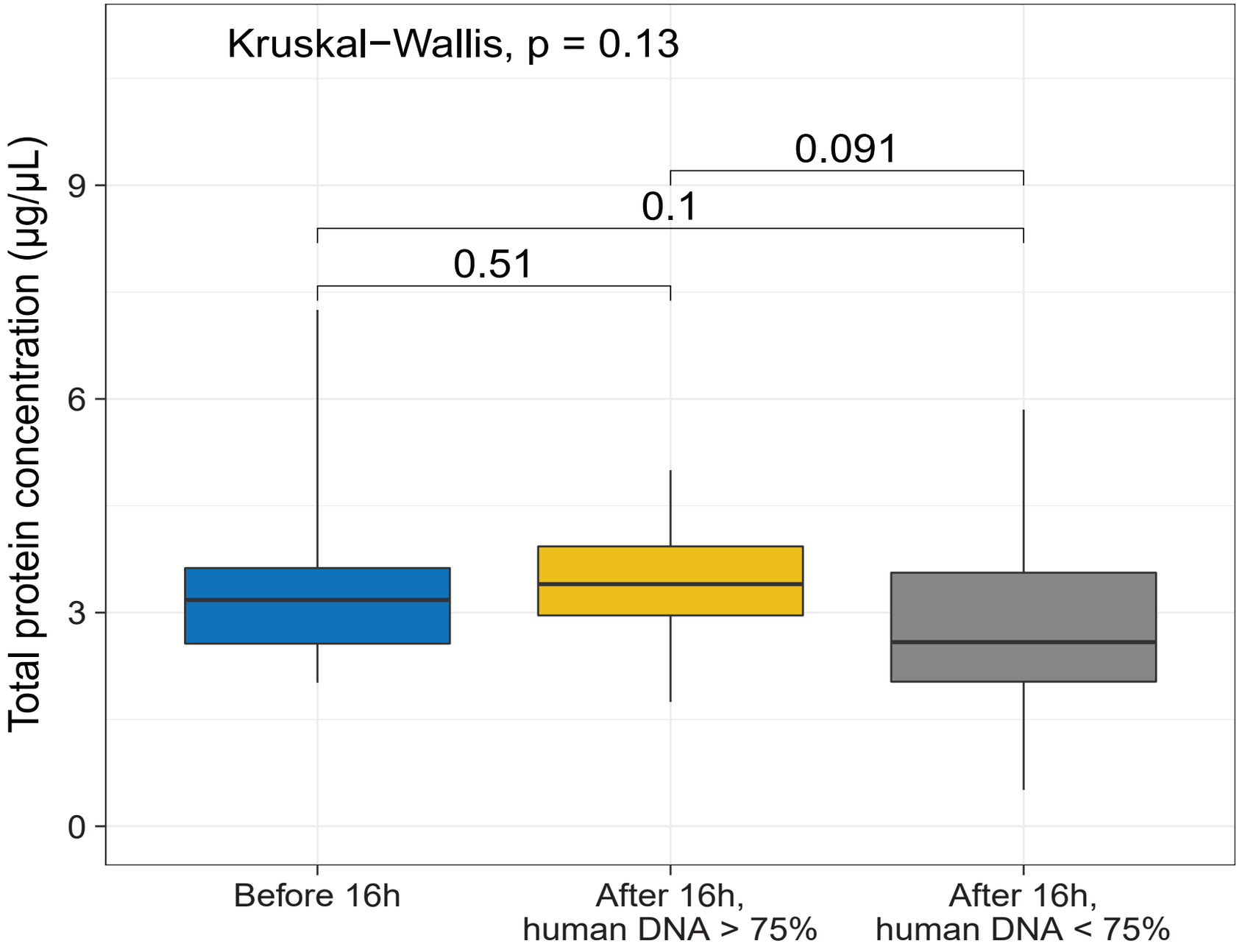
a



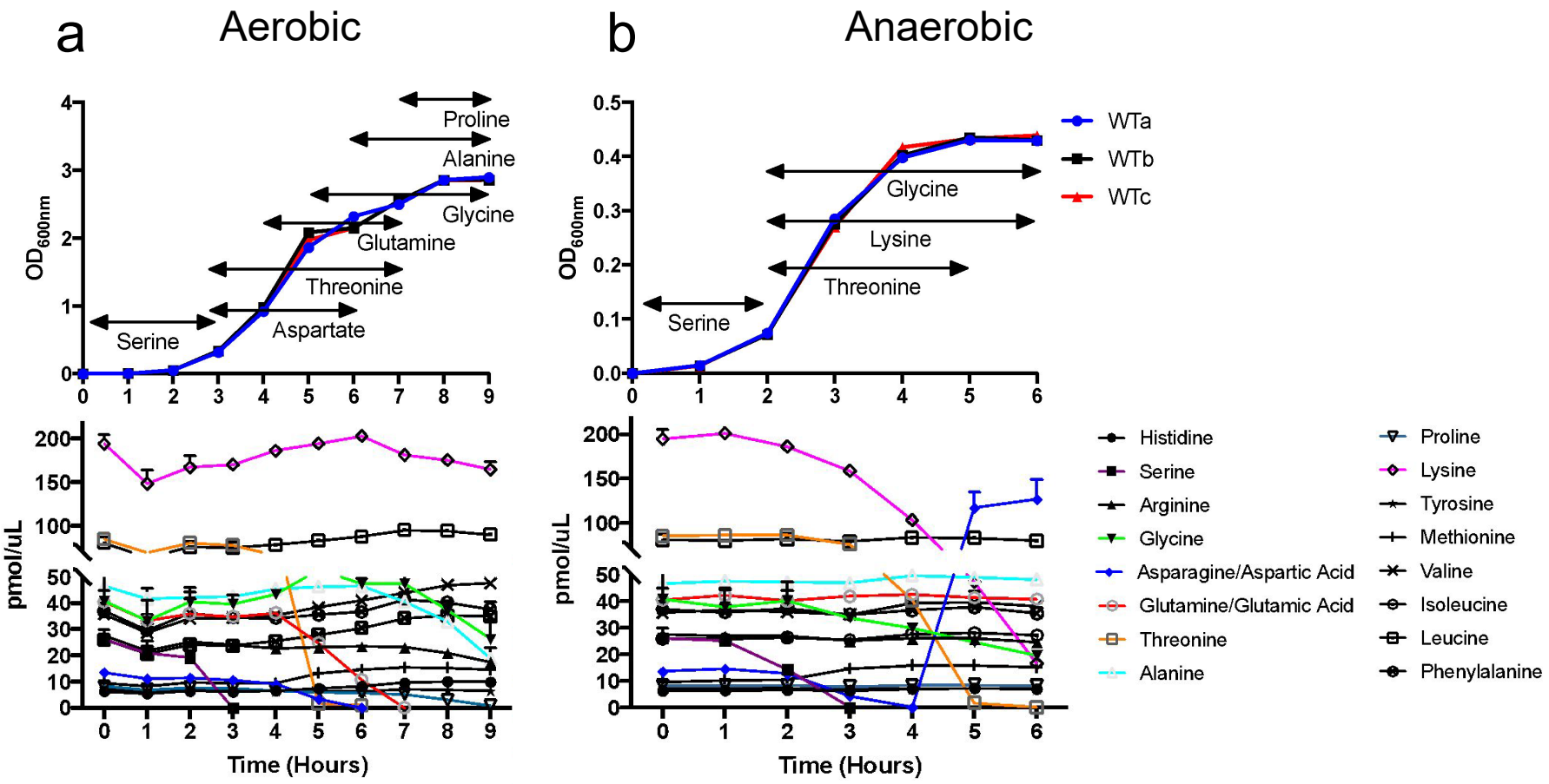
b



Supplementary Figure 12: **Proteins found to be differentially abundant between groups.** (a) The heatmap corresponds exactly to that in Figure 4C, only protein and subject identifiers are indicated. (b) Proteins identified as decreased after 16 hours with low human DNA were more likely to be assigned GO category of extracellular exosome (two-sided Fisher's exact test, P=0.02 after correction for false discovery rate, n_1 = 983 not identified as decreasing, n_2 = 53 identified as decreasing).



Supplementary Figure 13: **Total protein concentration.** The total protein concentration was not different between groups (Kruskal-Wallis test, $n_1 = 26$ before 16h, $n_2 = 12$ after 16h with human DNA > 75%, $n_3 = 24$ after 16h with human DNA < 75%). Boxes in indicate the median and interquartile distance, whiskers extend to the full range of the data.



Supplementary Figure 14: **Amino acid utilization by *E. coli* in culture.** Growth curves (top) and concentration of amino acids (bottom) of *E. coli* grown under (a) aerobic or (b) anaerobic conditions. Growth curves were obtained for three biological replicates of wild type (WT) E. coli: WTa, WTb, and WTc. Arrows indicate the range of time during which the amino acid decreases in concentration.