Supplementary Figure Legends

Supplementary Figure 1, related to Star Methods: Synthetic data modeling pipeline. We generated synthetic data from complete genomes of oral microbes, downloaded from the Human Oral Microbiome Database. We downloaded 467 genomes and used ART to generated synthetic read data from each one at 1X coverage. We then selected 95 of these genomes (the average number of species found across the samples in our oral microbiota dataset) and combined them into a synthetic metagenome at varying levels of coverage. We tried 3 coverage ranges (0 to 1X, 0 to 10X, and 10 to 20X) and created ten metagenomes per coverage range. We assembled each metagenome, predicted genes using PROKKA, identified false positives by alignments back to the 95 input genomes, and carried out three analyses. 1) Summary analysis: averaging gene and contig summary statistics across each iteration and identifying linear associations between false discovery rate and gene/contig features (i.e. length, coverage) 2) Gene-by-gene false positive analysis: a logistic regression-based analysis to determine if the odds of a gene being a false positive was contingent upon gene/contig coverage/length. 3) Gene-by-gene singleton analysis. We clustered the 10 synthetic metagenomes from each coverage range at the 50% identity level and modeled, again with logistic regression, the association between a gene being a singleton and a false positive.

Supplementary Figure 2, related to Star Methods: Visualization of synthetic data analysis. A-C) False discovery rate as a function of assembler parameters/type. A) Low coverage data (0-1X per genome in metagenome) B) Low-medium coverage (0-10X) C) High coverage (10-20X). D) Distributions of false and true positive genes within synthetic data. We show gene/contig length/coverage as a function of total sample sequencing depth and assembler type. E) Distributions of singleton and non-singleton genes within synthetic data. We show gene/contig length/coverage as a function of total sample sequencing depth and assembler type.

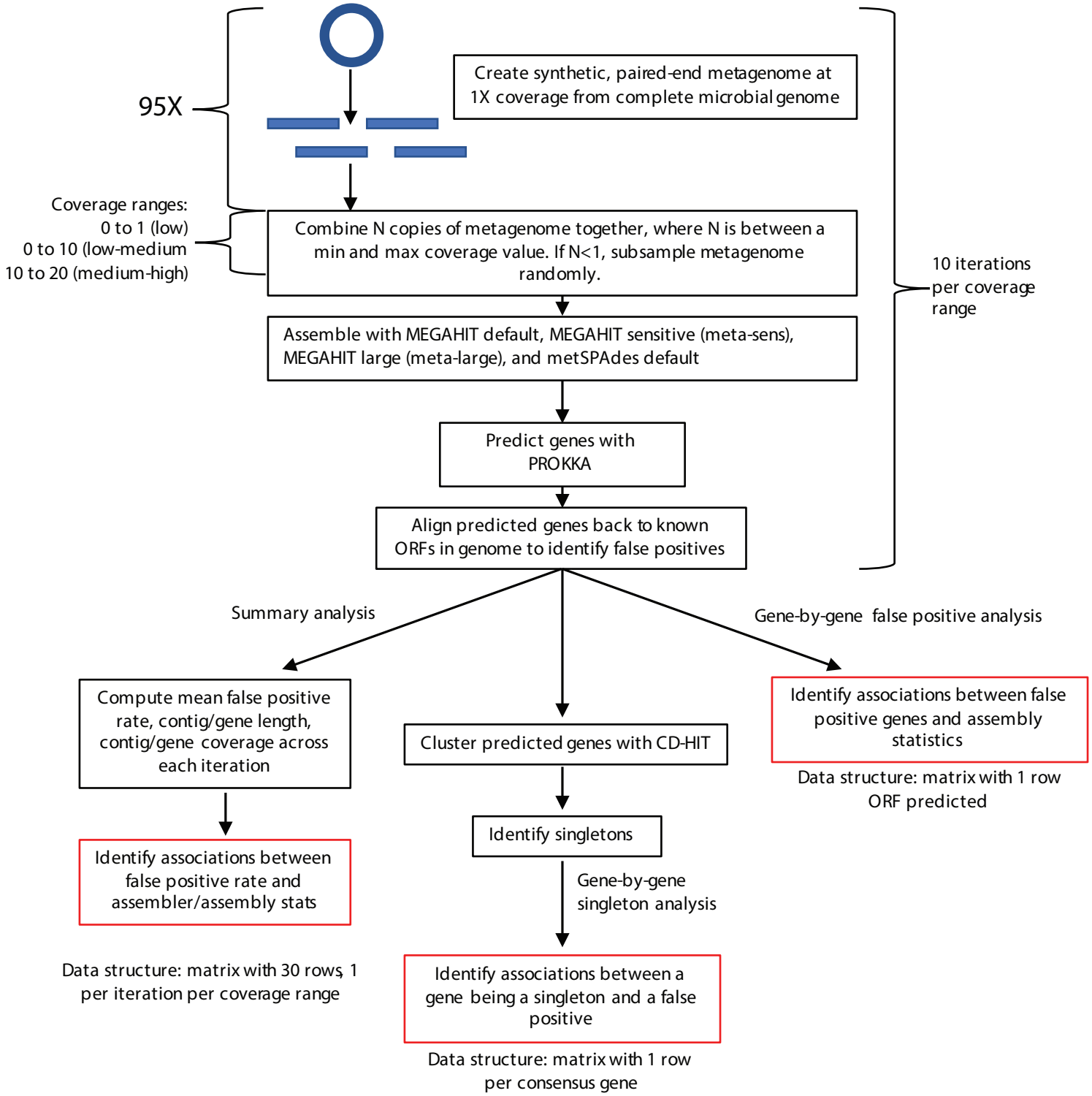
Supplementary Figure 3, related to Star Methods: Further testing the relationship between gene/contig length/coverage, sample sequencing depth, and gene singleton status. A) Distribution of oral coverage data by singleton status. B) Distribution of oral gene lengths by singleton status. C) Distribution of oral contig lengths by singleton status. D) Distribution of gut gene lengths by singleton status. E) Distribution of gut contig lengths by singleton status. F) Comparing the gene cluster sizes in the Metahit gene catalog (IGC) versus our own gut gene catalog. G-J) The relationship between singleton genes and depth of sequencing. The count (G) and fraction (H) of singleton genes per sample. I-J) Scatter plots showing the relationship (Spearman correlations and associated p-values) between total reads in a sample (sum of paired/single ended) and the total fraction/count of singletons for that sample.

Supplementary Figure 4, related to Figure 4 and Figure 5: Functional and taxonomic enrichment of singleton/non-singleton genes/contigs. A) Results of gene-level functional comparisons. Each point represents a unique minpath annotation. X-axis corresponds to total number of non-singleton genes annotated to a specific function, with the y-axis corresponds to the analogous value for singletons. Points above the horizontal line are enriched in singletons, below in non-singletons. B) Results from contig-level taxonomic binning analysis. Each point represents a different taxon, with the axes corresponding to counts of contigs consisting exclusively of singletons or non-singletons mapping to a specific taxa. C-D) Taxonomic annotation efficiency. Fraction of taxonomically annotated singletons (C) and non-singletons (D). E-L) The correlation between singleton and non-singleton taxonomies as as function of contig length and ORF number. We display here plots of normalized taxonomic counts for singleton and non-singleton taxonomies. Each point represents a different genus/species. Top row (E-H): only plotting contigs with at least 2 genes. Bottom row (I-L): only plotting contigs with at least 3 genes. Column 1 (E,I): min contig length of 500bp. Column 2 (F,J): min contig length of 1000bp. Column 3 (G,K): min contig length of 1500bp. Column 4 (H,L): min contig length of 2000 bp.

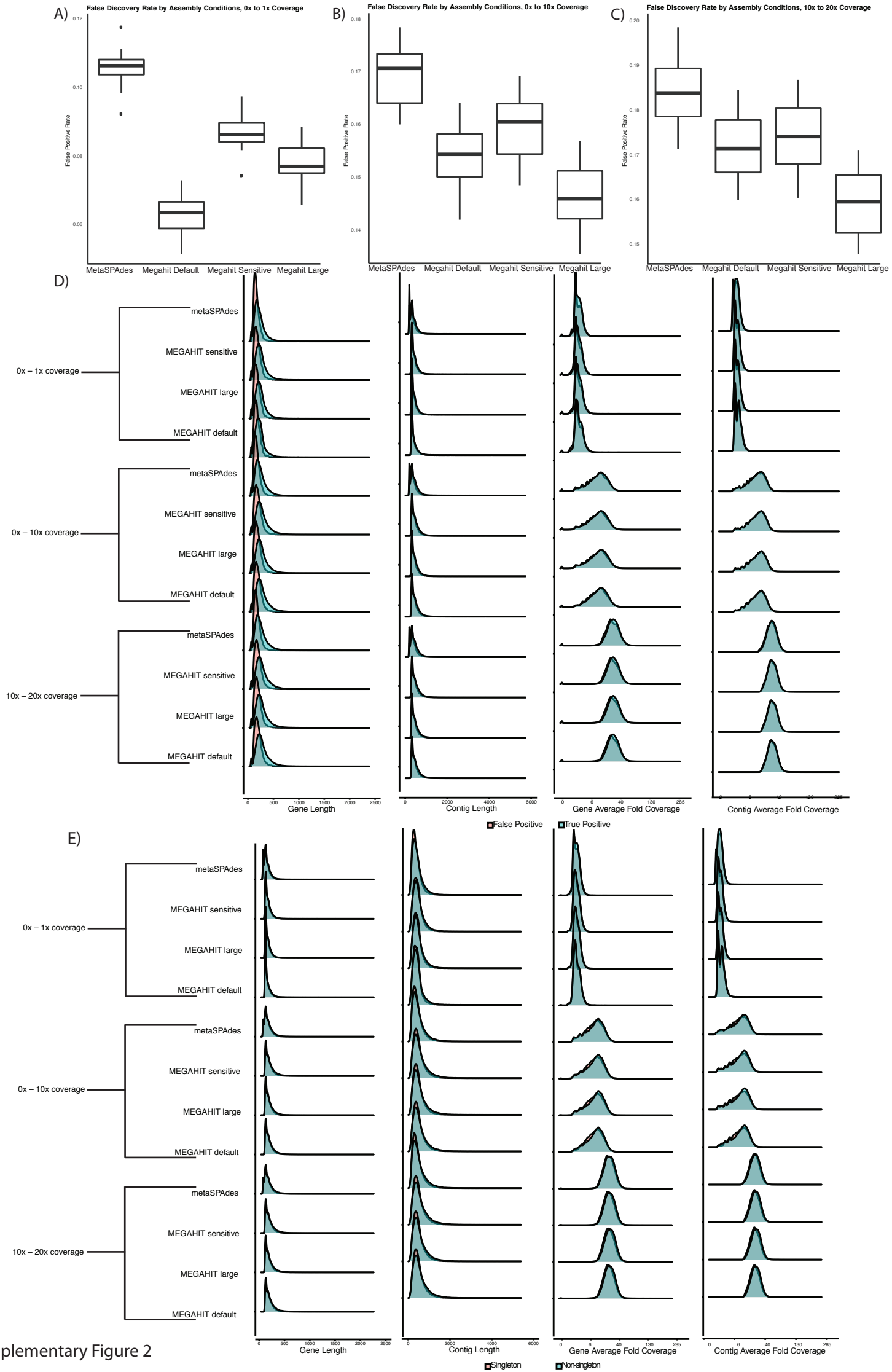
Supplementary Figure 5, related to Figure 5: Enrichment of species/genera in gut/oral niches for singletons/non-singletons. Here we display the top 25 most enriched species for oral singletons (A), oral non-singletons (B), gut singletons (C), and gut non-singletons (D) as well as the top 10 most enriched genera for oral singletons (E), oral non-singletons (F), gut singletons (G), and gut non-singletons (H).

Supplementary Figure 6, related to Figure 6: Rarefaction and singleton fraction curves, oral and gut. In A-B) we display the rarefaction curves in the oral (A) and gut (B) microbiomes for the accumulation of genes. Overall rarefaction curves are blue, singleton curves are black, and non-singleton curves are grey. C-D) The log of the fraction of singletons over time from our data (green lines), as well as the curve extrapolation fit to these data (red lines) for the oral (C) and gut (D) microbiomes.

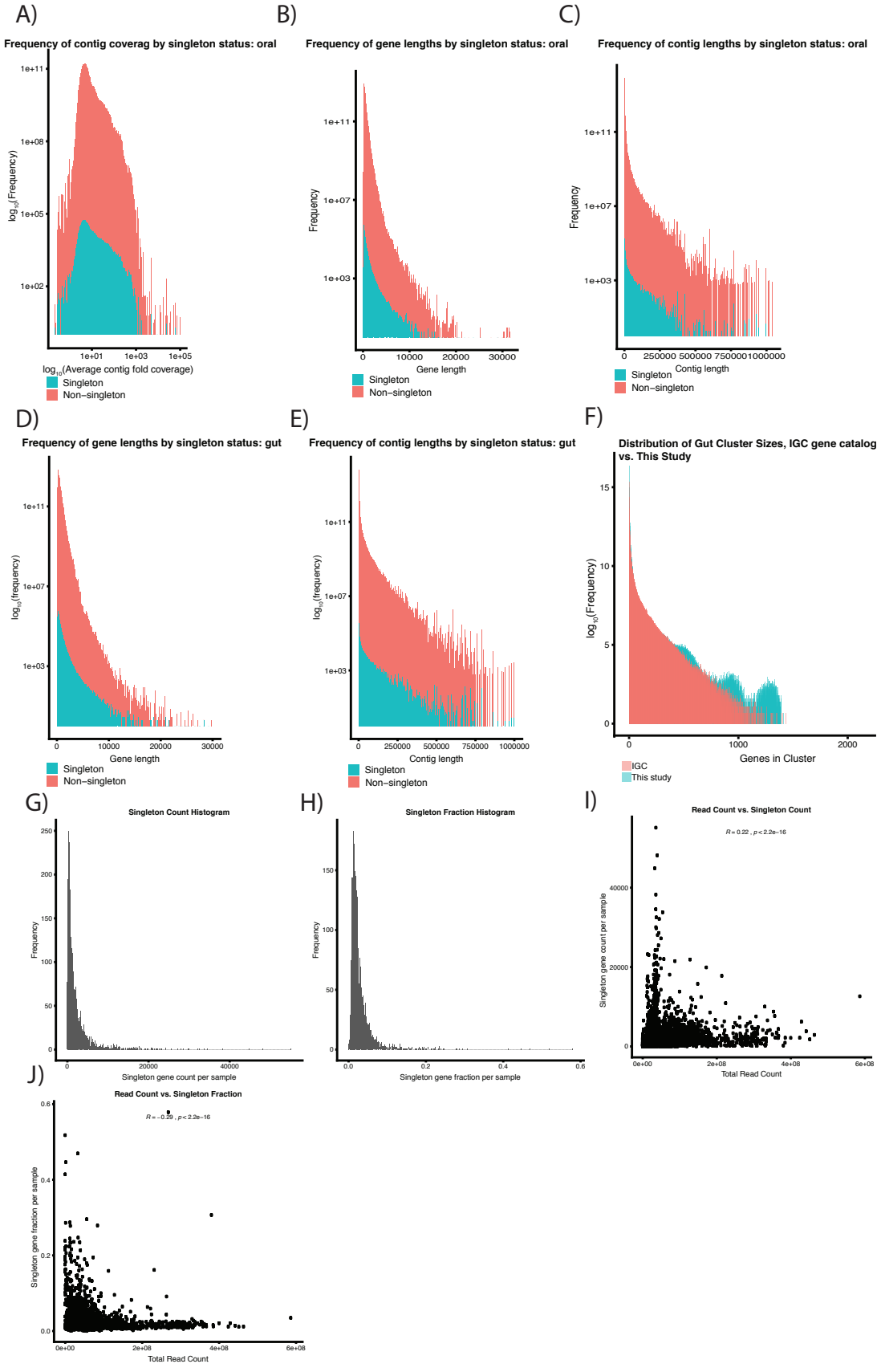
Supplementary Figures



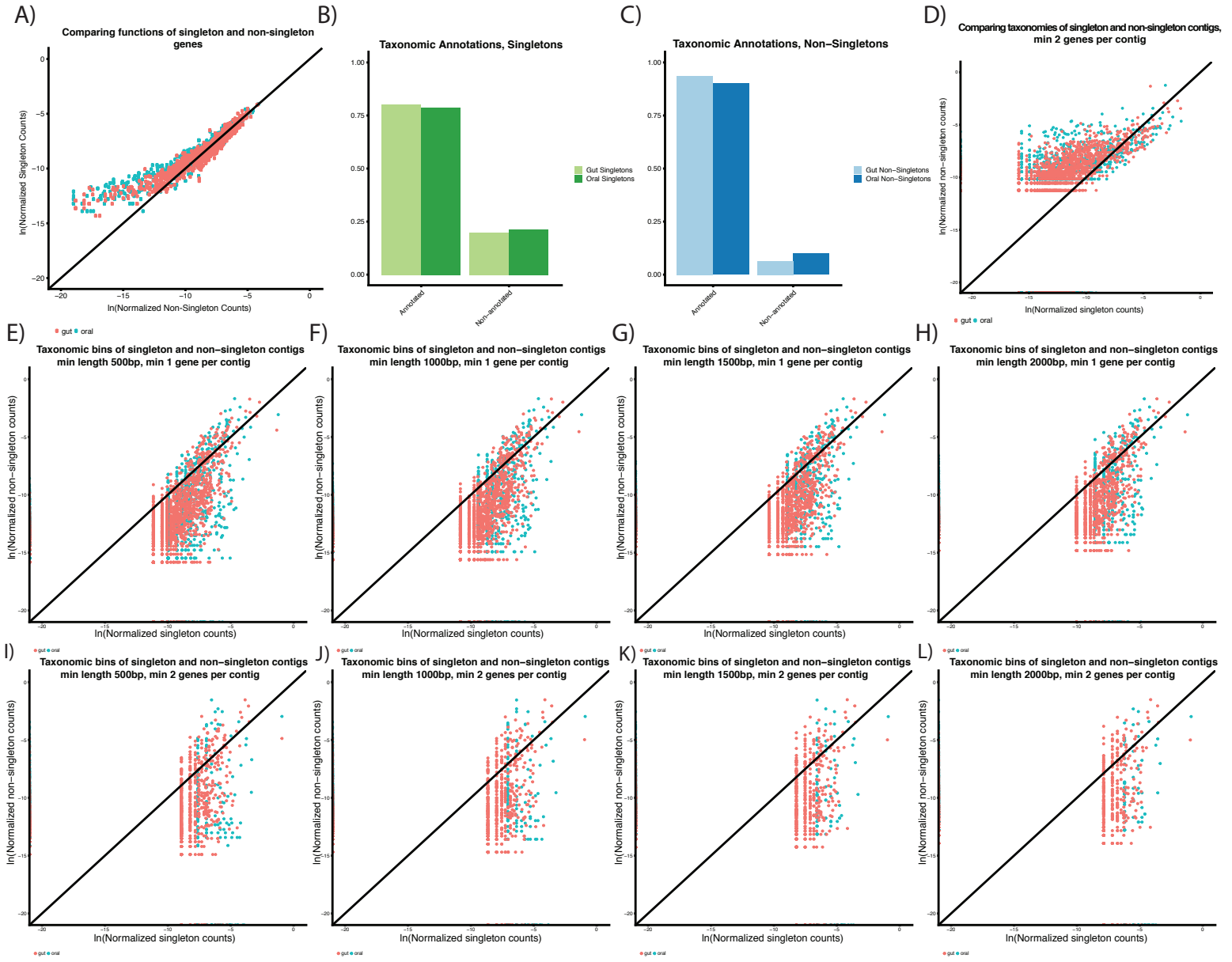
Supplementary Figure 1



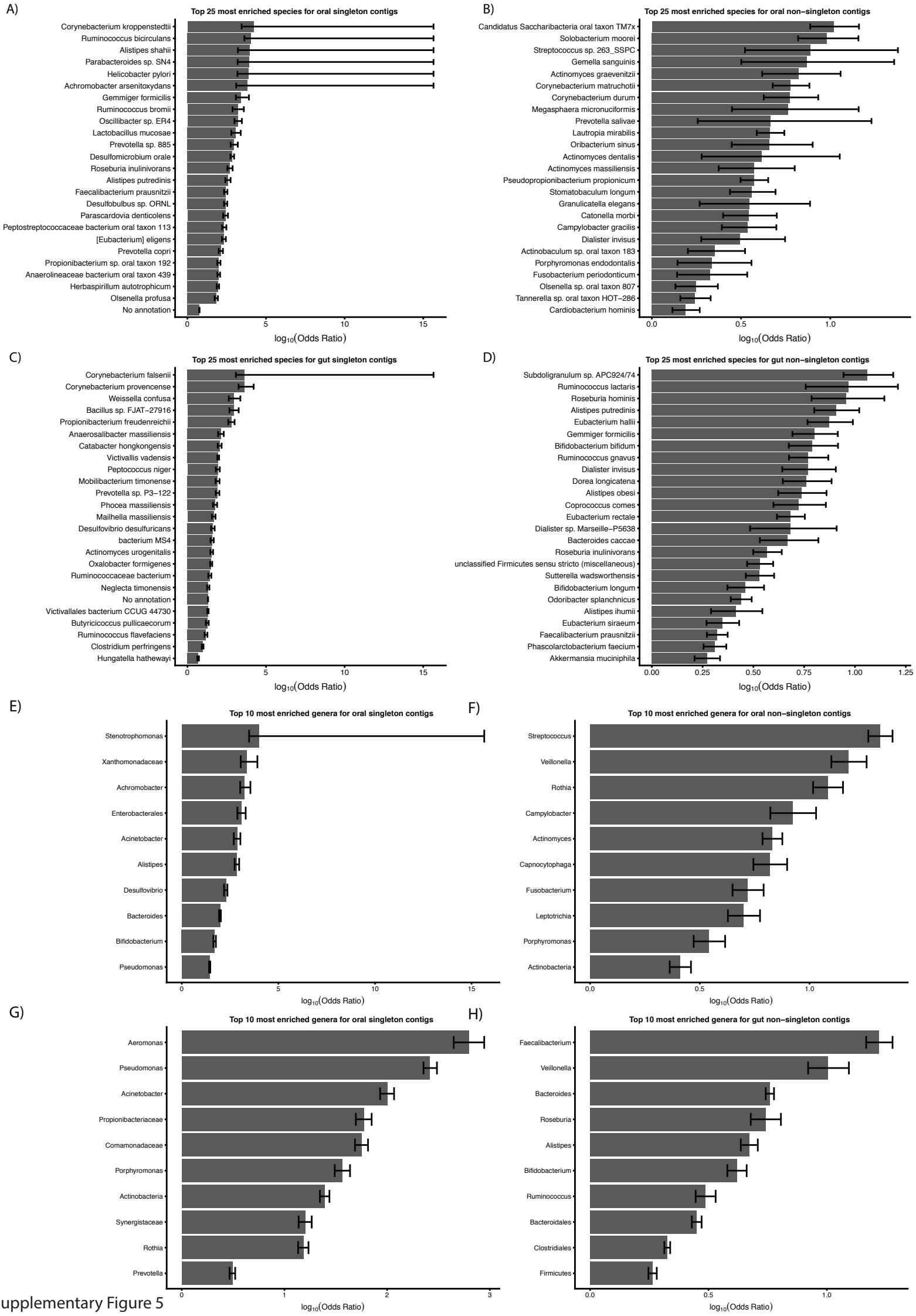
Supplementary Figure 2



Supplementary Figure 3

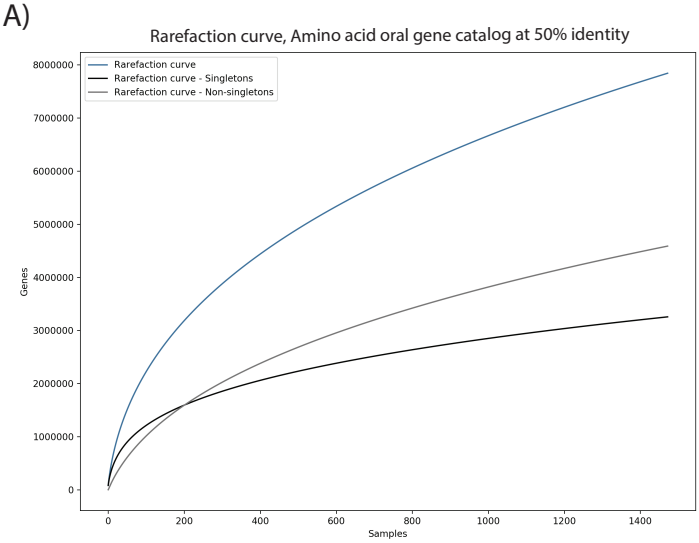


Supplementary Figure 4

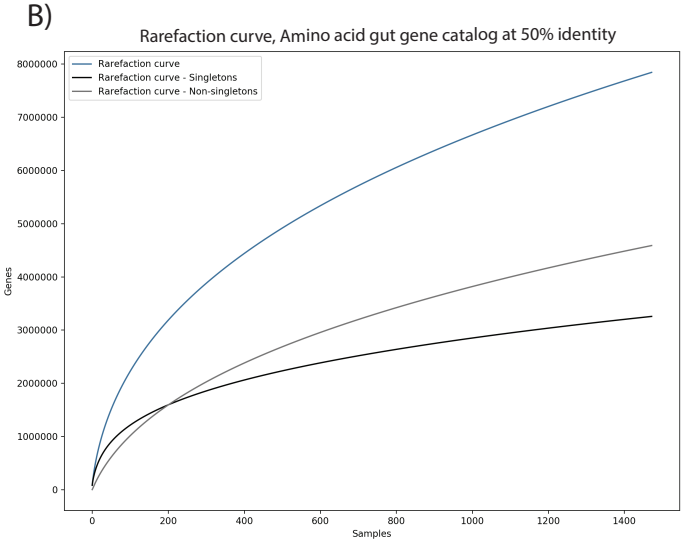


Supplementary Figure 5

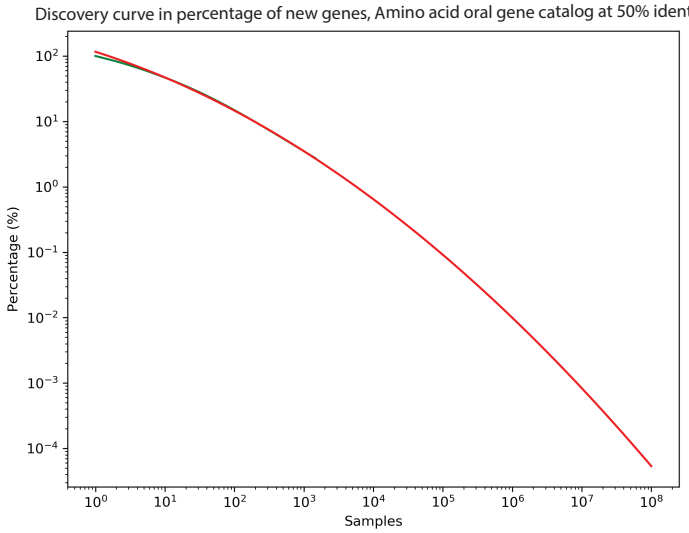
A)



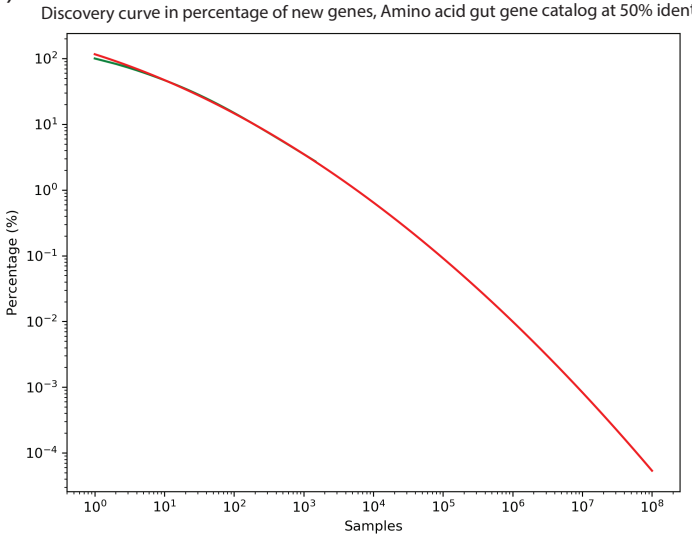
B)



C)



D)



Supplementary Figure 6