Supplementary Data for "Bacterial plasmid-associated and chromosomal proteins have fundamentally different properties in protein interaction networks"

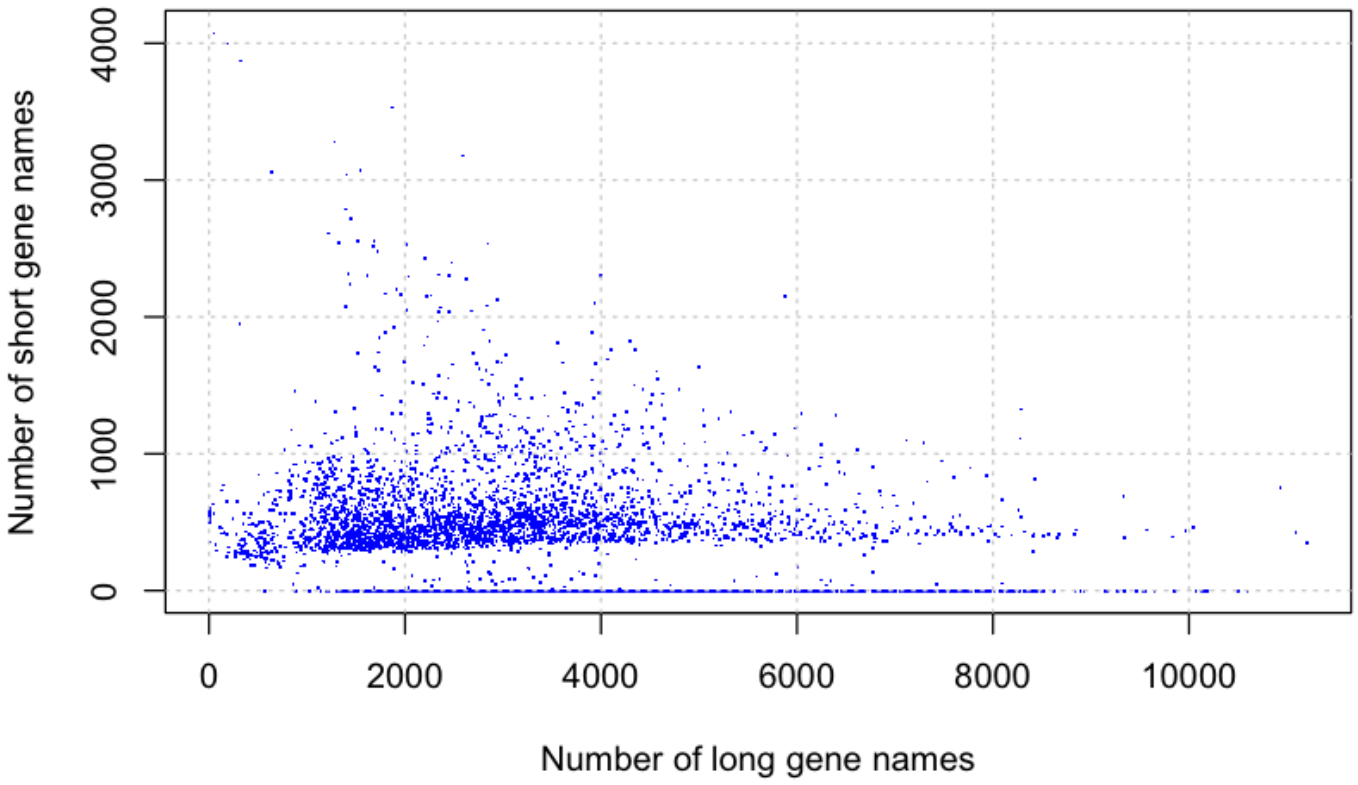

Fig S1. The number of long gene names (x-axis) compared to the number of short gene names (y-axis) per sample. The 4,429 samples had a median of 411±410 short names and a median of 2,931±1,744 long names each. Short gene names had four letters or less, and long ones had more than this.

Table S1. Each of the plasmid-related genes and their associated plasmid link(s) across 15,531 plasmid accessions. See full table FigShare doi: https://doi.org/10.6084/m9.figshare.19525453

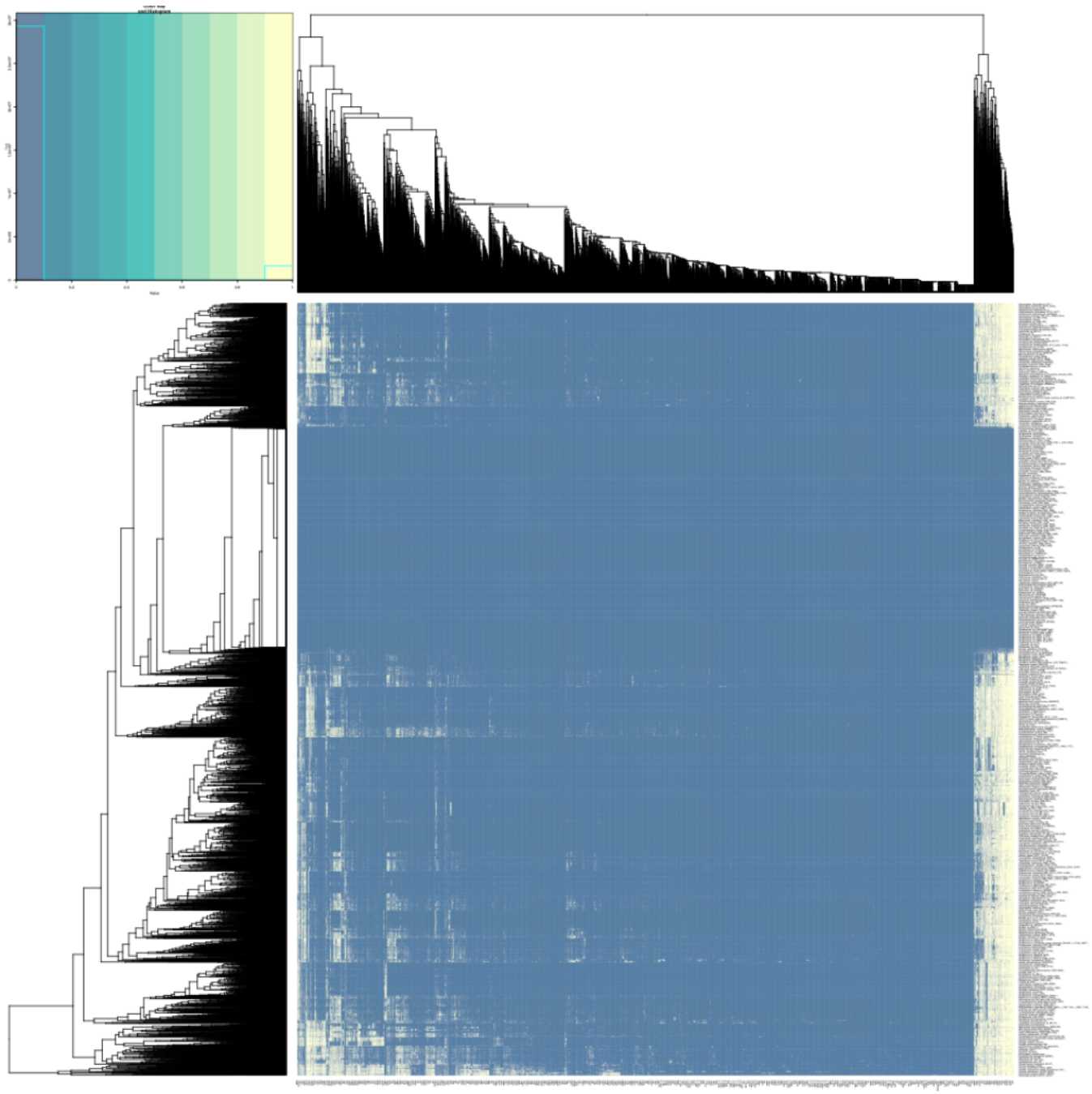

Figure S2. Heatmap visualising plasmid-related genes for 4,429 bacterial samples (y-axis) across 7,126 unique plasmid-derived genes (x-axis). Dendrograms indicate the similarity across the samples (left) and genes (top). Both axes were sorted based on similarity. Genes found to possess interactions are shown in yellow, and absence is show in blue. See full figure on FigShare at https://doi.org/10.6084/m9.figshare.19525453.

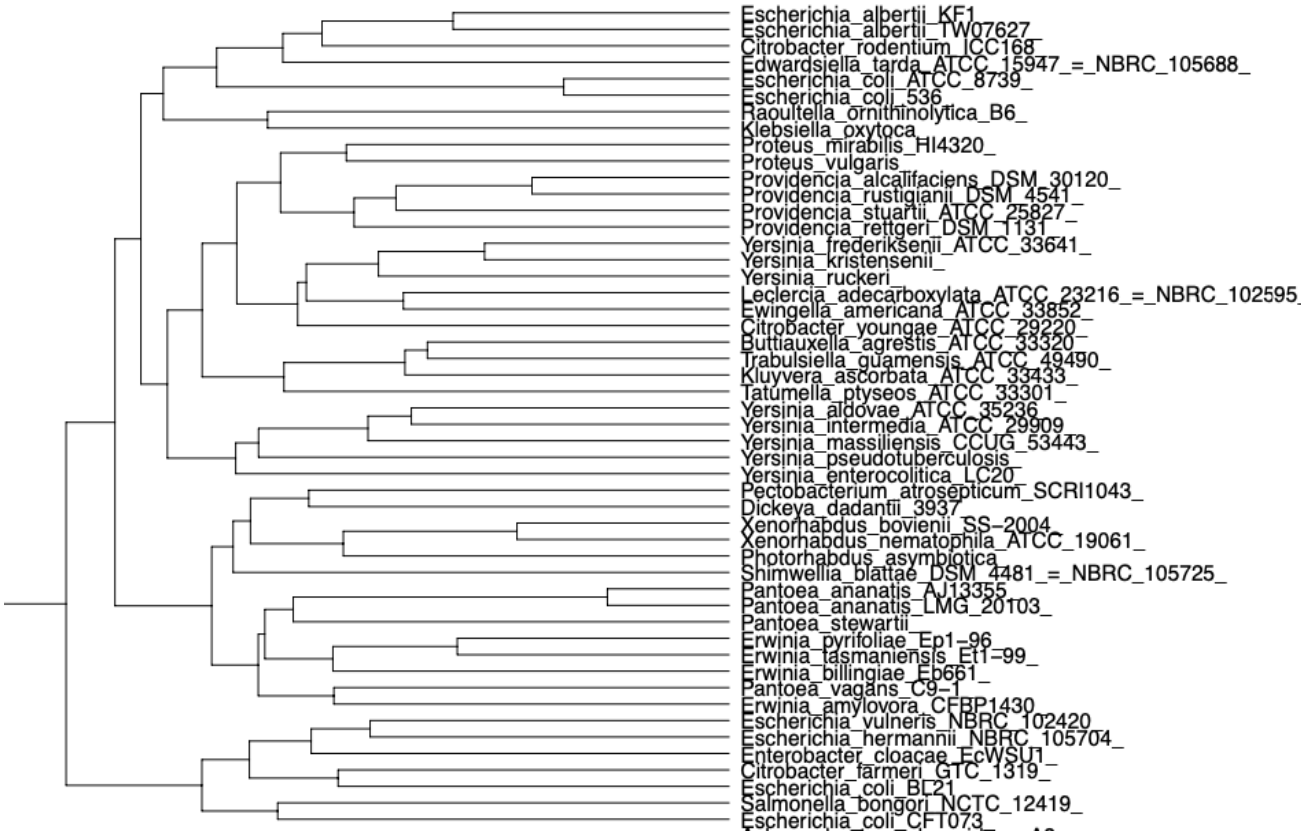

Figure S3. Dendrogram showing the similarity across 50 samples from the *Enterobacterales* order that were the most divergent clade among 3,146 bacteria with had 1+ plasmid-associated genes across 7,126 unique plasmid-derived genes. These corresponded to rank 6 in the NMF results.

Figure S4. PCA of the whole dataset showing the variation of: (A) the 4,429 samples based on their patterns across the 7,126 plasmid-associated genes, and (B) the 7,126 plasmid-associated genes based on their frequencies across the 4,429 samples. Dim1 indicates PC1 (29.2% across samples, 52.8% across genes), and Dim2 indicates PC2 (7.7% across samples, 4.6% across genes). Lower PCs had much less of the total variation. The colours are cosmetic and were to differentiate different points.

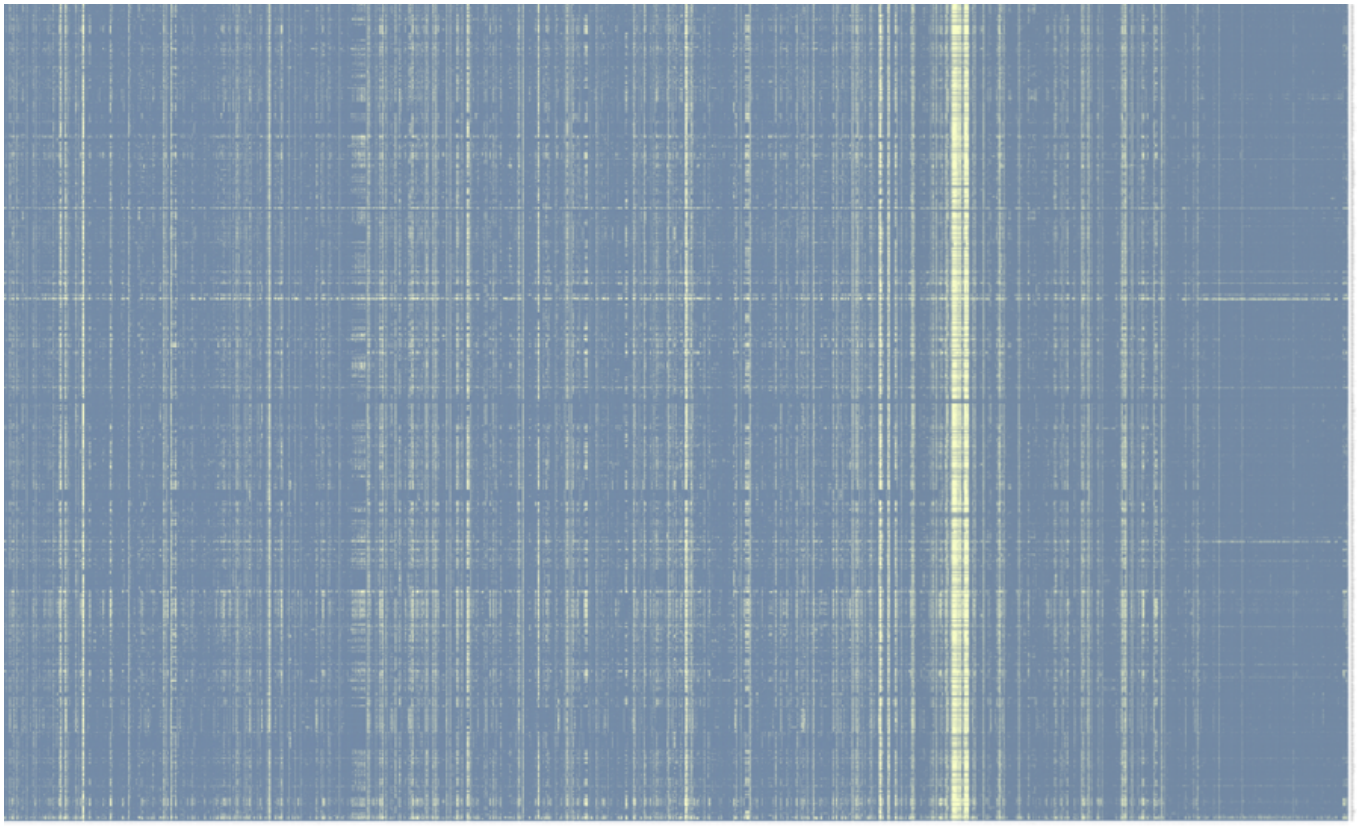

Figure S5. Heatmap visualising plasmid interactomes for 3,146 bacterial samples (y-axis) across 7,126 unique plasmid-derived genes (x-axis). The axes were sorted alphabetically. See full figure on FigShare at https://doi.org/10.6084/m9.figshare.19525465 and data at https://doi.org/10.6084/m9.figshare.19525471 .

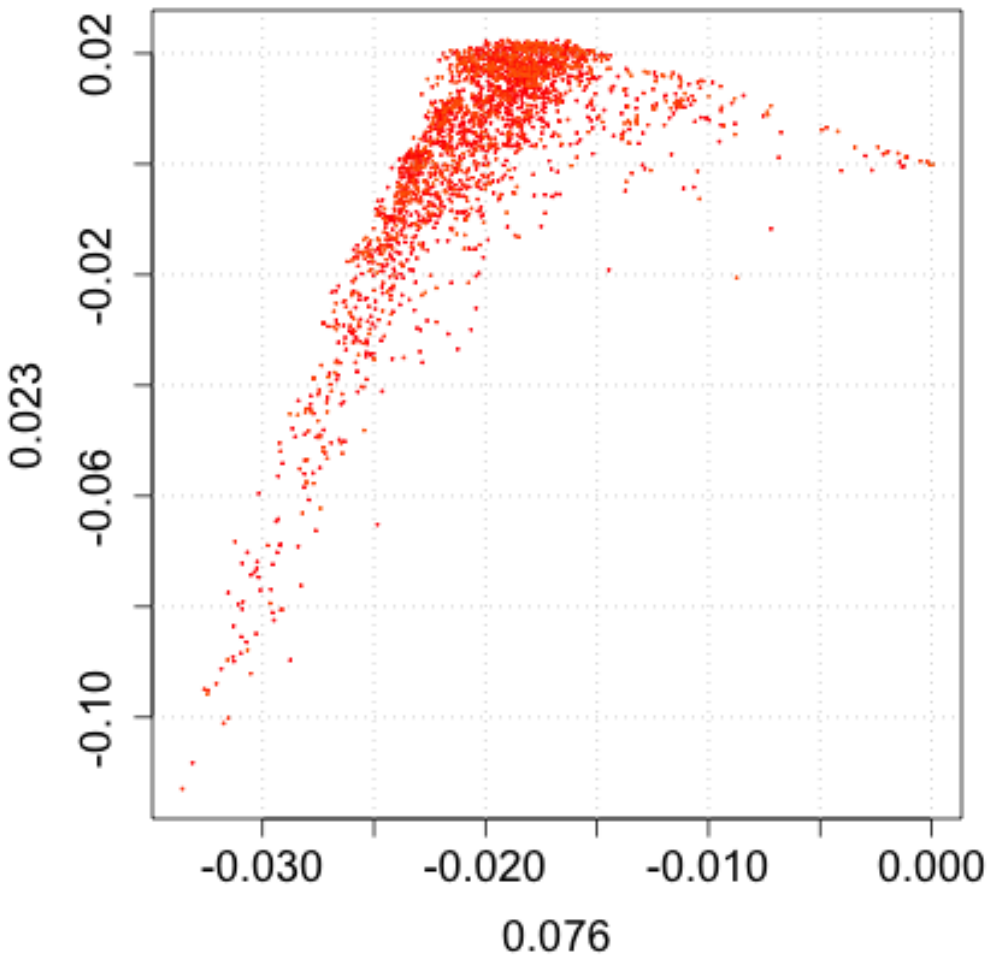

Figure S6. Visualisation of the 1st (x-axis, 7.6% of variation) and 2nd (y-axis, 2.3% of variation) dimensions from LSA of the 3,146 bacterial samples across 360 dimensions representing 7,126 unique plasmid-derived genes.

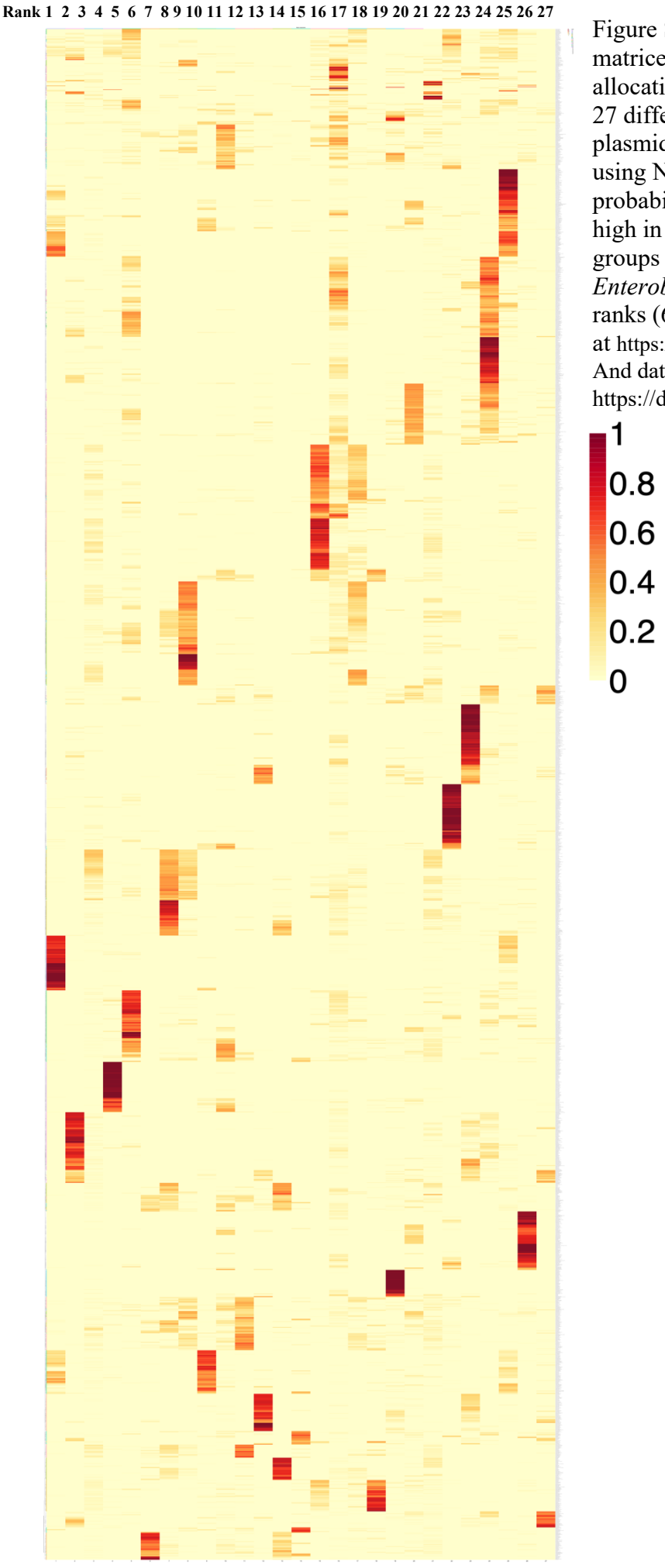

Figure S7. Heatmap of the basis coefficient matrices, reflecting the probability of allocation of the 3,164 samples (y-axis) to 27 different groups (x-axis) based on their plasmid gene presence-absence profiles using NMF. The legend below shows the probability of allocation (as per the legend: high in red, low in yellow) to the different groups ("basis"). Many of the *Enterobacterales* were allocated to the same ranks (6 and 14). See full figure on FigShare at https://doi.org/10.6084/m9.figshare.19525492 And data at https://doi.org/10.6084/m9.figshare.19525480 .

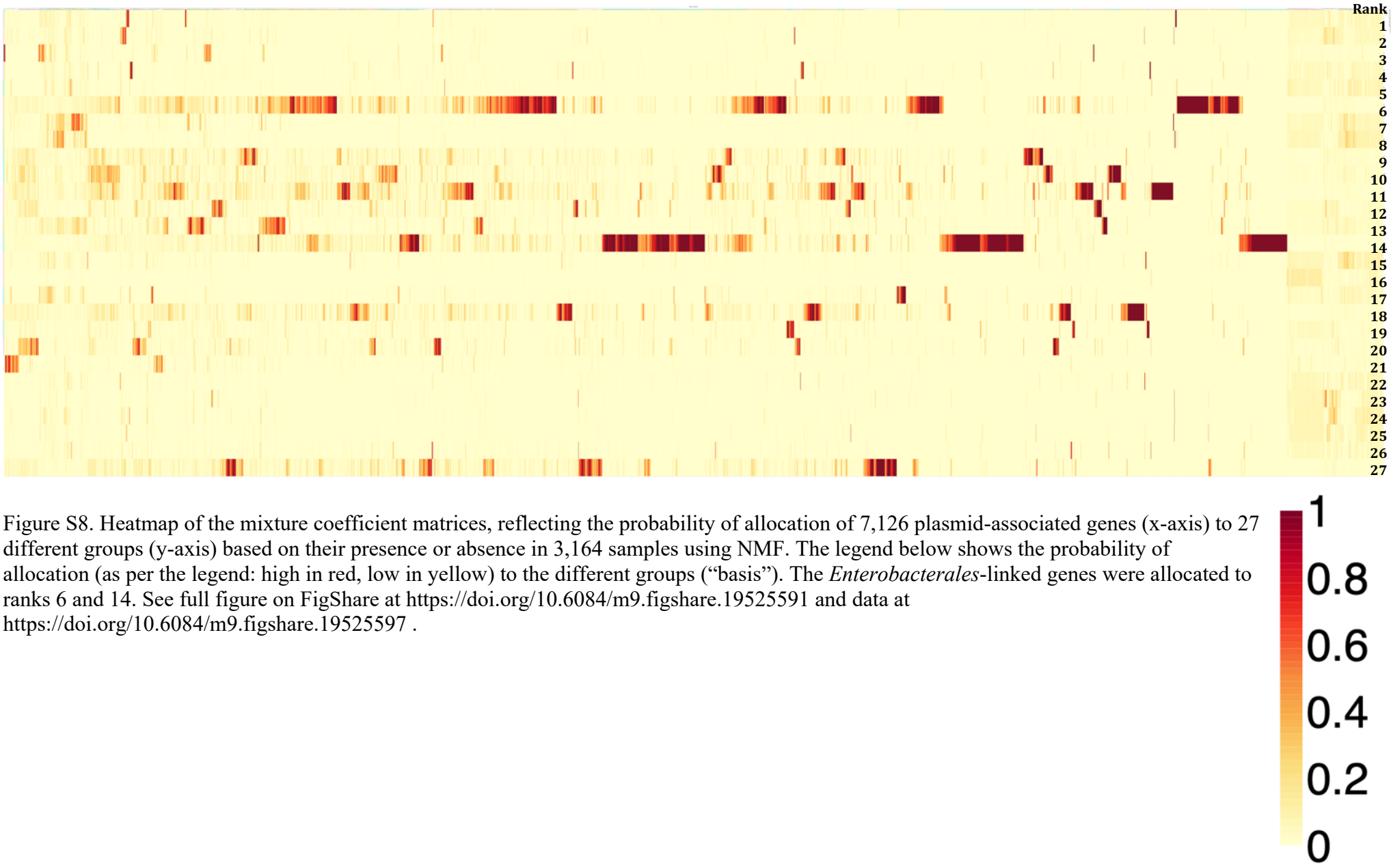

Figure S8. Heatmap of the mixture coefficient matrices, reflecting the probability of allocation of 7,126 plasmid-associated genes (x-axis) to 27 different groups (y-axis) based on their presence or absence in 3,164 samples using NMF. The legend below shows the probability of allocation (as per the legend: high in red, low in yellow) to the different groups ("basis"). The *Enterobacterales*-linked genes were allocated to ranks 6 and 14. See full figure on FigShare at https://doi.org/10.6084/m9.figshare.19525591 and data at https://doi.org/10.6084/m9.figshare.19525597 .

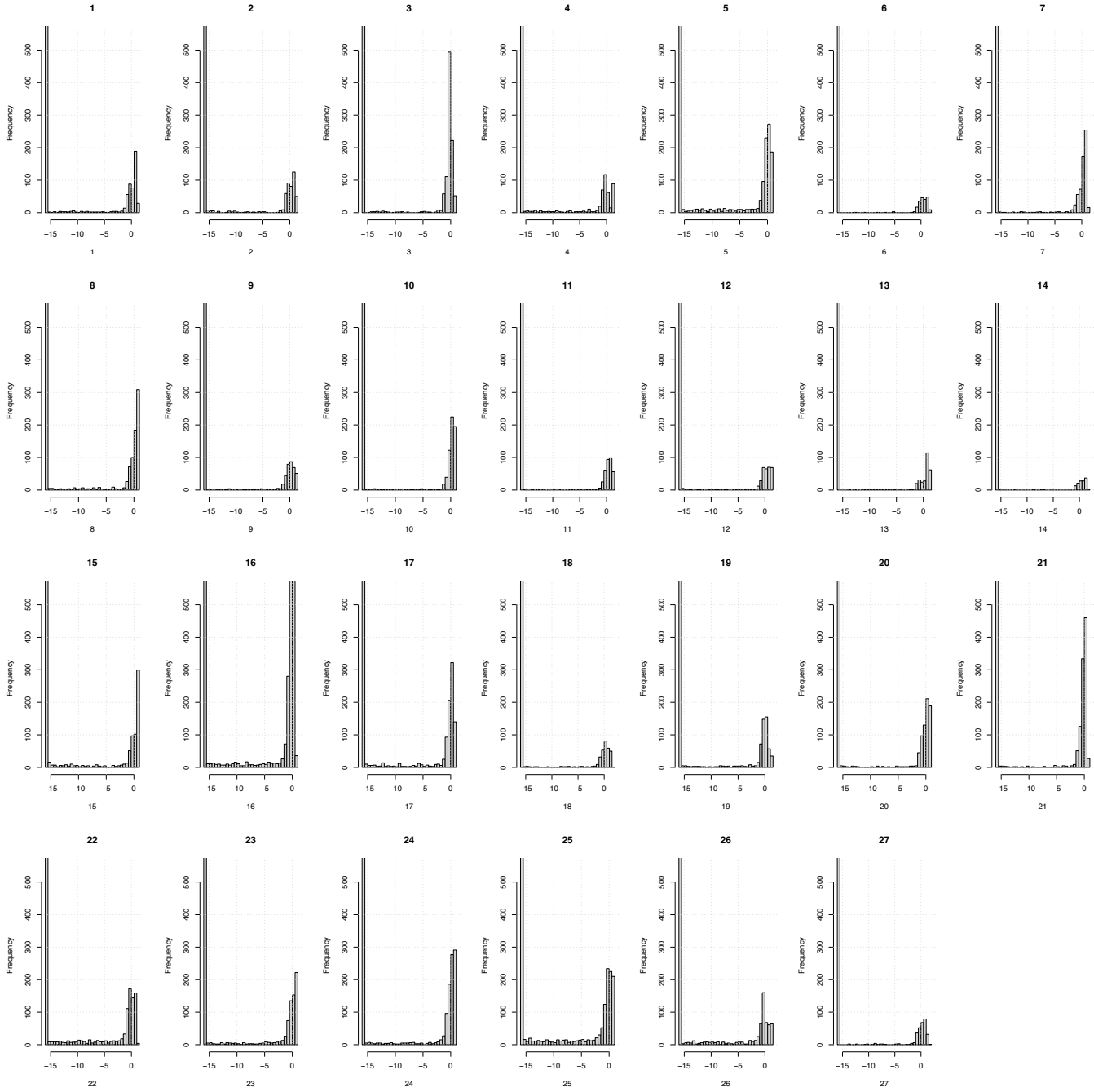

Figure S9. The frequency distributions of the rate of allocation of the 3,164 samples to each of 27 ranks from NMF based on their presence-absence rates in 7,126 plasmid-associated genes. The x-axes show the log10-scaled probability of allocation to each rank. Many of the *Enterobacterales* were allocated to the same rank (6).

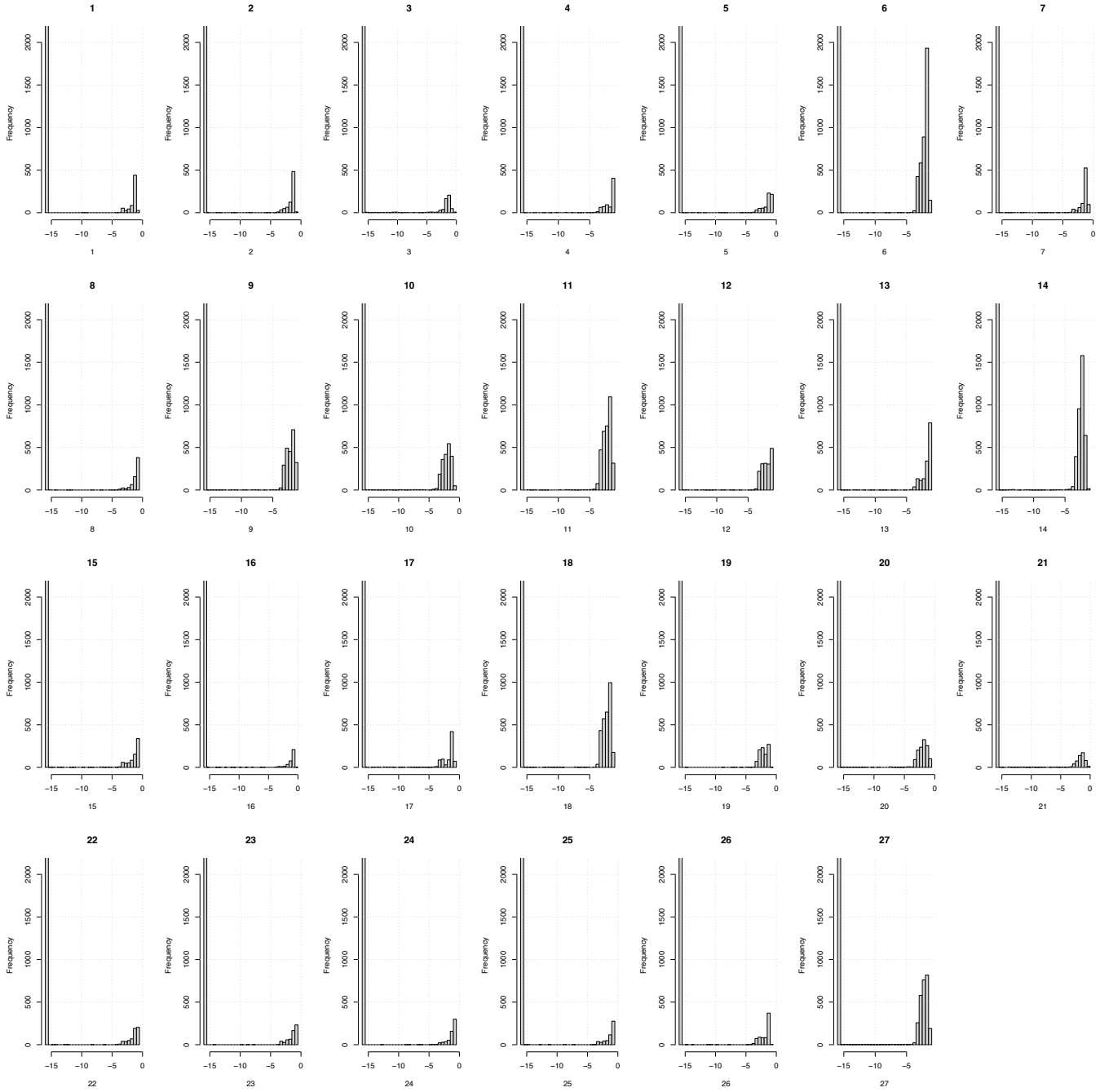

Figure S10. The frequency distributions of the rate of allocation of the 7,126 plasmid-associated genes to each of the 27 ranks from NMF based on presence or absence in 3,164 samples. The x-axes show the log10-scaled probability of allocation to each rank. Many *Enterobacterales*-related genes were allocated to the rank 6.

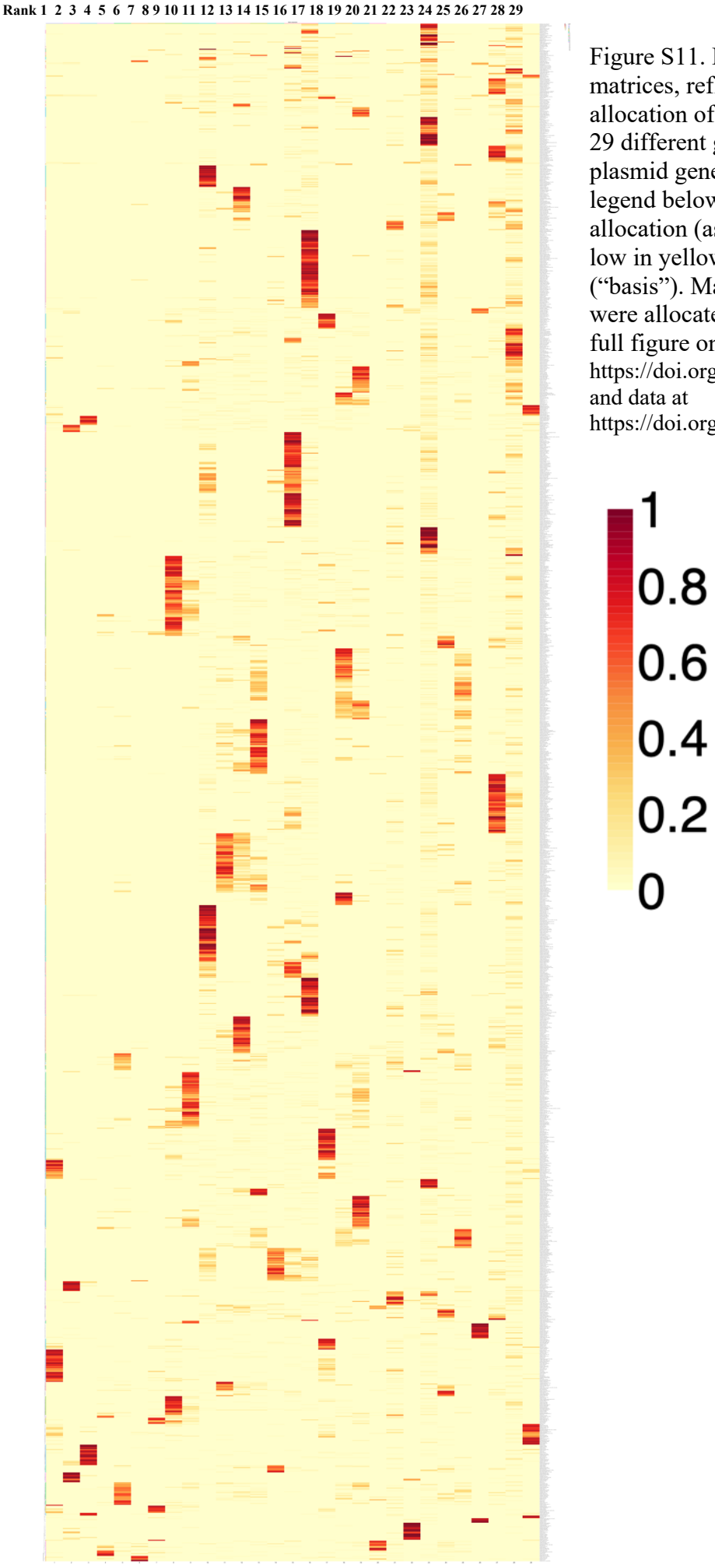

Figure S11. Heatmap of the basis coefficient matrices, reflecting the probability of allocation of the 1,713 samples (y-axis) to 29 different groups (x-axis) based on their plasmid gene PPI rates using NMF. The legend below shows the probability of allocation (as per the legend: high in red, low in yellow) to the different groups ("basis"). Many of the *Enterobacterales* were allocated to the same rank (20). See full figure on FigShare at https://doi.org/10.6084/m9.figshare.19611276 and data at https://doi.org/10.6084/m9.figshare.19611279.

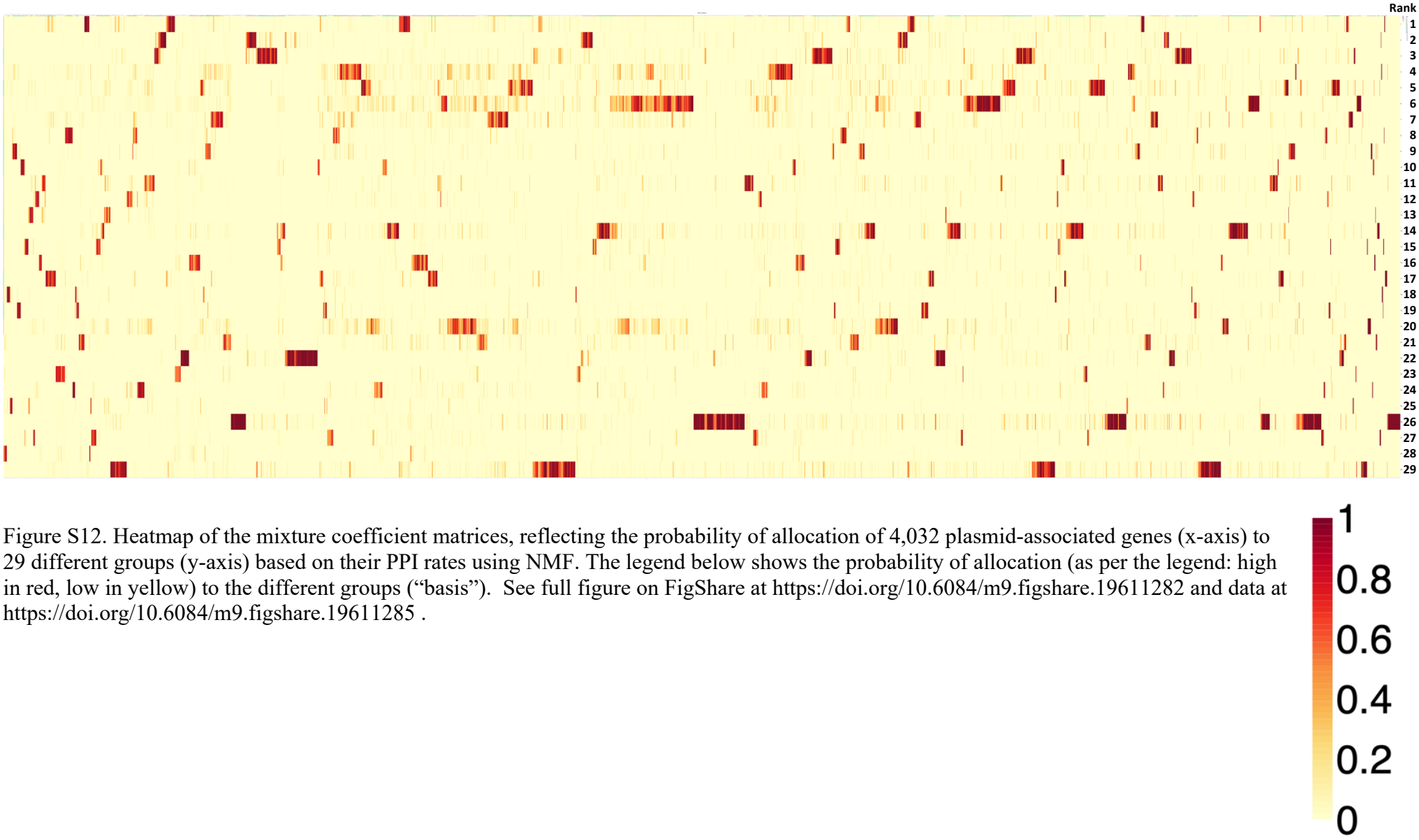

Figure S12. Heatmap of the mixture coefficient matrices, reflecting the probability of allocation of 4,032 plasmid-associated genes (x-axis) to 29 different groups (y-axis) based on their PPI rates using NMF. The legend below shows the probability of allocation (as per the legend: high in red, low in yellow) to the different groups ("basis"). See full figure on FigShare at https://doi.org/10.6084/m9.figshare.19611282 and data at https://doi.org/10.6084/m9.figshare.19611285 .

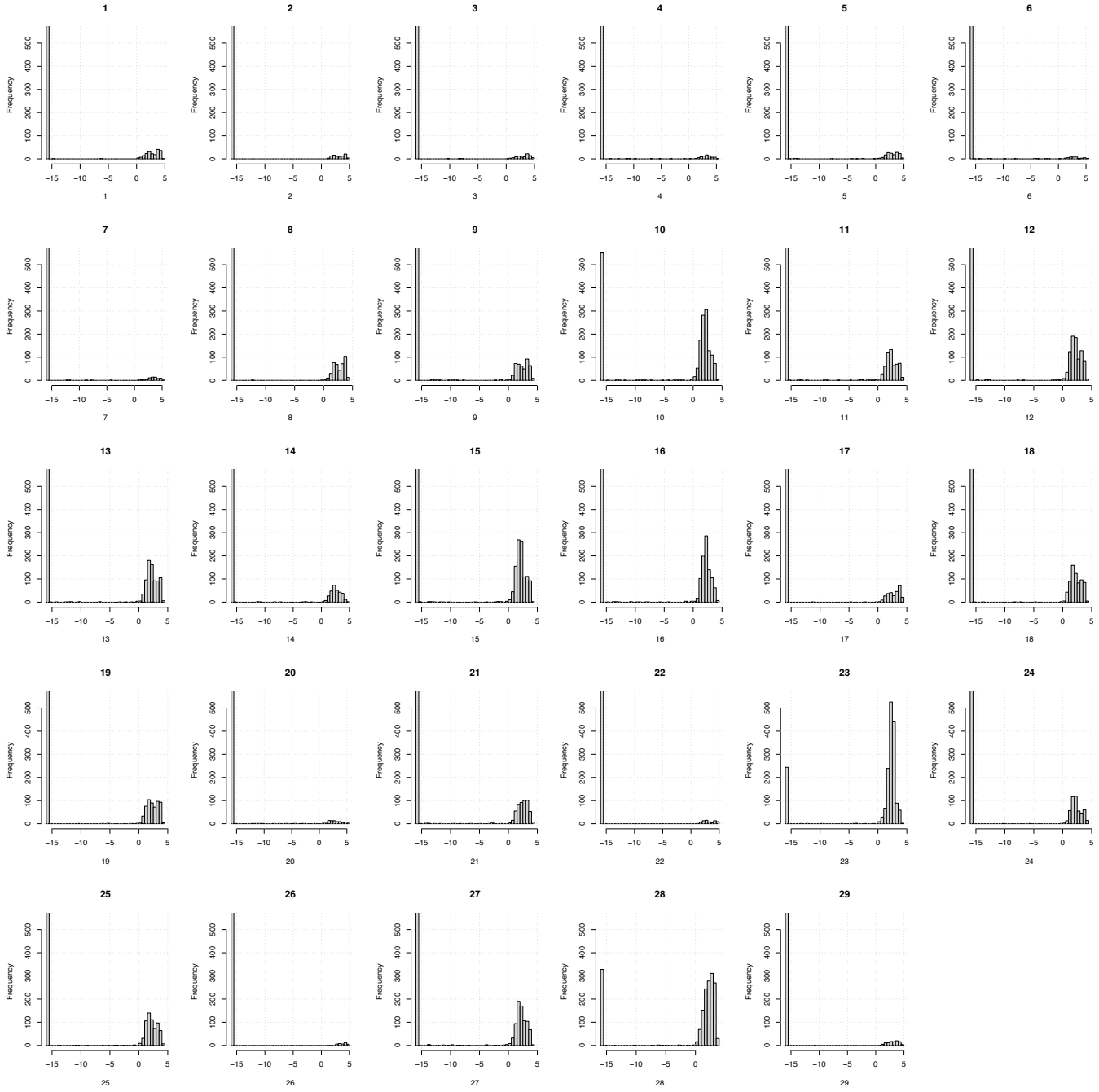

Figure S13. The frequency distributions of the rate of allocation of the 3,164 samples to each of 29 ranks from NMF based on their presence-absence rates in 1,713 plasmid-associated genes. The x-axes show the log10-scaled probability of allocation to each rank.

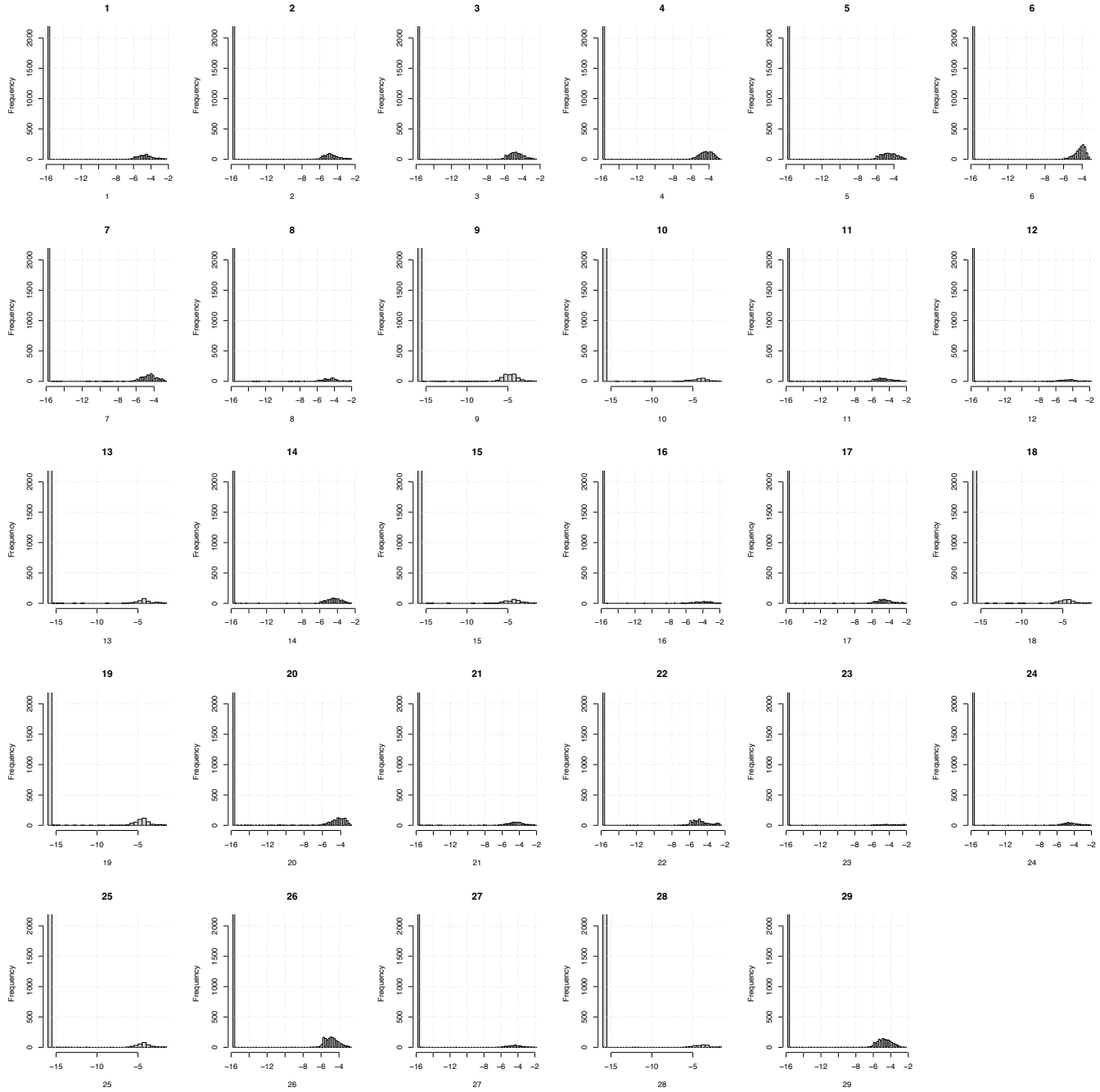

Figure S14. The frequency distributions of the rate of allocation of the 4,032 plasmid-associated genes to each of the 29 ranks from NMF based on presence or absence in 1,713 samples.

Sample	PMNLE	aPMNLE	%difference
Bifidobacterium thermophilum	0.5105	0.5112	-0.15%
Bifidobacterium cuniculi	0.4935	0.4934	0.00%
Acinetobacter baumannii	0.5213	0.5218	-0.10%
Bifidobacterium choerinum	0.5295	0.5296	-0.02%
Bifidobacterium coryneforme	0.5281	0.5282	-0.02%
Bifidobacterium magnum	0.5379	0.5420	-0.77%
Bifidobacterium minimum	0.4948	0.4948	-0.01%
Enterobacter aerogenes	0.5190	0.5186	0.07%
Klebsiella oxytoca	0.5422	0.5417	0.10%
Klebsiella pneumoniae	0.5435	0.5412	0.42%
E. coli 536	0.5639	0.5626	0.22%
E. coli ATCC8739	0.5461	0.5455	0.10%
E. coli BL21DE3	0.5706	0.5700	0.11%
E. coli CFT073	0.5554	0.5545	0.15%
E. coli K12	0.5907	0.5901	0.10%
E. coli O157H7	0.5659	0.5654	0.09%
Mean	0.5383	0.5382	0.02%
Standard deviation	0.0272	0.0269	0.25%

Table S2. The complete PMNLE and approximate PMNLE (aPMNLE) values for 16 samples' PPI data from StringDB across all proteins. The scaled mean difference (%difference) between the complete and approximate PMNLE values was small (0.02% with a standard deviation of 0.25%).

Table S3. Key features of the bacterial samples. Plasmid_PPI indicates the number of plasmid-related PPIs; Chrom_PPI indicates the number of chromosomal PPIs; Fraction indicates the fraction of PPIs that were plasmid-related; inputProteins indicates the number of proteins per sample; interactions indicates the total number of PPIs; Plasmid_genes indicates the number of plasmid-related genes per sample; aPMNLE indicates the chromosomal proteins' aPMNLE; cTri indicates the number of chromosomal protein trios; loops indicates the number of indirect connections for chromosomal proteins; ccs indicates the number of connected components for chromosomal proteins; aPMNLE_all indicates the aPMNLE for all proteins; cTri_all indicates the number of protein trios; loops_all indicates the number of indirect connections; ccs_all indicates the number of connected components. Note some PMNLE values were not computational tractable and so the cells are empty. See FigShare doi: https://doi.org/10.6084/m9.figshare.19525708.

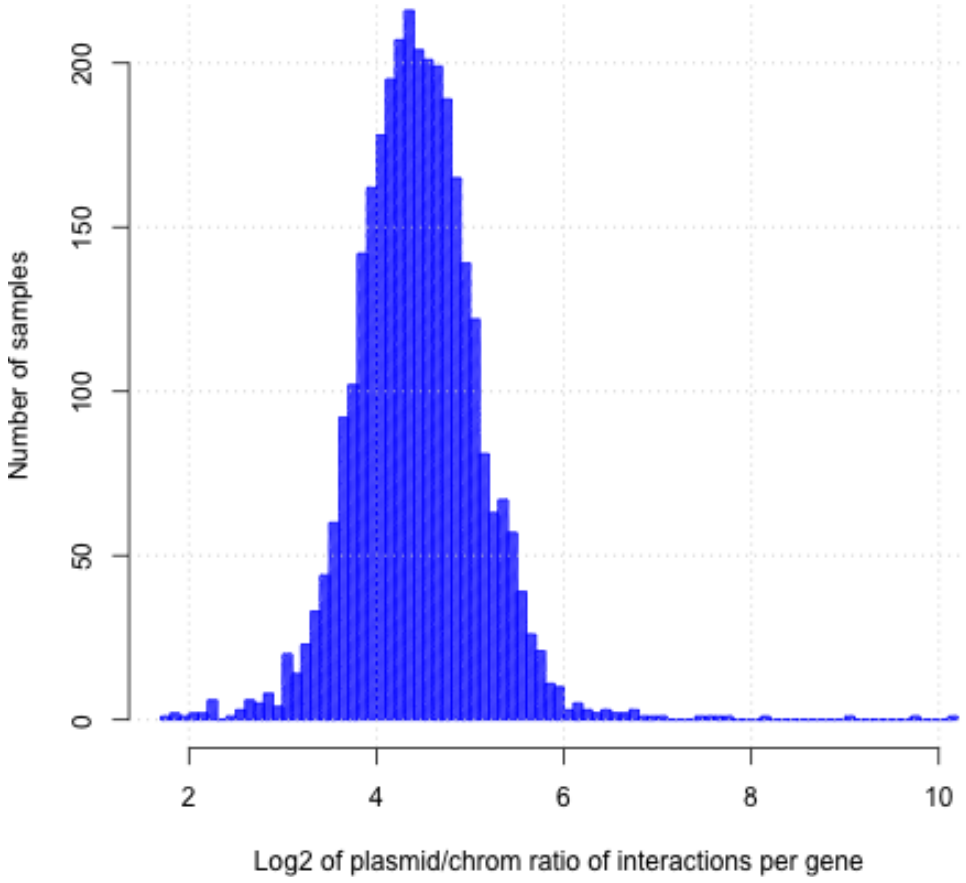

Figure S15. The variation in the log2-scaled average number of PPIs per gene for plasmid-associated versus chromosomal genes across the 3,164 samples with a mean PPIs per gene > 0.5. The median (±SD) ratio was 4.5±0.9, highlighting much higher PPIs among plasmid-encoded genes.

Table S4. The data for the number of chromosomal and plasmid-related PPIs per gene for 4,394 samples. See FigShare doi: https://doi.org/10.6084/m9.figshare.19525681.

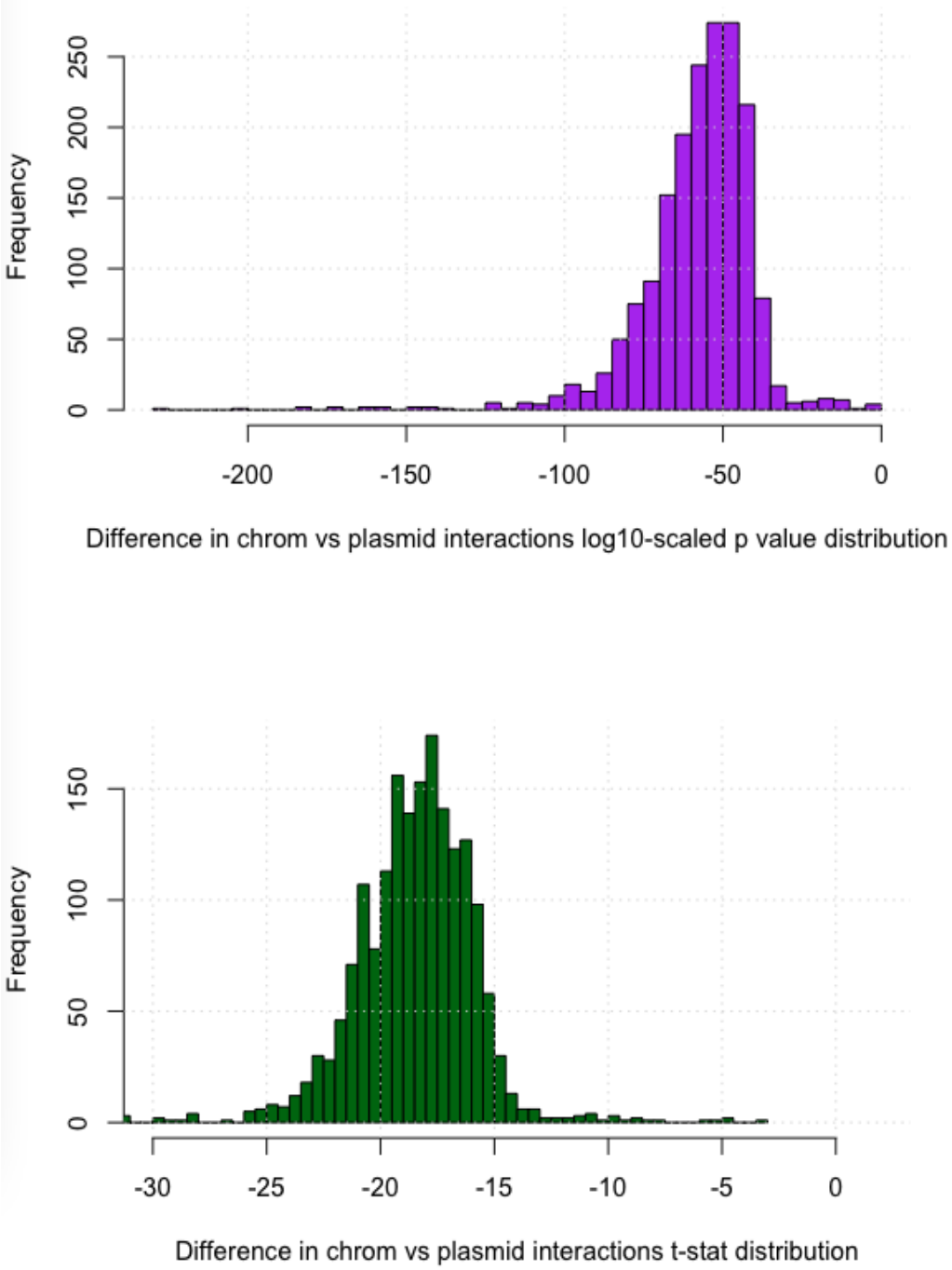

Figure S16. More interactions per protein for plasmid-associated proteins compared to chromosomal proteins for each sample based on t-test log10-scaled p value distribution (top, purple) and the t-statistic distribution (bottom, green).

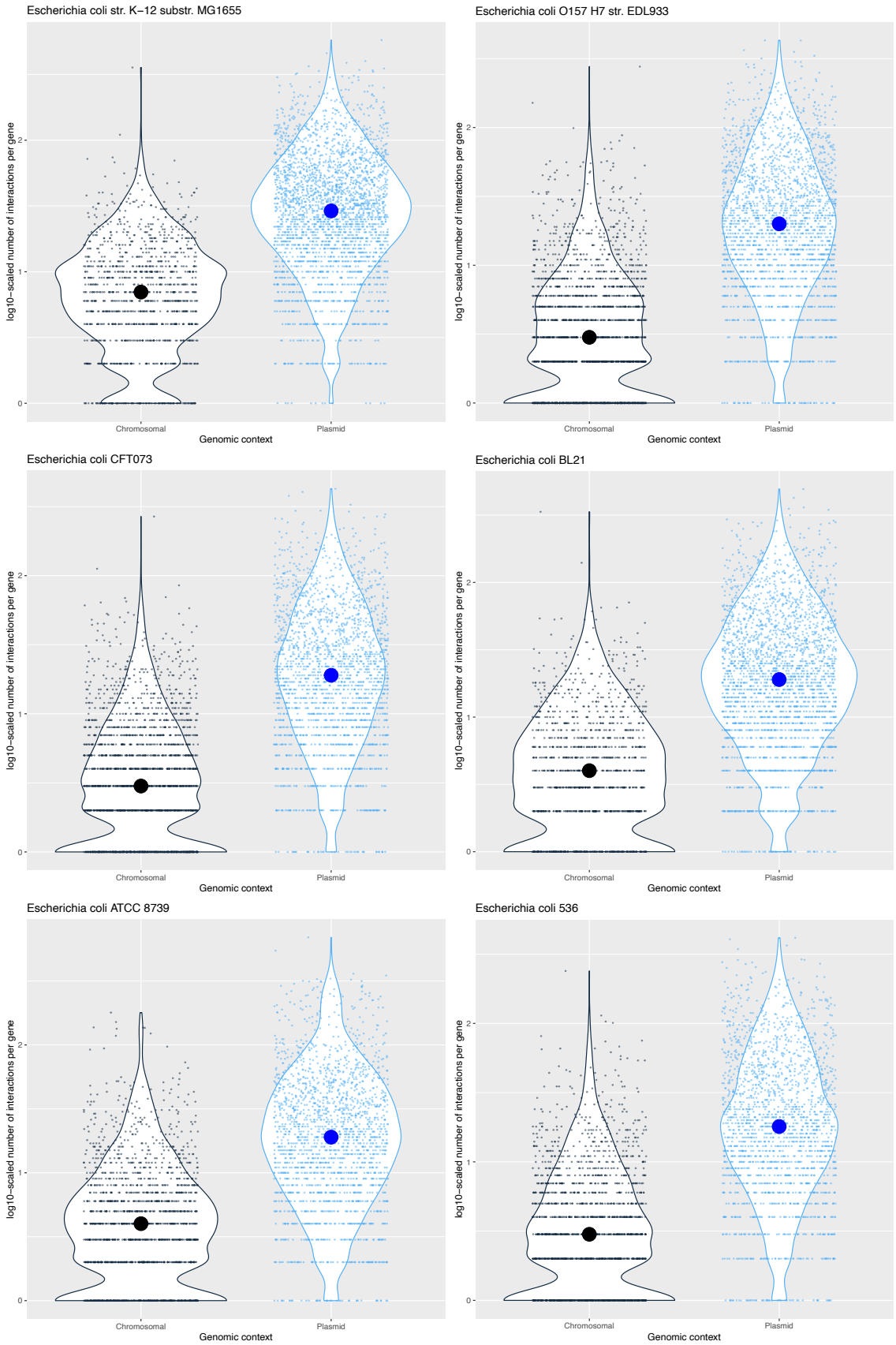

Figure S17. The log10-scaled numbers of chromosomal (black) and plasmid (blue) PPIs per gene for six *E. coli*: K-12 substrain MG1665 (top left), O157:H7 substrain EDL933 (top right), CFT073 (middle left), BL21 (middle right), ATCC8793 (bottom left), and 536 (bottom right). The data for these and the full set of chromosomal and plasmid interactions per gene for 4,377 samples are available as CSV files and 3,008 PDF image files (images were generated if there was sufficient PPI data to plot) on FigShare at doi: https://doi.org/10.6084/m9.figshare.19576408.The points show the observed values on top of the distribution shapes and the medians shown by large circles.

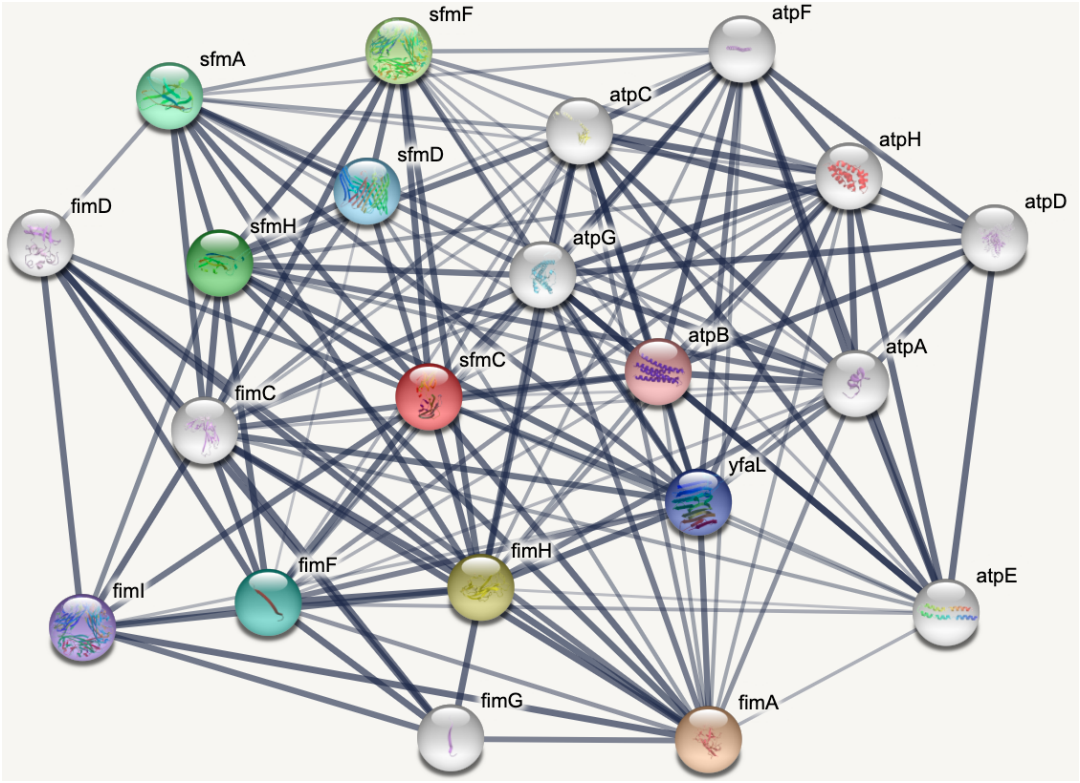

Figure S18. An *E. coli* K12 MG1655 (String ID 511145) PPI network from the StringDB website of 21 proteins in Table S4 centred on SfmC (a putative periplasmic pilus chaperone that is part of the *sfmACDHF* fimbrial operon) with 147 PPIs.

Gene	Context	Number of PPIs
atpA	Plasmid	82
atpB	Plasmid	32
atpC	Plasmid	43
atpD	Plasmid	65
atpE	Plasmid	32
atpF	Plasmid	43
atpG	Plasmid	60
atpH	Plasmid	53
fimA	Plasmid	12
fimC	Plasmid	8
fimD	Plasmid	6
fimF	Plasmid	4
fimG	Plasmid	2
fimI	Plasmid	10
sfmA	Chromosome	20
sfmC	Chromosome	12
sfmD	Chromosome	11
sfmF	Chromosome	54
sfmH	Chromosome	24
yfaL	Chromosome	25

Table S5. The 21 genes encoding proteins in the PPI network in Figure S18 along with their genomic context (plasmid or chromosome) and the number of PPIs per protein across *E. coli* K12 MG1655's entire set of proteins. In this example, the plasmid-related PPIs have 32.3±26.1 (mean±SD) PPIs compared to 24.3±15.7 for the chromosomal ones.

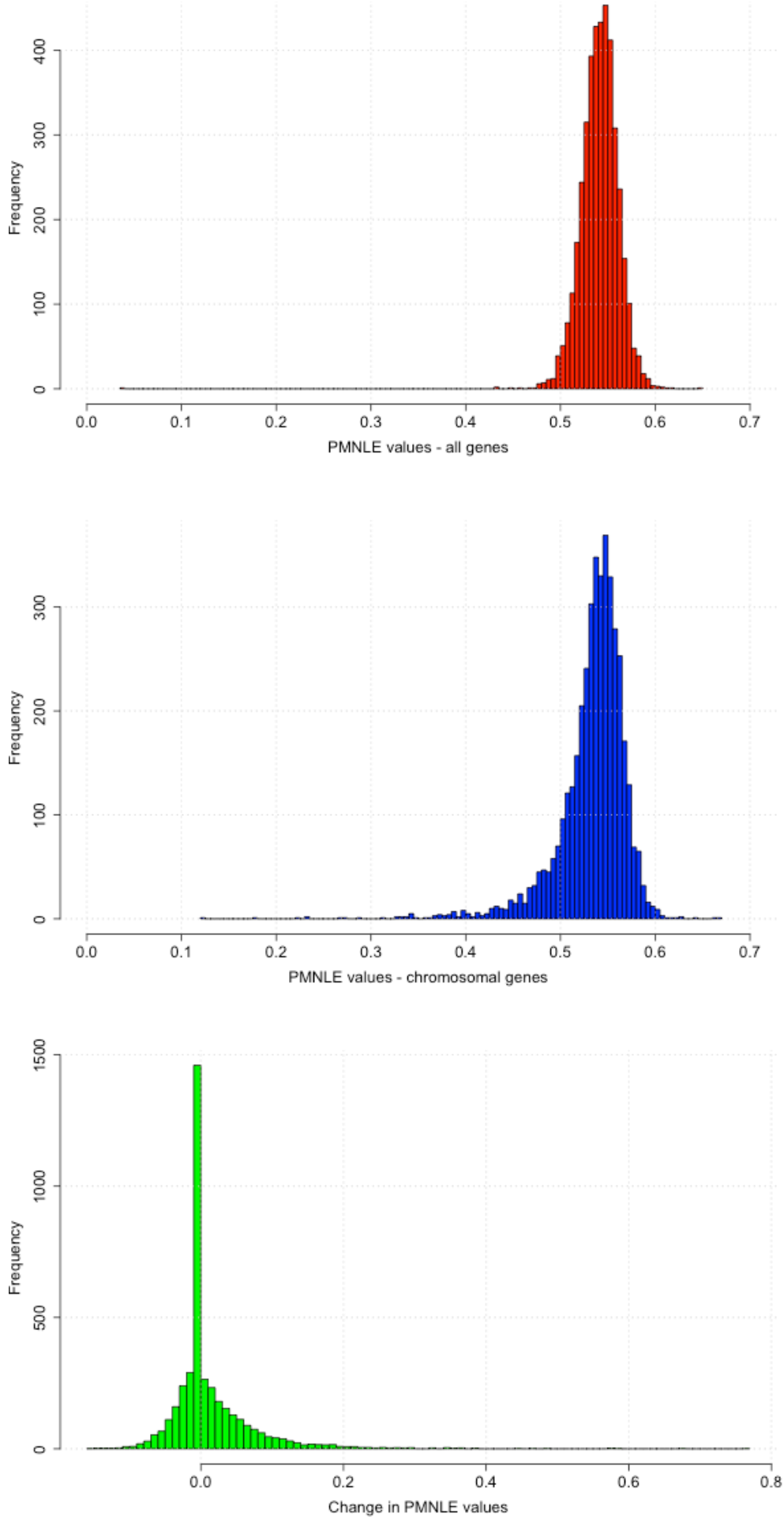

Figure S19. The distribution of aPMNLE values (x-axis) for all proteins (top, red), chromosomal ones (middle, blue) and the scaled difference between all and chromosomal proteins (bottom, green) across the bacterial samples. Values for 4,187 samples with valid aPMNLE values for all proteins and the chromosomal proteins. Note that the y-axes scales differ.

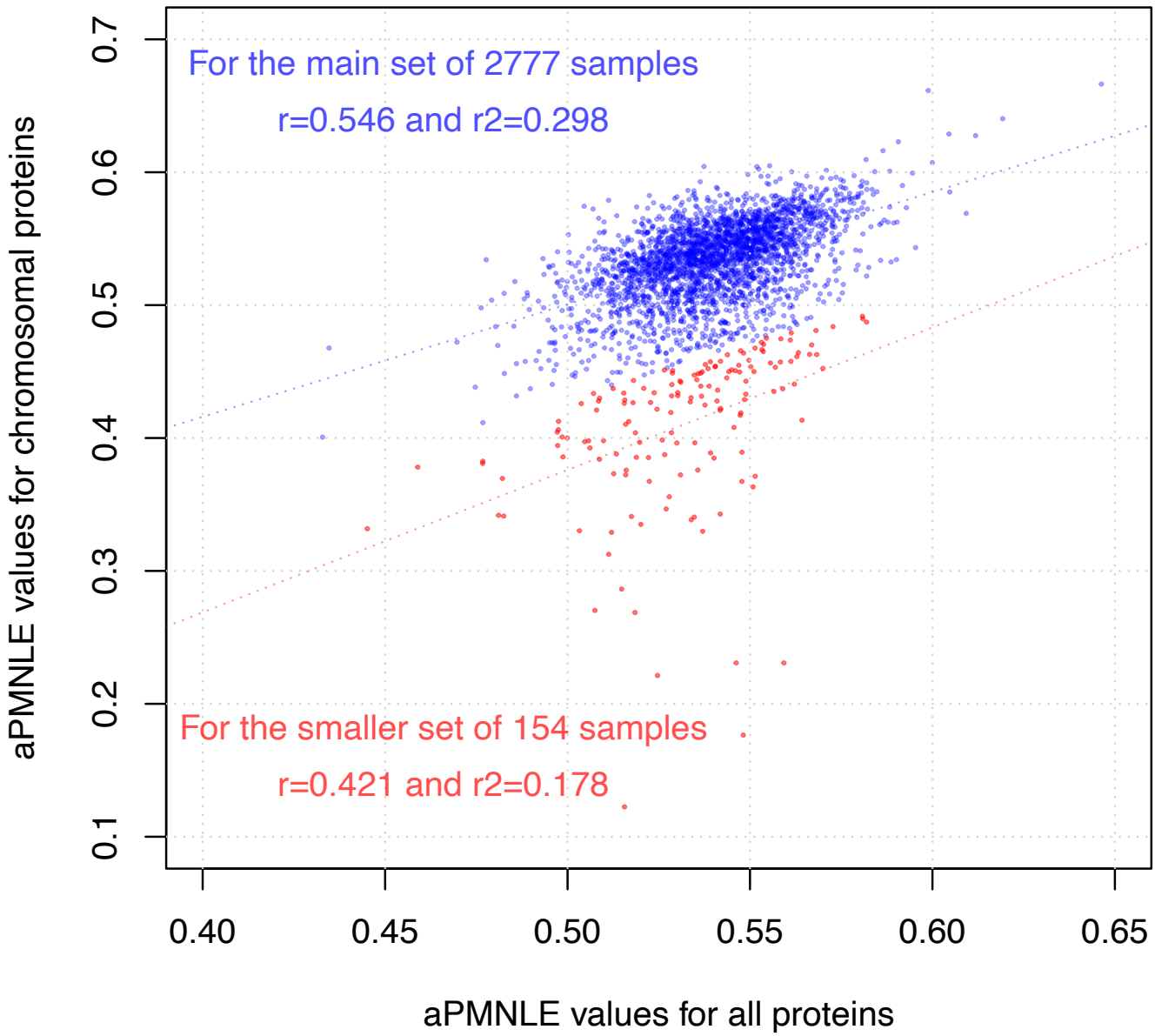

Figure S20. The association between the aPMNLE values for all proteins (x-axis) compared to those for chromosomal proteins (y-axis) for the group with no significant difference in these aPMNLE values (n=2,777 samples, blue) and the group where there was a much lower chromosomal aPMNLE value (n=154, red). The correlation between these aPMNLE metrics was higher for the main group of 2,777 samples (r=0.55) compared to the group with the change in aPMNLE value (r=0.42). The group of 154 had proportionally lower chromosomal aPMNLE values when compared to the other 2,777 samples.

Table S6. Bacterial samples (154, 5% of total) with a much higher aPMNLE for all proteins than the chromosomal ones alone. This indicated that the plasmid-related proteins contributed more indirect PPIs. See FigShare doi: https://doi.org/10.6084/m9.figshare.19525681.

Sample	Plasmid_PPI	Chrom_PPI	Fraction	Proteins	PPIs	Plasmid_genes	chrom_PMNLE	PMNLE	diffPMNLE
Streptococcus agalactiae LMG 14747	97,558	1,828	0.982	2,186	99,385	368	0.597	0.521	-0.145
Candidatus Amoebophilus asiaticus 5a2	15,291	901	0.944	1,335	16,191	213	0.586	0.509	-0.152
Arcanobacterium sp S3PF19	17,337	210	0.988	1,003	17,546	238	0.611	0.532	-0.148

Table S7. Bacterial samples (3, 0.1% of total) with a much lower aPMNLE for all proteins than the chromosomal ones alone. This indicated that the plasmid-related proteins lowered the rates of indirect PPIs by switching them to direct PPIs, suggesting they were mixed in the PPI network with the chromosomal proteins.

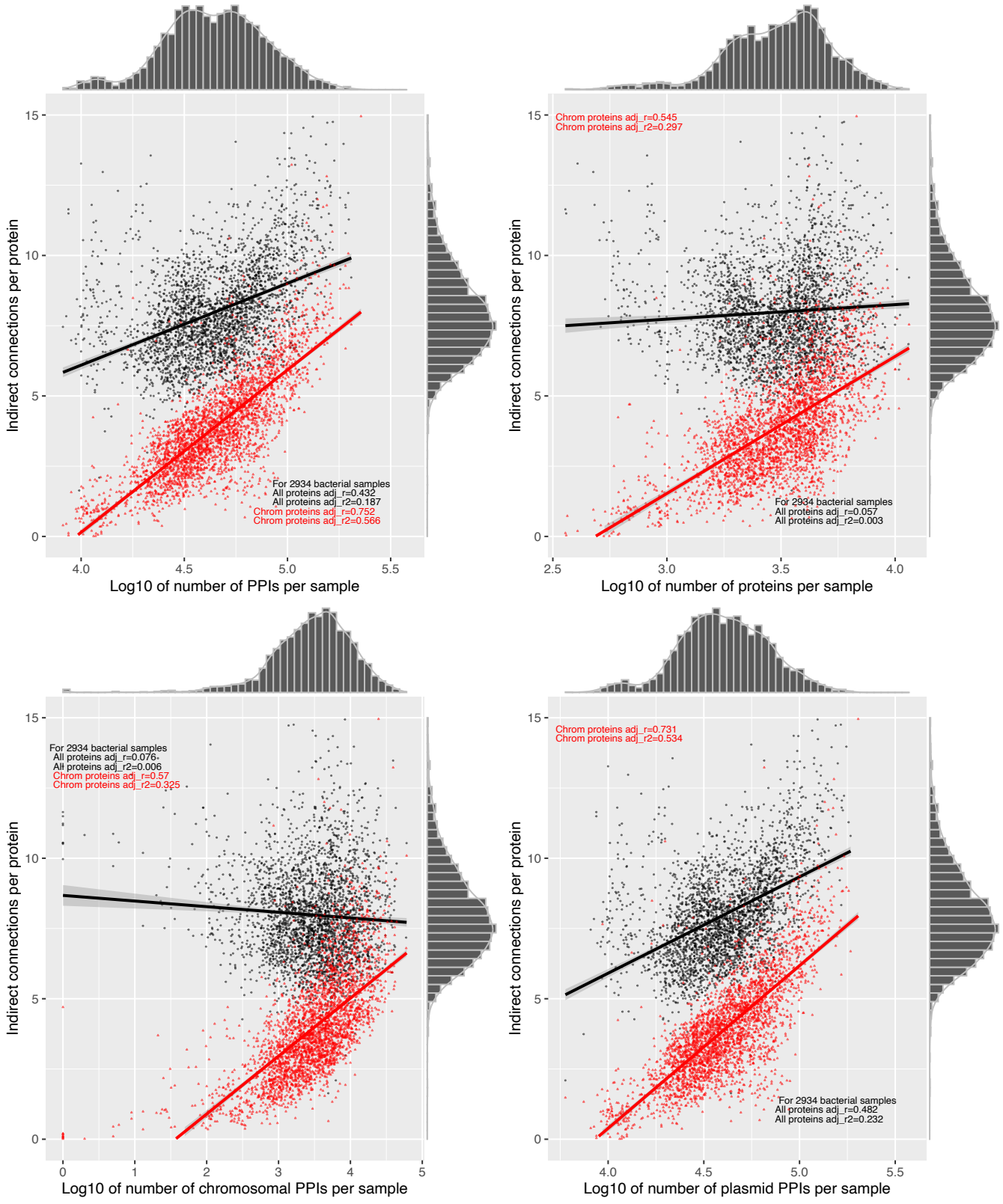

Figure S21. Varying levels of a positive correlation between (A) the log10-scaled number of PPIs, (B) log10-scaled number of plasmid-related PPIs, (C) log10-scaled number of chromosome-restricted PPIs, (D) log10-scaled number of proteins per sample, and number of indirect connections per protein (y-axis) for all proteins (black) and chromosomal proteins (red). The data plotted is for 2,934 samples with at least one plasmid-related PPI and complete connected component data. The line of best fit (black for all proteins, red for chromosomal ones) shows a linear correlation of each variable with the number of indirect connections per protein. The histograms indicate the marginal densities per axis.

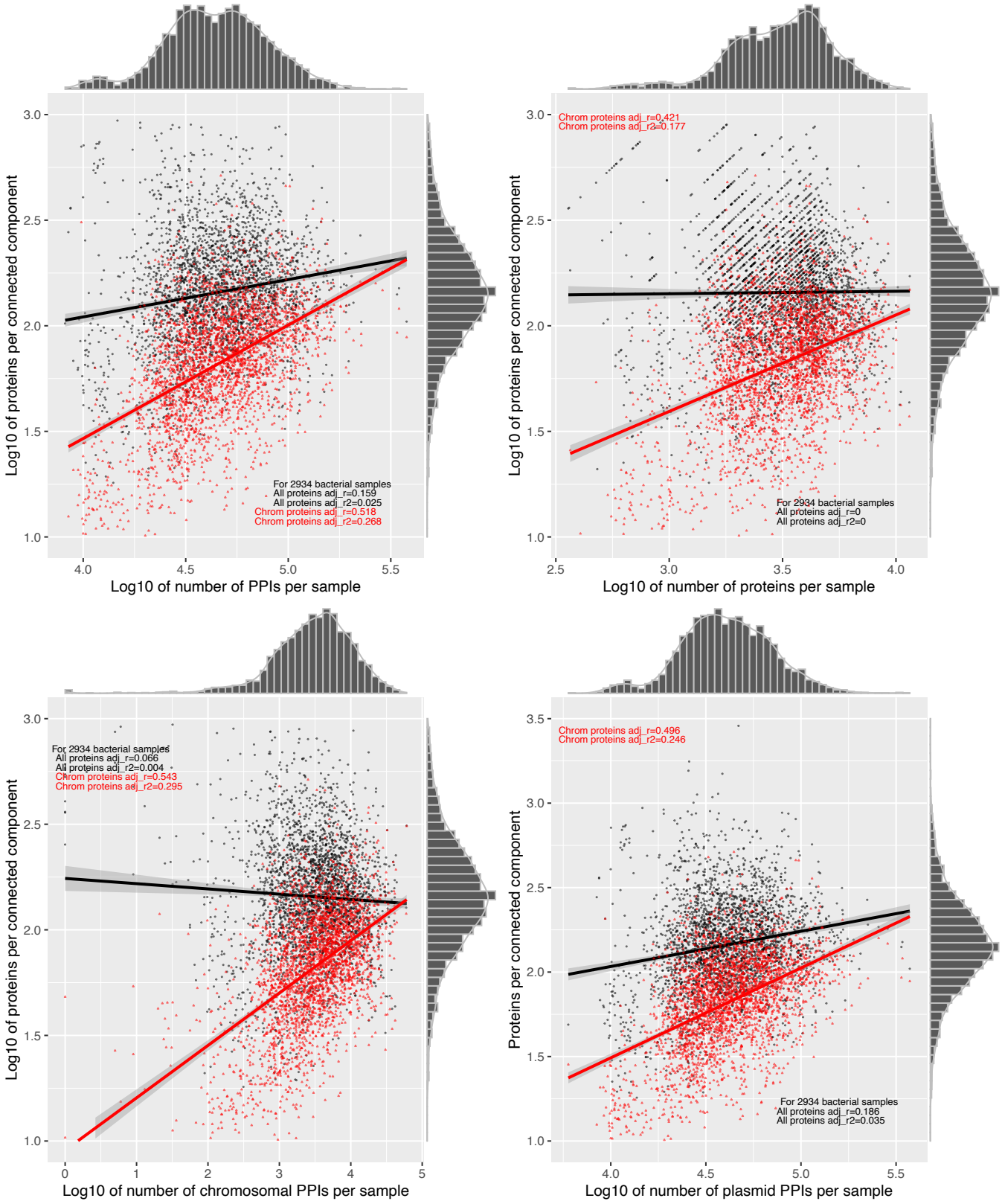

Figure S22. Varying levels of a positive correlation between (A) the log10-scaled number of PPIs, (B) log10-scaled number of plasmid-related PPIs, (C) log10-scaled number of chromosome-restricted PPIs, (D) log10-scaled number of proteins per sample, and number of proteins per connected components (y-axis) for all proteins (black) and chromosomal proteins (red). The data plotted is for 2,934 samples with at least one plasmid-related PPI and complete connected component data. The line of best fit (black for all proteins, red for chromosomal ones) shows a linear correlation of each variable with the number of indirect connections per protein. The histograms indicate the marginal densities per axis.