

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

Page	Item	Description
1	Fig S1	The number of long gene names (x-axis) compared to the number of short gene names (y-axis) per sample.
1	Table S1	Each of the plasmid-related genes and their associated plasmid link(s) across 15,531 plasmid accessions.
2	Figure S2	Heatmap visualising plasmid-related genes for 4,429 bacterial samples (y-axis) across 7,126 unique plasmid-derived genes (x-axis).
3	Figure S3	Dendrogram showing the similarity across 50 samples from the <i>Enterobacterales</i> order.
4	Figure S4	PCA of the whole dataset showing the variation of: (A) the 4,429 samples across the 7,126 plasmid-associated genes found in these samples, and (B) these genes across these samples.
5	Figure S5	Heatmap visualising plasmid interactomes for 3,146 bacterial samples (y-axis) across 7,126 unique plasmid-derived genes (x-axis).
6	Figure S6	Visualisation of the 1 st (x-axis, 7.6% of variation) and 2 nd (y-axis, 2.3% of variation) dimensions from LSA.
7	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 profiles using NMF.
8	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 prevalence in 3,164 samples using NMF.
9	Figure S9	The frequency distributions of the rate of allocation of the 3,164 samples to each of the 27 ranks from NMF.
10	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.
11	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 PPI rates using NMF.
12	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 in 1,713 samples using NMF.
13	Figure S13	The frequency distributions of the rate of allocation of the 1,713 samples to each of the 29 ranks from NMF of the plasmid PPI rates.
14	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 of the plasmid PPI rates.
15	Table S2	The complete PMNLE and approximate PMNLE (aPMNLE) values for 16 samples' PPI data from StringDB across all proteins.
16	Table S3	Key features of the bacterial samples.
16	Table S4	The data for the number of chromosomal and plasmid-related PPIs per gene.
16	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.
17	Figure S16	More interactions per protein for plasmid-associated proteins compared to chromosomal proteins for each sample.
18	Figure S17	The numbers of chromosomal and plasmid interactions per gene for the six <i>E. coli</i> samples.
19	Figure S18	An <i>E. coli</i> K12 MG1655 (String ID 511145) PPI network from the StringDB website of 21 proteins in Table S5 centred on SfmC.
19	Table S5	The 21 genes encoding proteins in the PPI network in Figure S18 along with their genomic context (plasmid or chromosome).
20	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.
21	Figure S20	The association between the aPMNLE values for all proteins (x-axis) compared to those for chromosomal proteins (y-axis).
22	Table S6	Bacterial samples (155, 5% of total) with a much higher aPMNLE for all proteins than the chromosomal ones alone.
22	Table S7	Bacterial samples (3, 0.1% of total) with a much lower aPMNLE for all proteins than the chromosomal ones alone.
23	Figure S21	Varying levels of a positive correlation with the number of indirect connections per protein.
24	Figure S22	Varying levels of a positive correlation with the number of proteins per connected component.

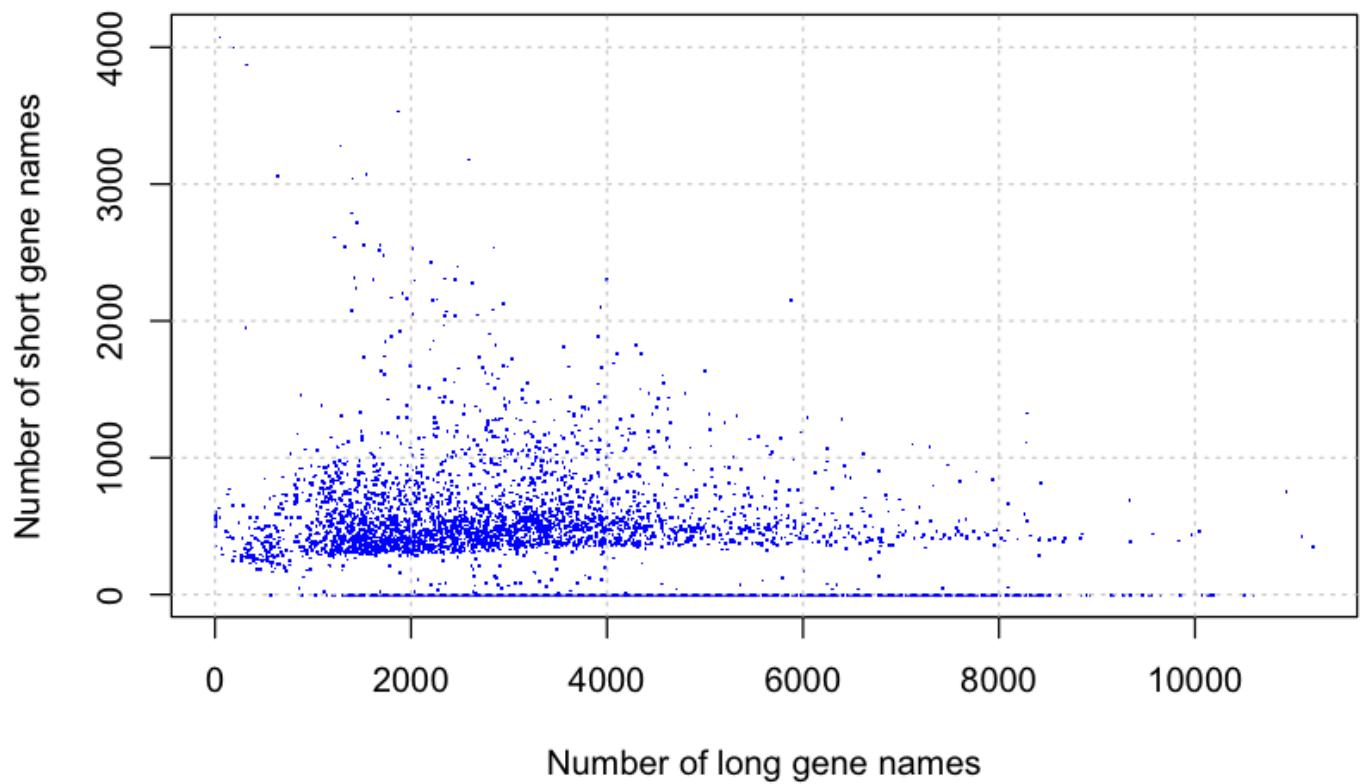


Fig S1. The number of long gene names (x-axis) compared to the number of short gene names (y-axis) per sample. The 4,429 samples had a median of 411 ± 410 short names and a median of $2,931 \pm 1,744$ long names each. Short gene names had four letters or less, and long ones had more than this.

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>

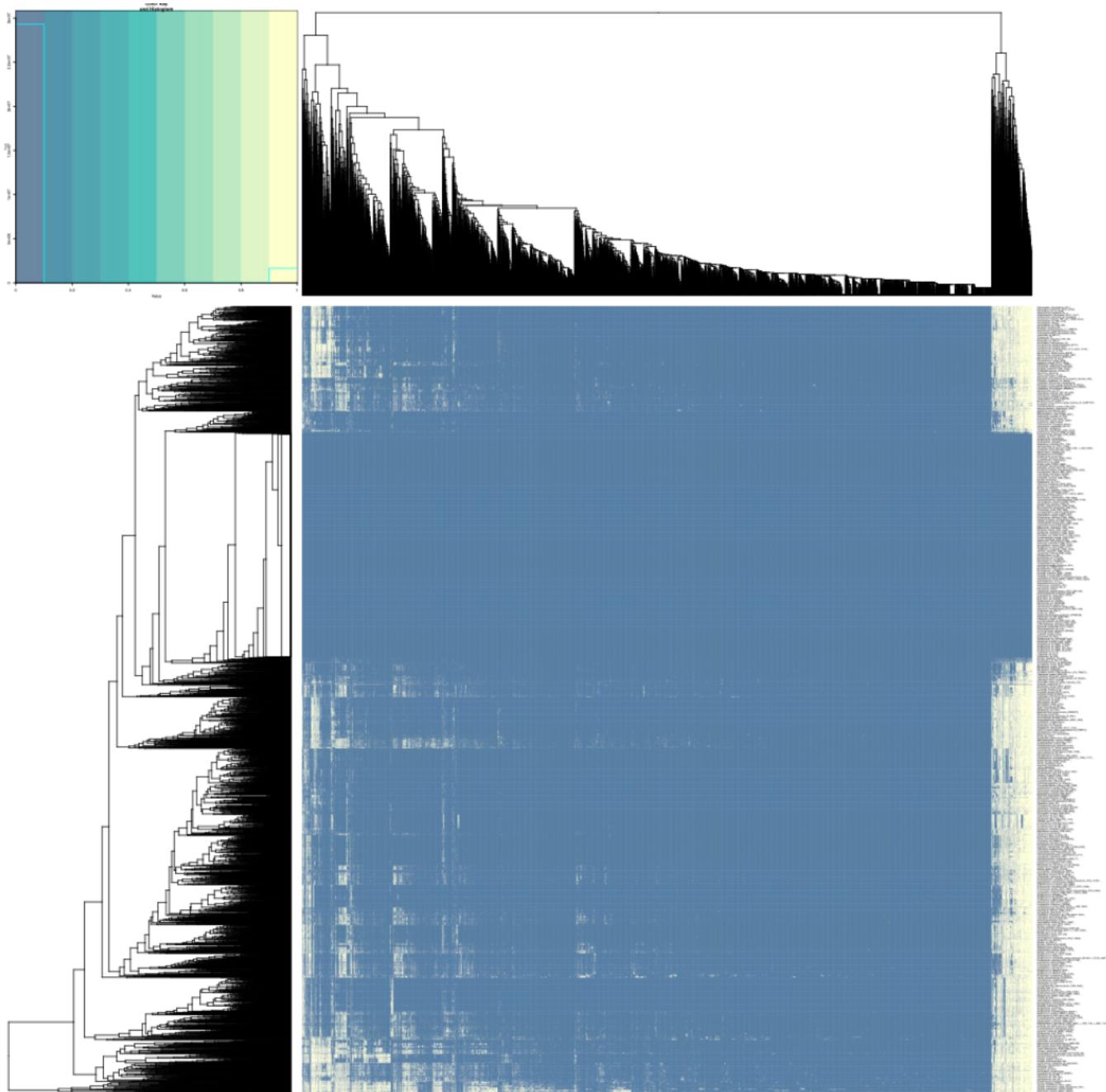


Figure S2. Heatmap visualising plasmid-related genes for 4,429 bacterial samples (y-axis) across 7,126 unique plasmid-derived genes (x-axis). Dendograms indicate the similarity across the samples (left) and genes (top). Both axes were sorted based on similarity. Genes found to possess interactions are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

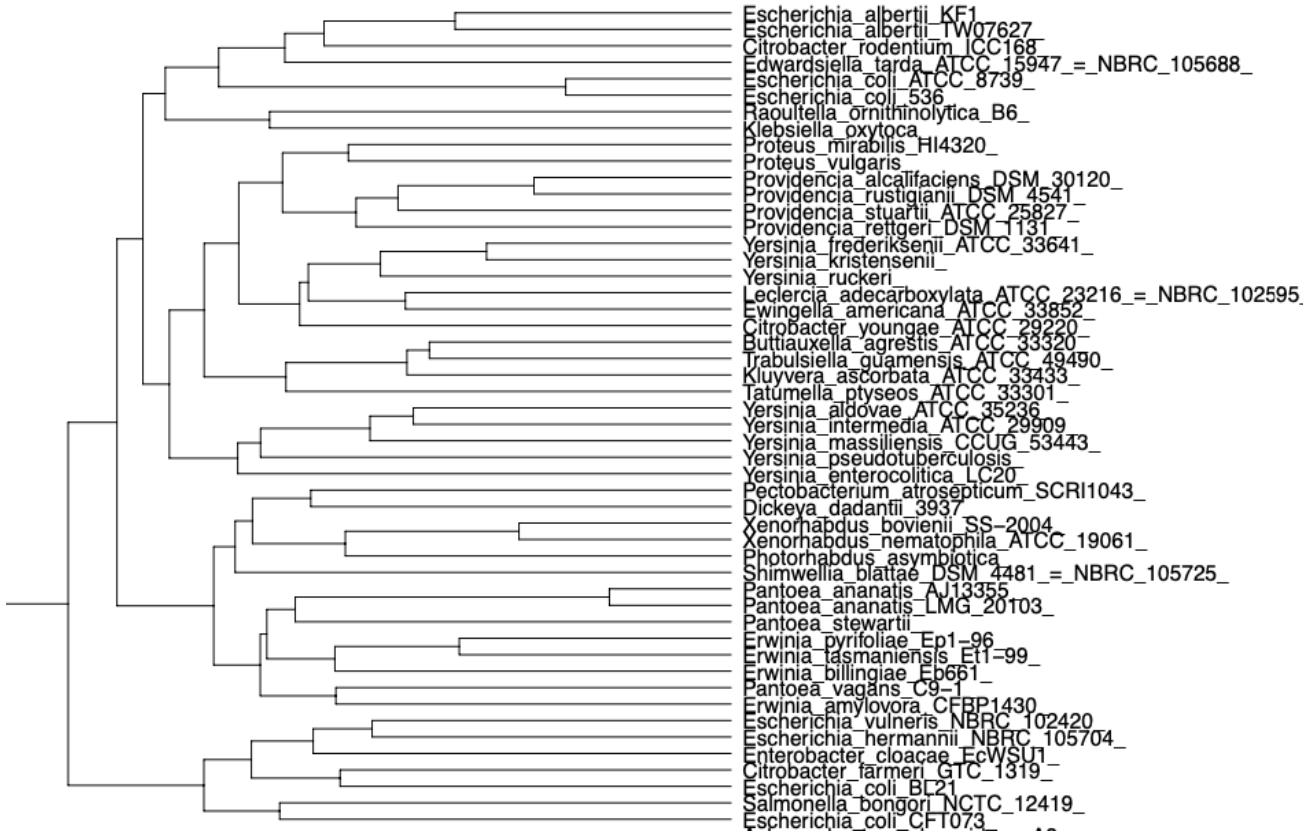


Figure S3. Dendrogram showing the similarity across 50 samples from the *Enterobacterales* order that were the most divergent clade among 3,146 bacteria with had 1+ plasmid-associated genes across 7,126 unique plasmid-derived genes. These corresponded to rank 6 in the NMF results.



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.

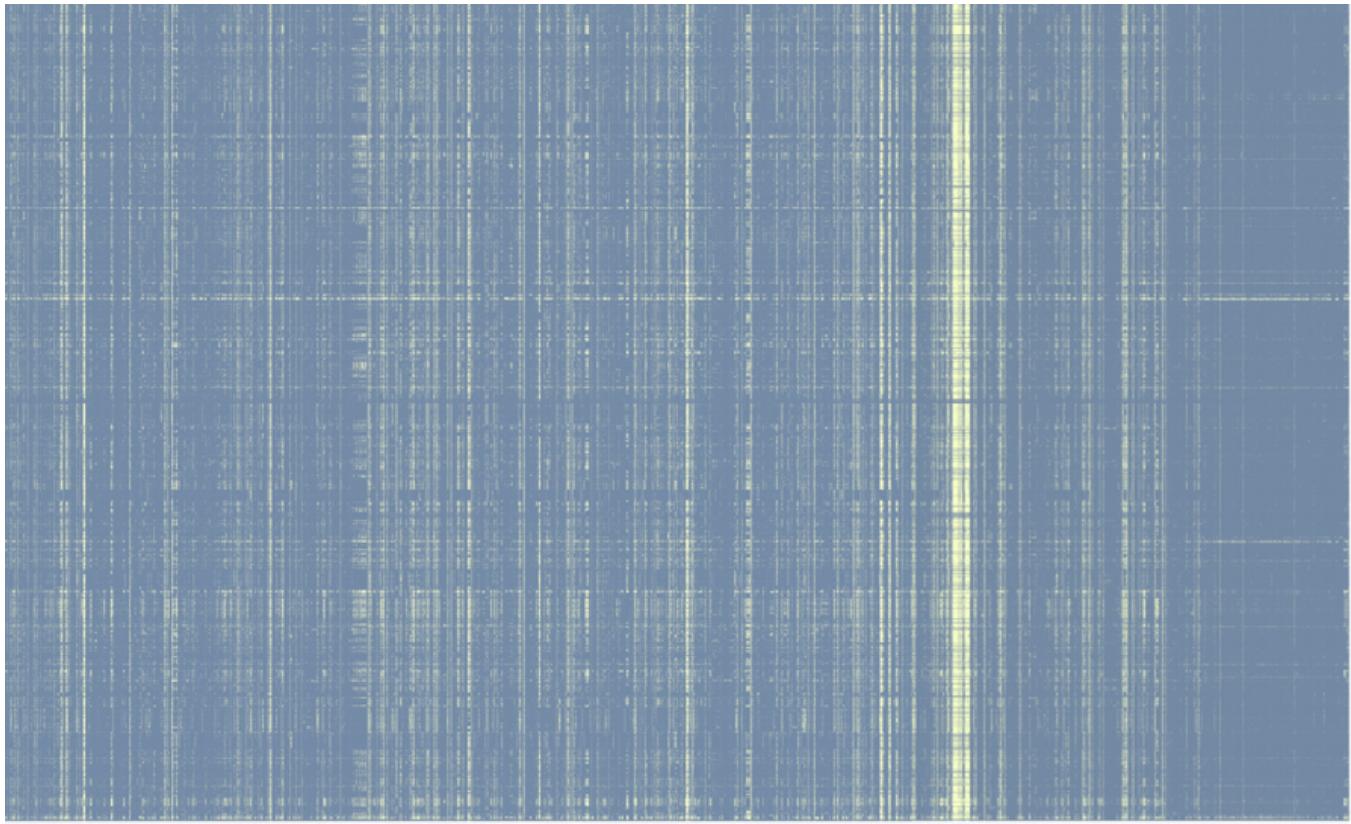


Figure S5. Heatmap visualising plasmid interactomes for 3,146 bacterial samples (y-axis) across 7,126 unique plasmid-derived genes (x-axis). The axes were sorted alphabetically. See full figure on FigShare at <https://doi.org/10.6084/m9.figshare.19525465> and data at <https://doi.org/10.6084/m9.figshare.19525471>.

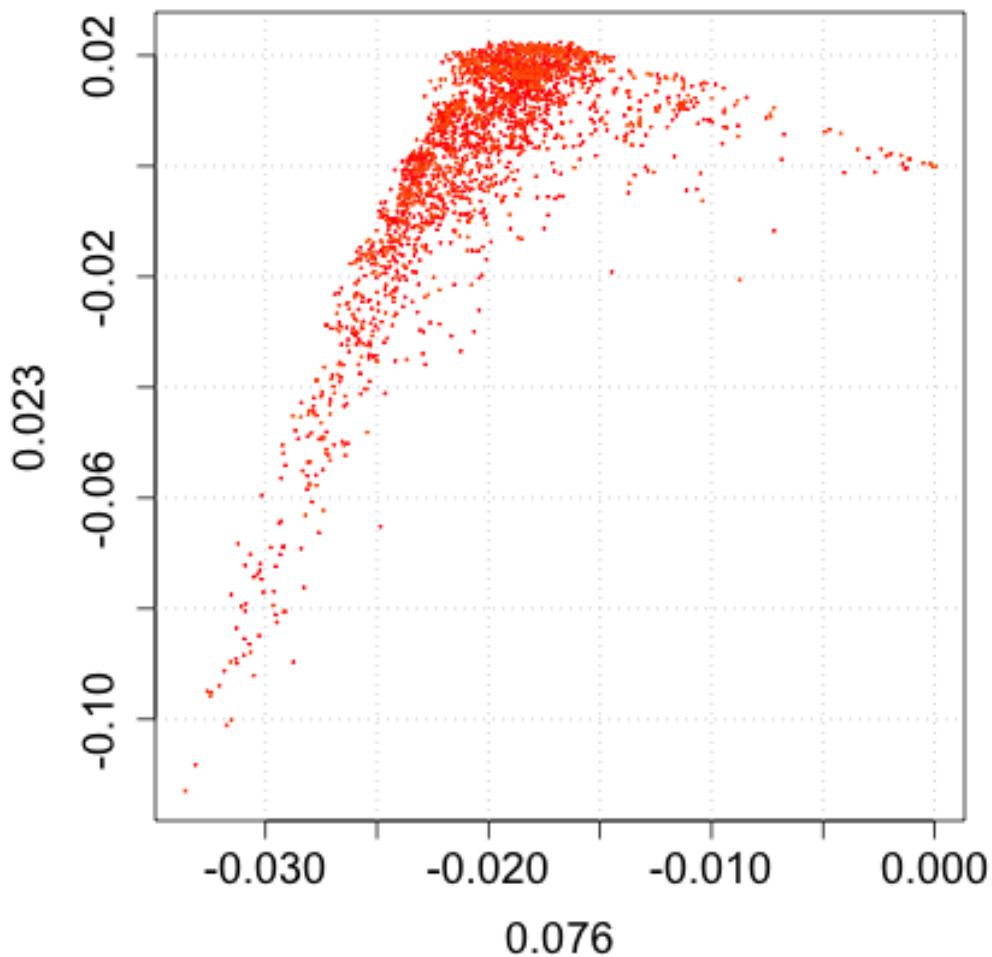


Figure S6. Visualisation of the 1st (x-axis, 7.6% of variation) and 2nd (y-axis, 2.3% of variation) dimensions from LSA of the 3,146 bacterial samples across 360 dimensions representing 7,126 unique plasmid-derived genes.

Rank 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27

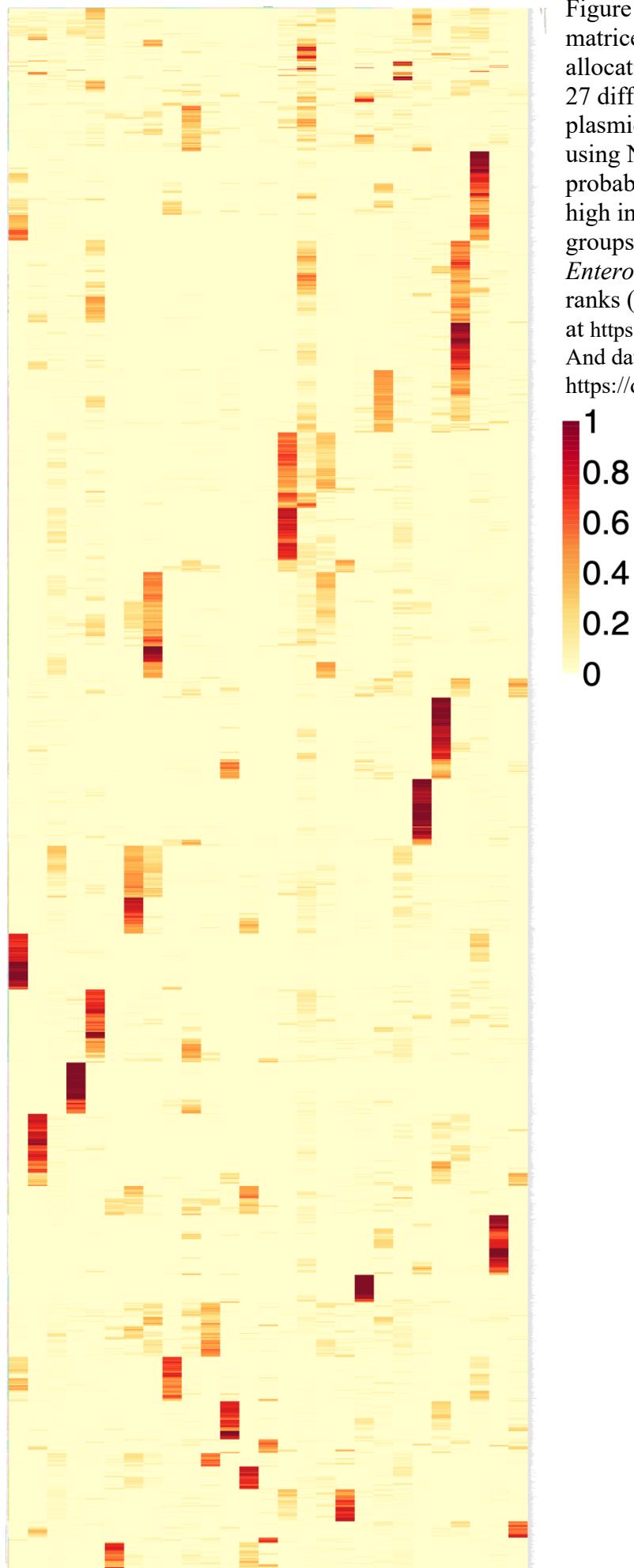


Figure S7. Heatmap of the basis coefficient matrices, reflecting the probability of allocation of the 3,164 samples (y-axis) to 27 different groups (x-axis) based on their plasmid gene presence-absence profiles using NMF. The legend below shows the probability of allocation (as per the legend: high in red, low in yellow) to the different groups (“basis”). Many of the *Enterobacteriales* 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>.



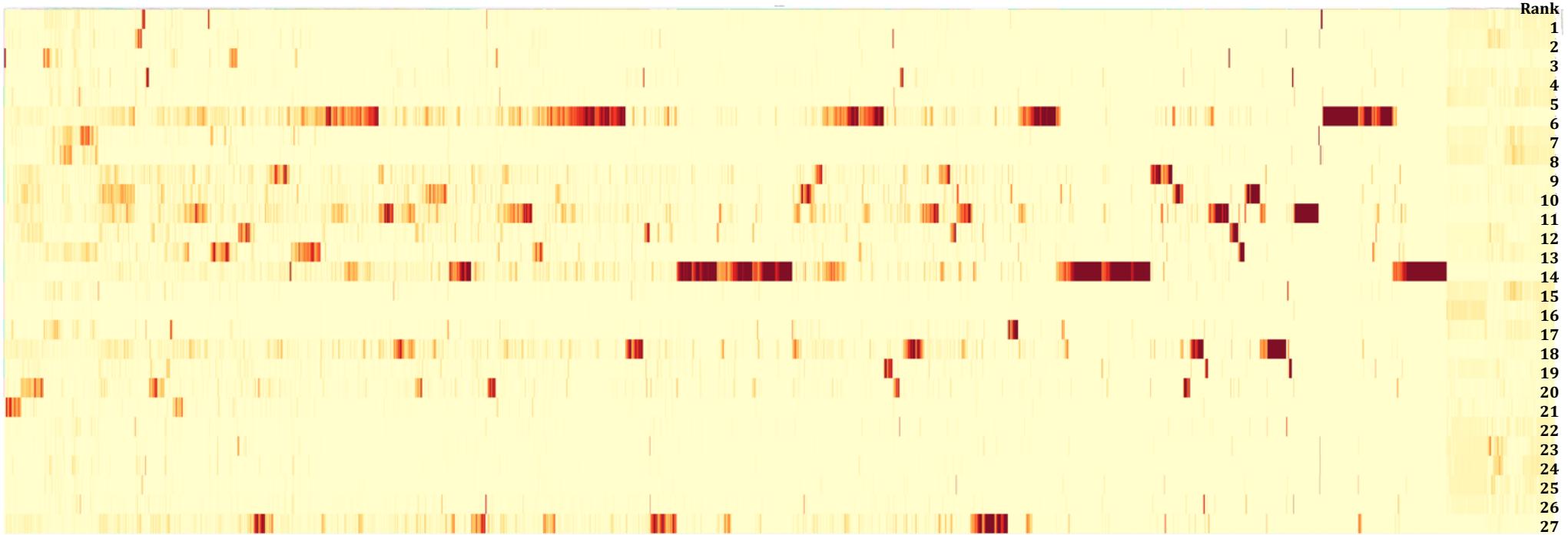


Figure S8. Heatmap of the mixture coefficient matrices, reflecting the probability of allocation of 7,126 plasmid-associated genes (x-axis) to 27 different groups (y-axis) based on their presence or absence in 3,164 samples using NMF. The legend below shows the probability of allocation (as per the legend: high in red, low in yellow) to the different groups (“basis”). The *Enterobacteriales*-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>.



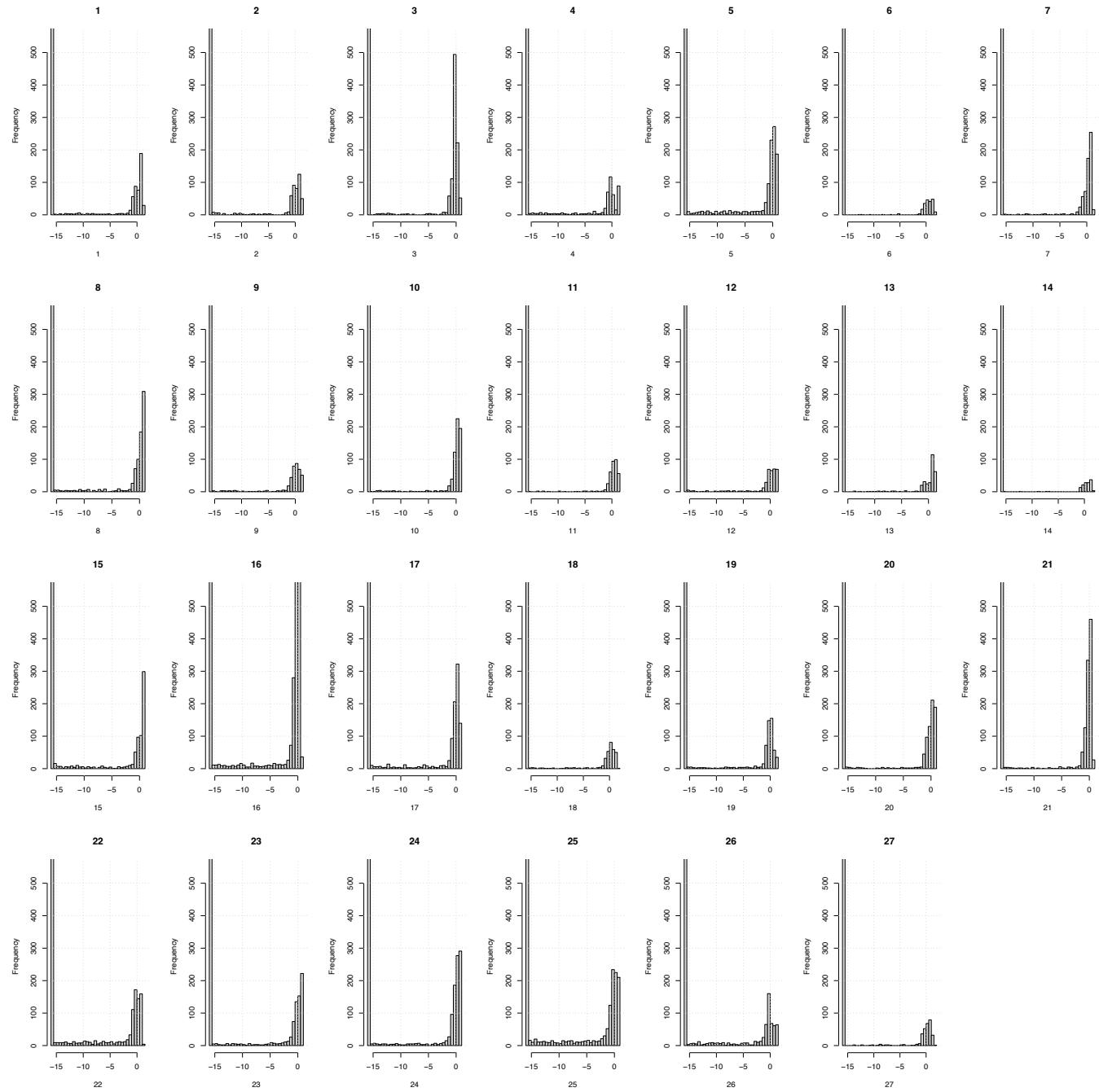


Figure S9. The frequency distributions of the rate of allocation of the 3,164 samples to each of 27 ranks from NMF based on their presence-absence rates in 7,126 plasmid-associated genes. The x-axes show the log₁₀-scaled probability of allocation to each rank. Many of the *Enterobacteriales* were allocated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

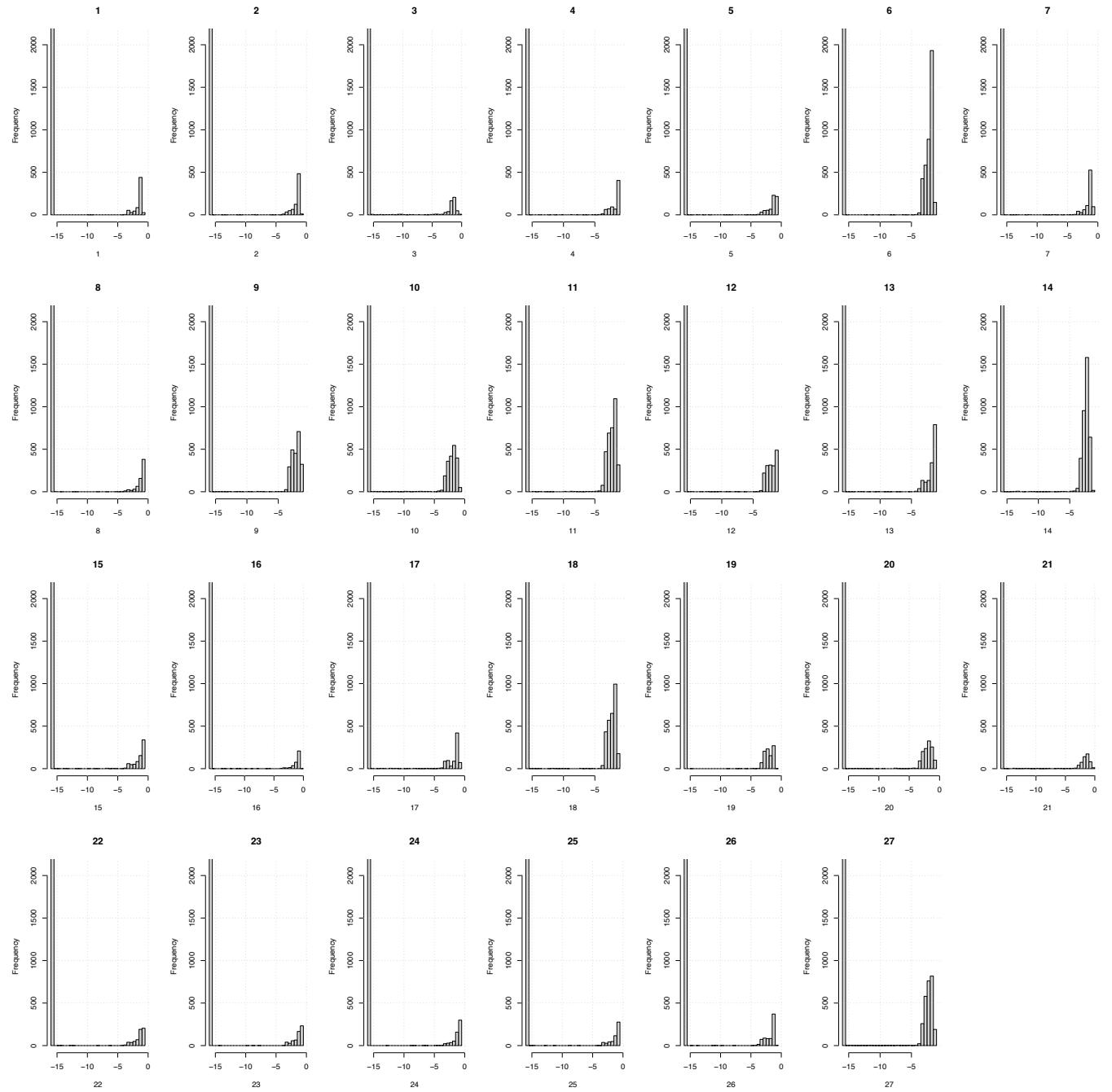


Figure S10. The frequency distributions of the rate of allocation of the 7,126 plasmid-associated genes to each of the 27 ranks from NMF based on presence or absence in 3,164 samples. The x-axes show the log10-scaled probability of allocation to each rank. Many *Enterobacteriales*-related genes were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

Rank 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29

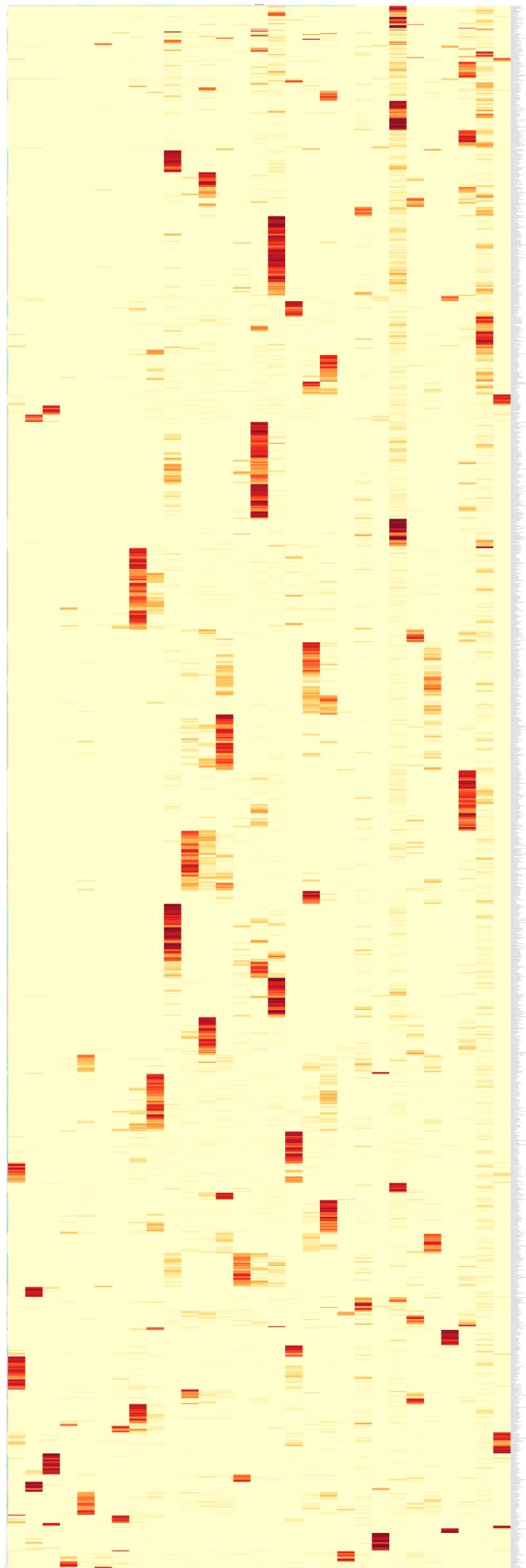
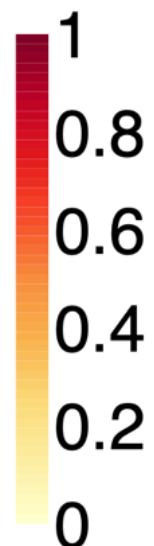
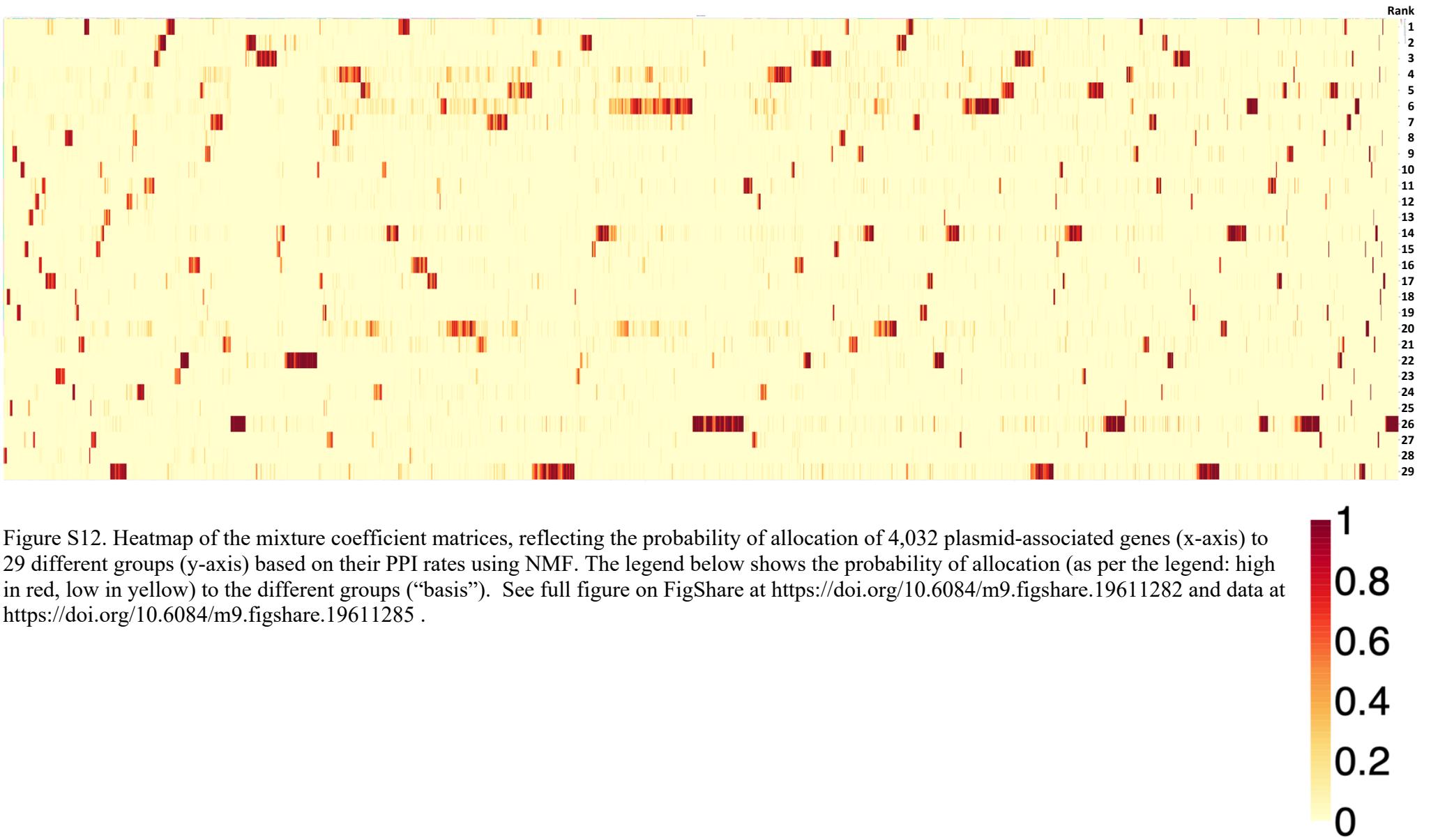


Figure S11. Heatmap of the basis coefficient matrices, reflecting the probability of allocation of the 1,713 samples (y-axis) to 29 different groups (x-axis) based on their plasmid gene PPI rates using NMF. The legend below shows the probability of allocation (as per the legend: high in red, low in yellow) to the different groups (“basis”). Many of the *Enterobacteriales* 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>.





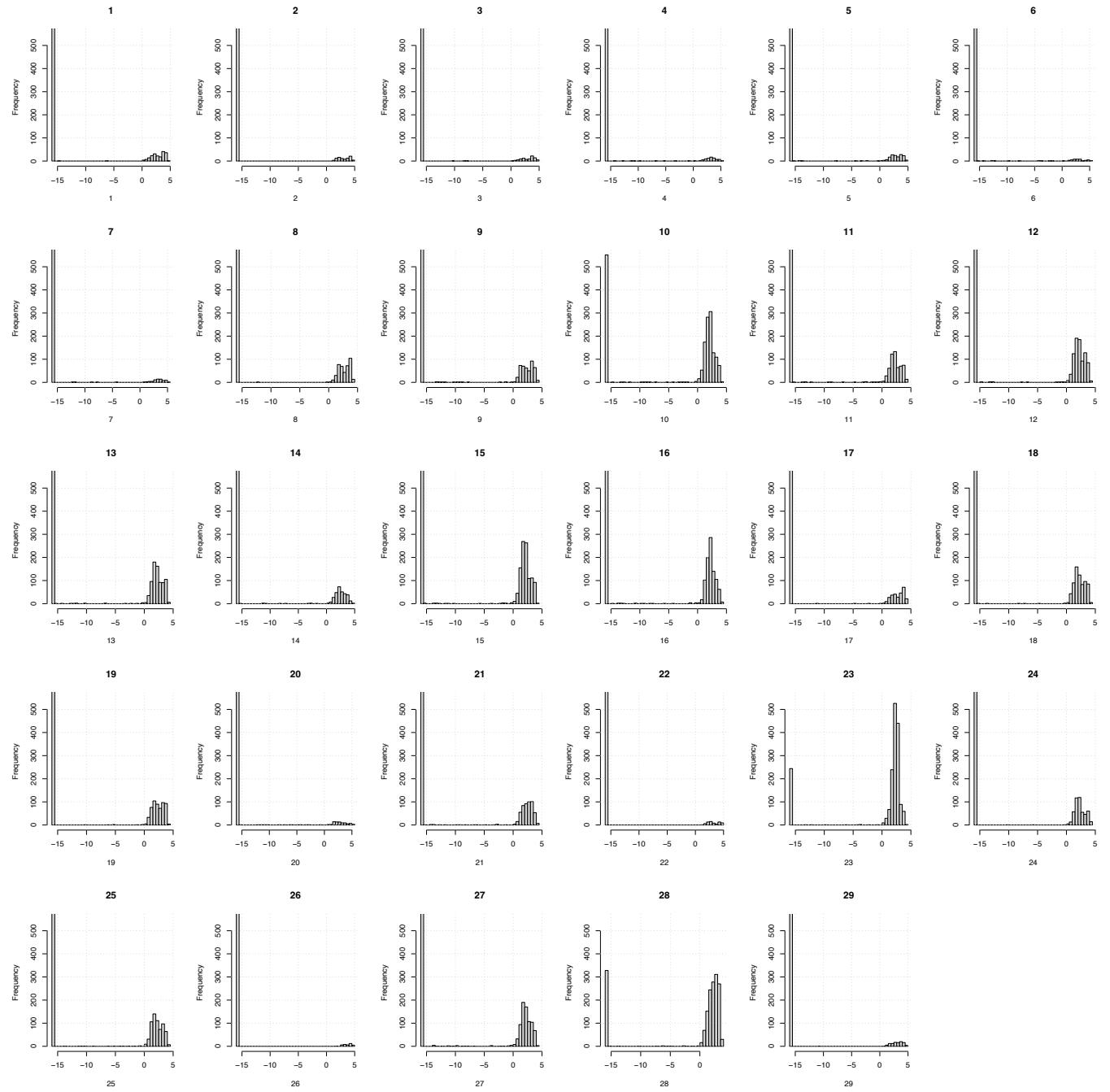


Figure S13. The frequency distributions of the rate of allocation of the 3,164 samples to each of 29 ranks from NMF based on their presence-absence rates in 1,713 plasmid-associated genes. The x-axes show the log10-scaled probability of allocation to each rank.

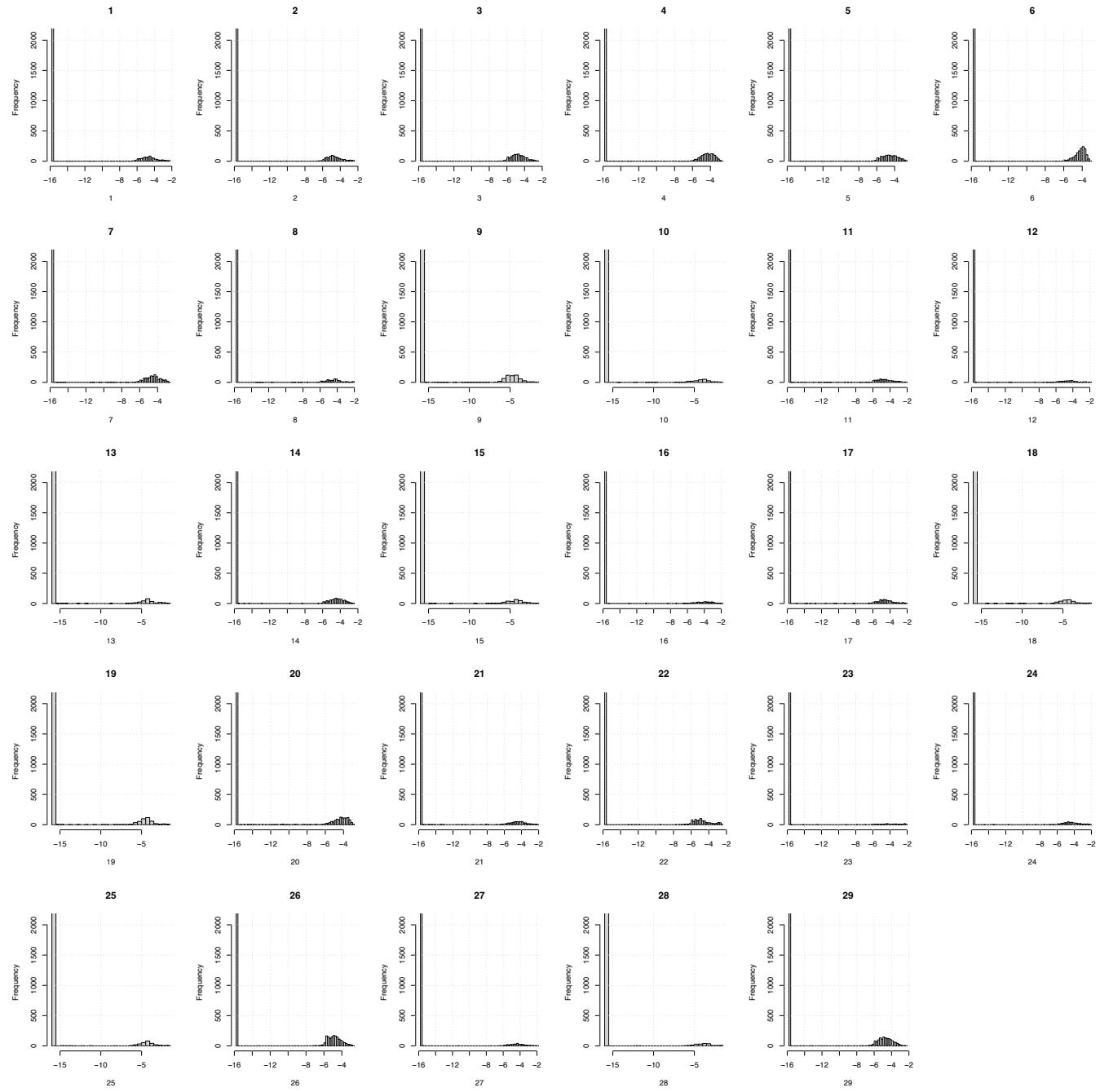


Figure S14. The frequency distributions of the rate of allocation of the 4,032 plasmid-associated genes to each of the 29 ranks from NMF based on presence or absence in 1,713 samples.

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

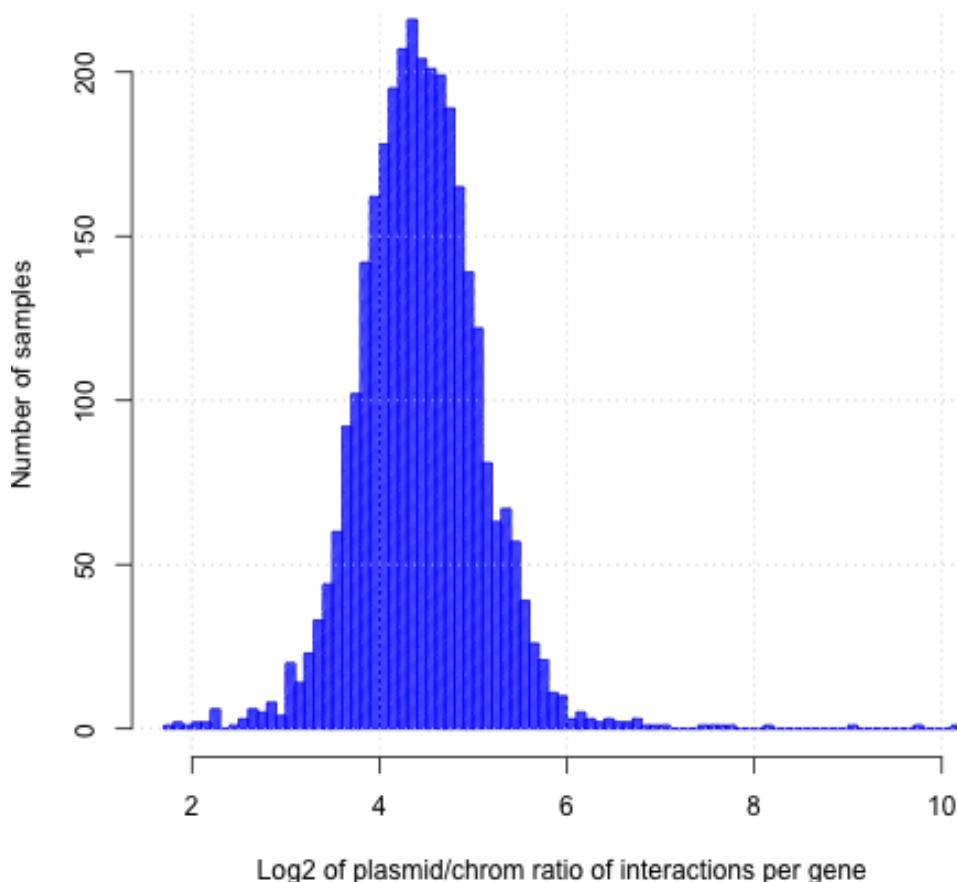
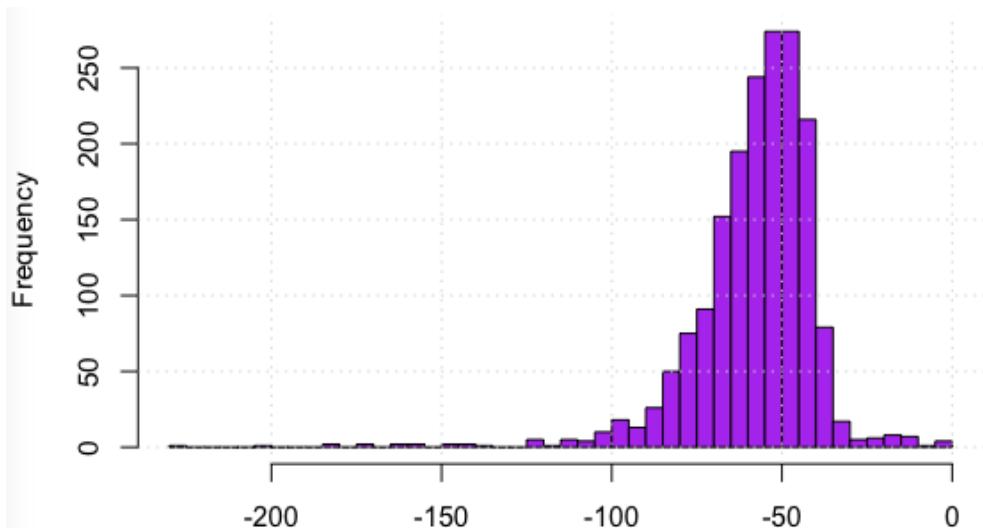
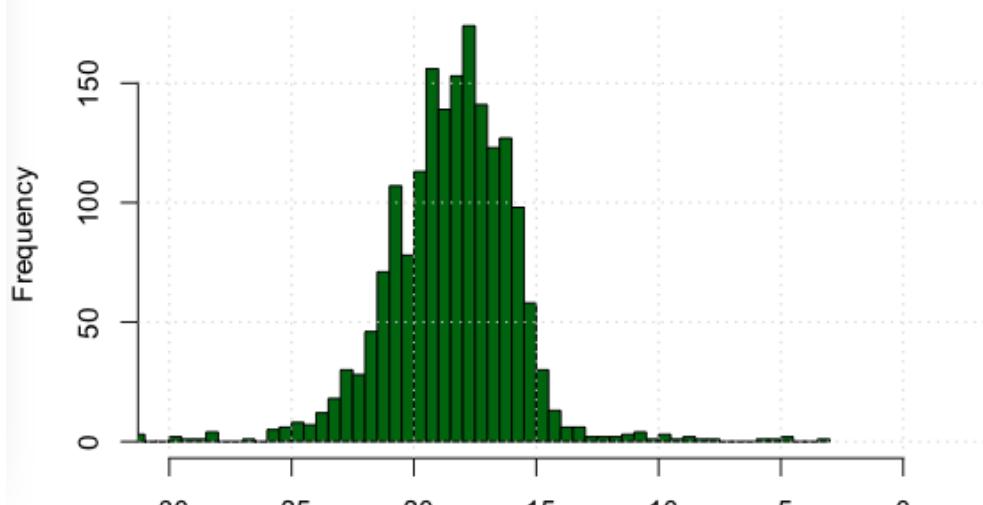


Figure S15. The variation in the log2-scaled average number of PPIs per gene for plasmid-associated versus chromosomal genes across the 3,164 samples with a mean PPIs per gene > 0.5 . The median ($\pm SD$) ratio was 4.5 ± 0.9 , highlighting much higher PPIs among plasmid-encoded genes.

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>.



Difference in chrom vs plasmid interactions log10-scaled p value distribution



Difference in chrom vs plasmid interactions t-stat distribution

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).

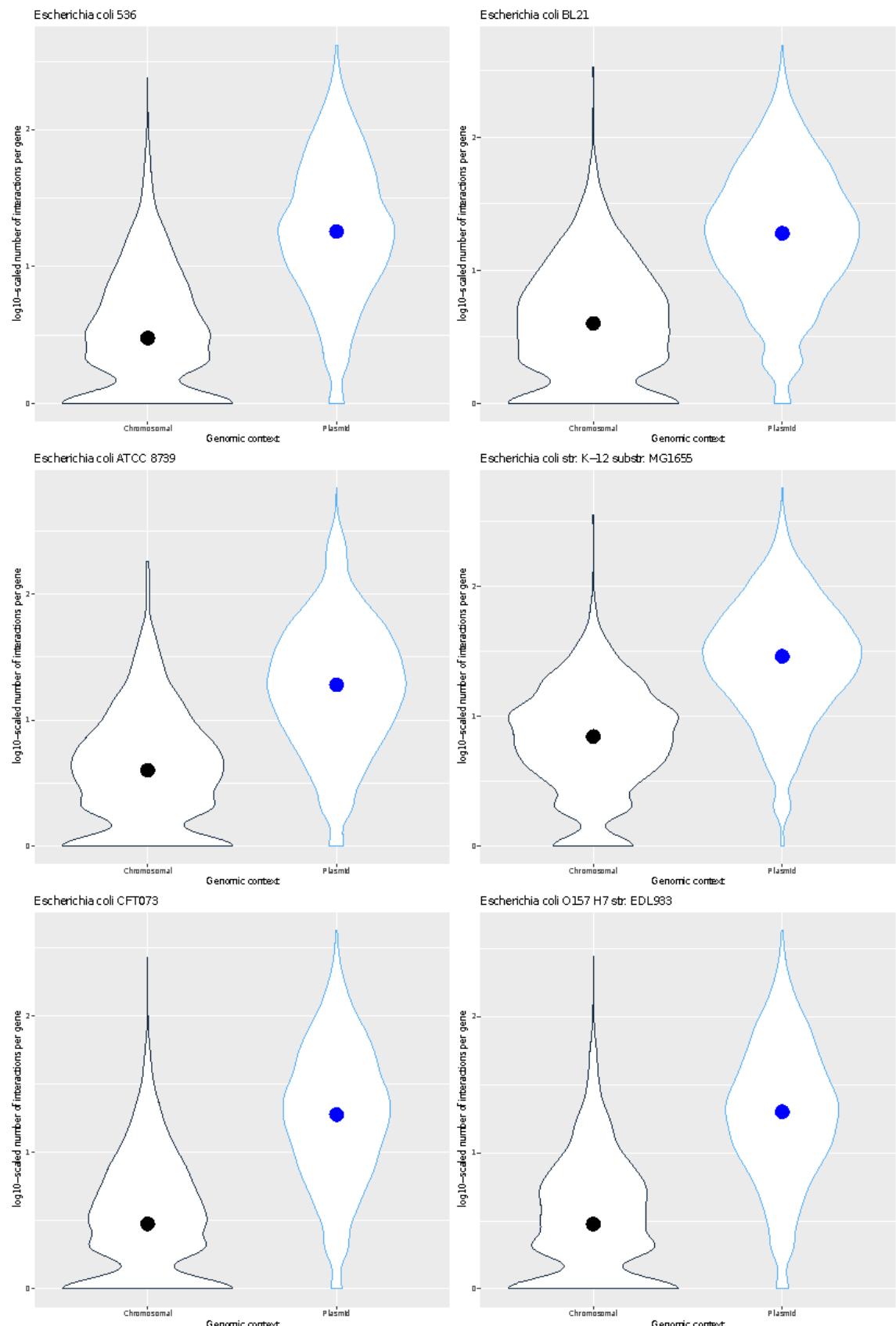


Figure S17. The \log_{10} -scaled numbers of chromosomal (black) and plasmid (blue) interactions per gene for the six *E. coli* samples: 536 (top left), BL21 (top right), ATCC8739 (middle left), K-12 substrain MG1655 (middle right), CFT073 (bottom left), O157:H7 substrain EDL933 (bottom right). The points show the observed values on top of the distribution shapes and the medians shown by large circles. The data for these six and the full set of chromosomal and plasmid interactions per gene for 4,377 samples are available as CSV files and as 3,008 PDF image files (the images were generated where there was sufficient PPI data to plot) on FigShare at doi: <https://doi.org/10.6084/m9.figshare.1957640>.

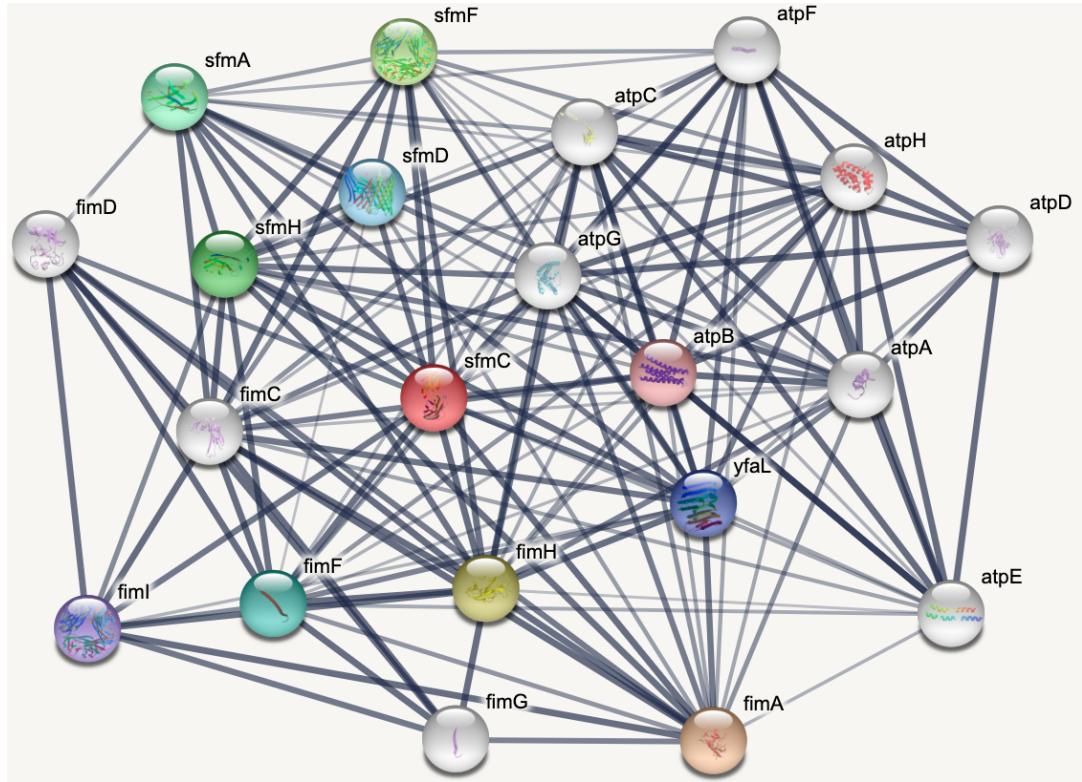


Figure S18. An *E. coli* K12 MG1655 (String ID 511145) PPI network from the StringDB website of 21 proteins in Table S4 centred on SfmC (a putative periplasmic pilus chaperone that is part of the *sfmACDHF* fimbrial operon) with 147 PPIs.

Gene	Context	Number of PPIs
<i>atpA</i>	Plasmid	82
<i>atpB</i>	Plasmid	32
<i>atpC</i>	Plasmid	43
<i>atpD</i>	Plasmid	65
<i>atpE</i>	Plasmid	32
<i>atpF</i>	Plasmid	43
<i>atpG</i>	Plasmid	60
<i>atpH</i>	Plasmid	53
<i>fimA</i>	Plasmid	12
<i>fimC</i>	Plasmid	8
<i>fimD</i>	Plasmid	6
<i>fimF</i>	Plasmid	4
<i>fimG</i>	Plasmid	2
<i>fimI</i>	Plasmid	10
<i>sfmA</i>	Chromosome	20
<i>sfmC</i>	Chromosome	12
<i>sfmD</i>	Chromosome	11
<i>sfmF</i>	Chromosome	54
<i>sfmH</i>	Chromosome	24
<i>yfaL</i>	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 \pm SD) PPIs compared to 24.3 ± 15.7 for the chromosomal ones.

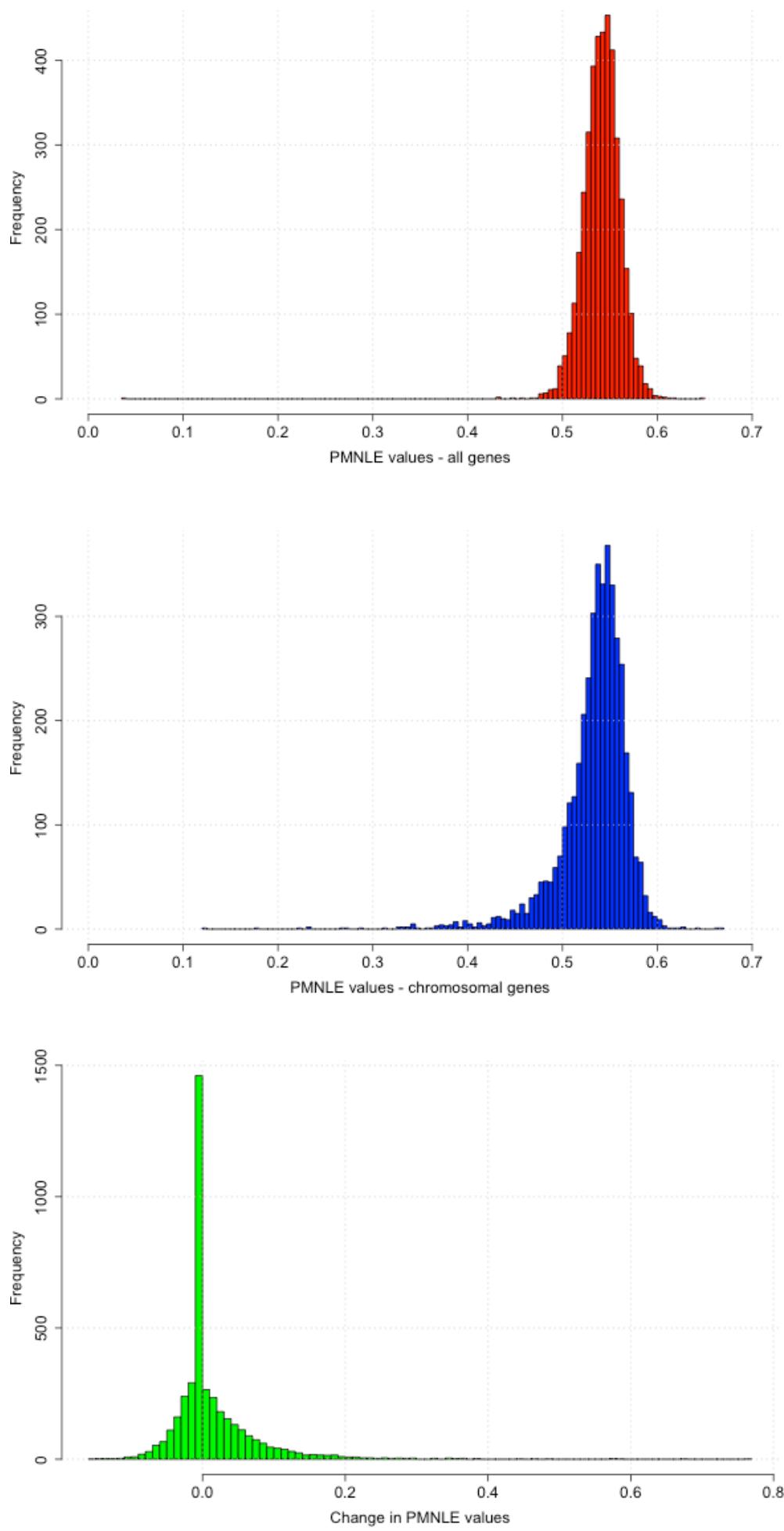


Figure S19. The distribution of aPMNLE values (x-axis) for all proteins (top, red), chromosomal ones (middle, blue) and the scaled difference between all and chromosomal proteins (bottom, green) across the bacterial samples. Values for 4,102 samples with valid aPMNLE values for all proteins and the chromosomal proteins. Note that the y-axes scales differ.

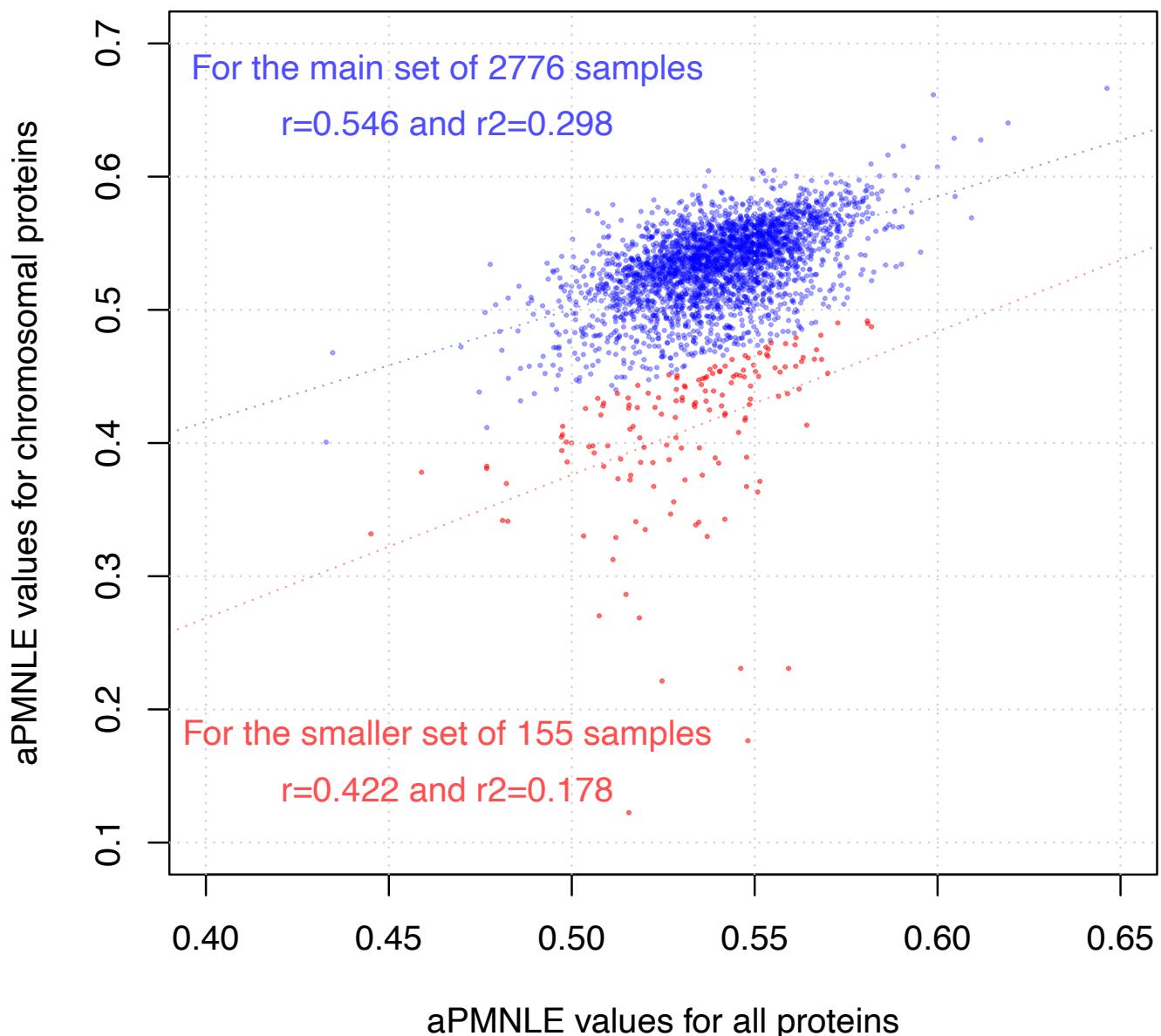


Figure S20. The association between the aPMNLE values for all proteins (x-axis) compared to those for chromosomal proteins (y-axis) for the group with no significant difference in these aPMNLE values ($n=2,776$ samples, blue) and the group where there was a much lower chromosomal aPMNLE value ($n=155$, red). The correlation between these aPMNLE metrics was higher for the main group of 2,776 samples ($r=0.55$) compared to the group with the change in aPMNLE value ($r=0.42$). The group of 155 had proportionally lower chromosomal aPMNLE values when compared to the other 2,776 samples (0.408 vs 0.535, t-test corrected $p=1.6e-58$).

Table S6. Bacterial samples (155, 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
<i>Streptococcus agalactiae</i> LMG 14747	97,558	1,828	0.982	2,186	99,385	368	0.597	0.521	-0.145
<i>Candidatus Amoebophilus asiaticus</i> 5a2	15,291	901	0.944	1,335	16,191	213	0.586	0.509	-0.152
<i>Arcanobacterium sp</i> 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.

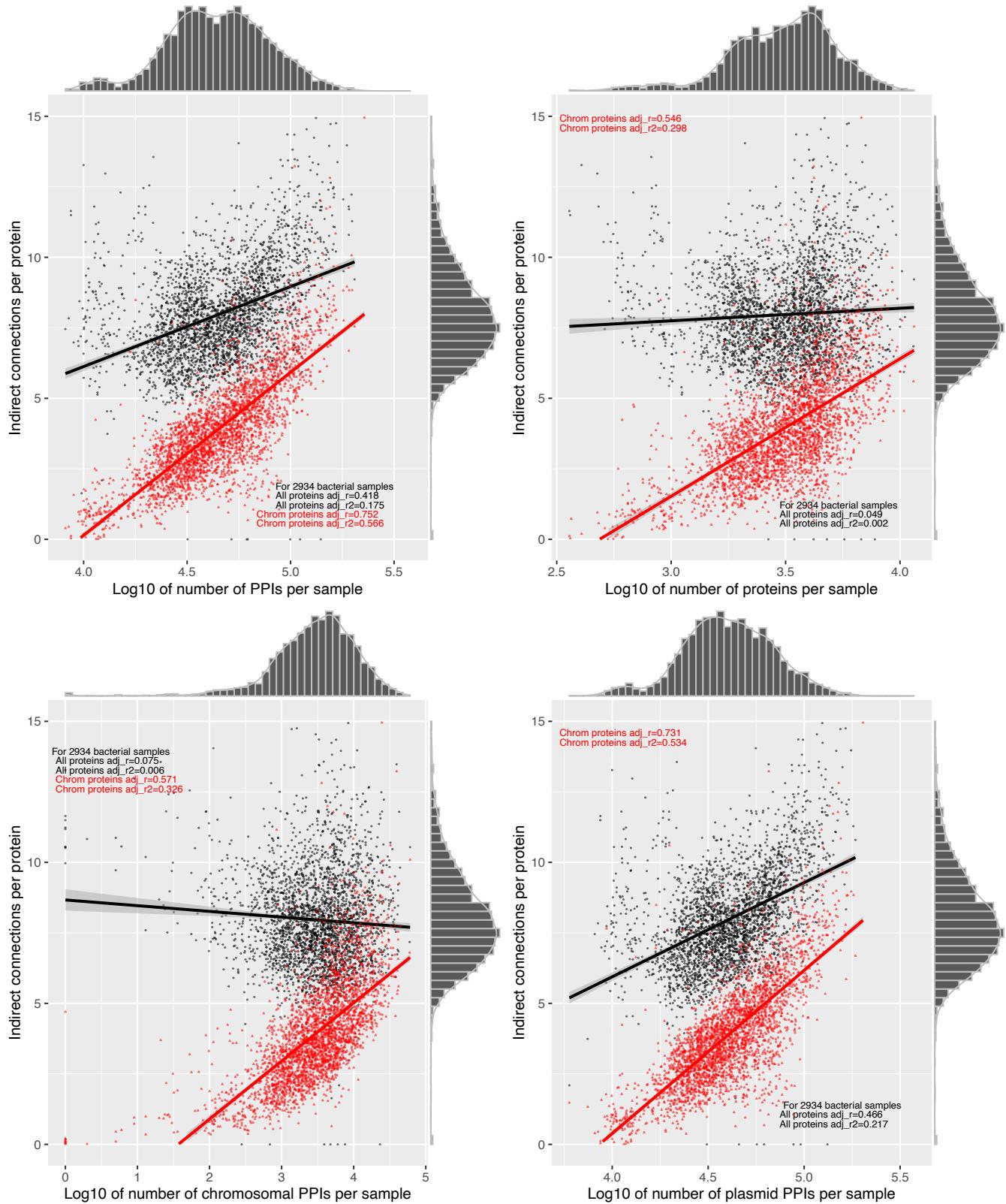


Figure S21. Varying levels of a positive correlation between (A) the log10-scaled number of PPIs, (B) log10-scaled number of plasmid-related PPIs, (C) log10-scaled number of chromosome-restricted PPIs, (D) log10-scaled number of proteins per sample, and number of indirect connections per protein (y-axis) 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.

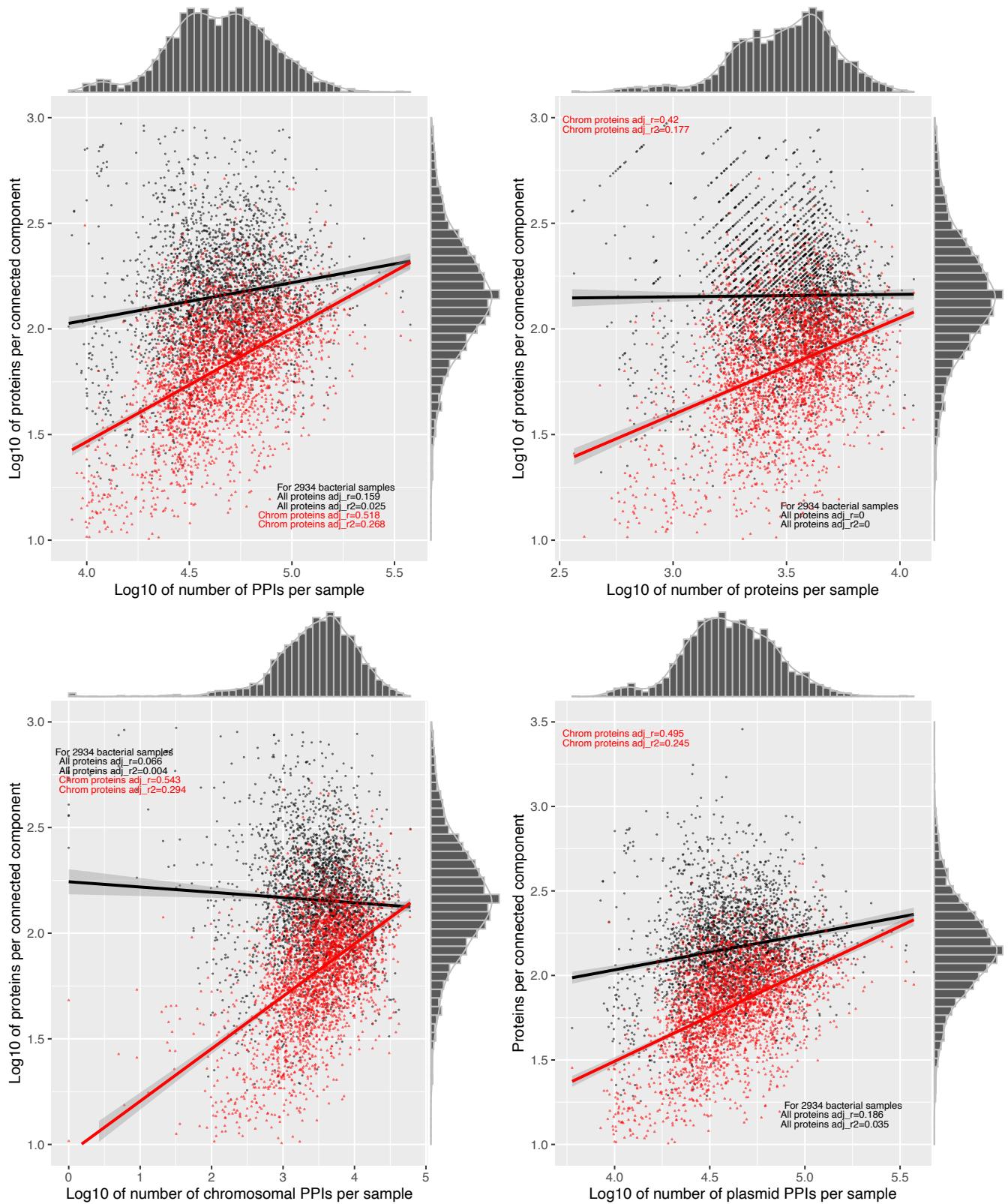


Figure S22. Varying levels of a positive correlation between (A) the log10-scaled number of PPIs, (B) log10-scaled number of plasmid-related PPIs, (C) log10-scaled number of chromosome-restricted PPIs, (D) log10-scaled number of proteins per sample, and number of proteins per connected components (y-axis) 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.