Bacterial plasmid-associated and chromosomal proteins have fundamentally different properties in protein interaction networks

Tim Downing[1], Alexander Rahm[2].

[1] School of Biotechnology, Dublin City University, Dublin, Ireland. Corresponding author. Email: Tim.Downing@dcu.ie. Orcid: 0000-0002-8385-6730.
[2] GAATI Lab, University of French Polynesia, Tahiti, French Polynesia. Email: Alexander.Rahm@upf.pf. Orcid: 0000-0002-5534-2716.

Running head: understanding what bacteria may acquire new plasmids

Keywords: AMR, plasmid, genome, protein-protein interaction, network.

Abstract

Plasmid-encoded genes facilitate the spread of antimicrobial resistance (AMR) though horizontal gene transfer (HGT). HGT driven by plasmids enables the diversification of pathogens into new anatomical and environmental niches, implying that plasmid-bourne genes can cooperate well with chromosomal genes. It is hypothesised that such mobile genes are functionally different to chromosomal ones due to this ability to encode non-essential functions like AMR and traverse distinct host cells. The effect of plasmid-related gene gain on protein-protein interaction network topology is an important question in this area. Moreover, the extent to which these chromosomally- and plasmid-encoded proteins interact with proteins from their own groups compared to the levels with the other group remains unclear. Here, we examined the incidence and protein-protein interactions of all known plasmid-linked genes (n=9,172) across representative specimens from all bacteria (n=4,419) using all known unique plasmids. We found that such plasmid-related genes constitute on average of 10.5% of the total number of genes per bacterial sample, and that plasmid genes preferentially associated with different species such that they provide moderate taxonomical power. Surprisingly, plasmid-related proteins had both more direct and indirect protein-protein interactions compared to chromosomal proteins, countering the hypothesis that genes with higher mobility rates should have fewer protein-level interactions. Nonetheless, topological analysis and investigation of the protein-protein interaction networks' connectivity and change in the number of independent components demonstrated that the plasmid-related proteins had limited overall impact in most bacterial samples (>94%). This paper assembled extensive data on plasmid-related proteins, their interactions and associations with diverse bacterial specimens that is available for the community to investigate in more detail.

Significance as a bioresource to the community

This paper outlines the compilation and analysis of bacterial plasmid-encoded genes and protein-level interactions that will be a useful resource for the community. All input data, output data, code and high-resolution figures is available on FigShare and Github. We extracted protein-protein interaction data for all available valid bacterial samples (n=4,419) to differentiate plasmid-related genes and PPIs from chromosomal ones. We compiled all known unique plasmids (n=18,628) to identify 9,172 unique plasmid-related genes using gene name information, which may require refinement is a useful resource for the community because we found that these had some power to resolve known taxonomic differences, suggesting host-gene co-evolution. The co-varying plasmid-related genes and samples are available for more specific testing. We used this data to identify protein-protein interactions involving plasmid proteins per sample, and showed that plasmid-related proteins tended here to have more protein-protein interactions than chromosomal ones. However, this varied across samples, and thus data allows the community to explore features of selected bacteria and/or genes. We identified a small number of samples (n=158) where the plasmid-related proteins had major effects on protein-protein interaction network structure, which is possible to explore more through this bioresource.

Data Summary

Item	Brief description	FigShare link
Table S1	Plasmid-related genes and their associated plasmid link(s)	https://doi.org/10.6084/m9.figshare.19525630
Table S3	Key features of the bacterial samples	https://doi.org/10.6084/m9.figshare.19525708
Table S4	Plasmid and chromosomal genes' interaction rates per sample	https://doi.org/10.6084/m9.figshare.19575820
Table S6	155 bacterial samples with large aPMNLE changes	https://doi.org/10.6084/m9.figshare.19525681
Figure S2	Heatmap of the bacterial plasmid-related genes	https://doi.org/10.6084/m9.figshare.19525453
Figure S3	Dendrogram of the bacteria based on plasmid-related gene similarity	https://doi.org/10.6084/m9.figshare.19525435
Figure S5	Heatmap of the bacteria and plasmid-related genes alphabetically sorted	https://doi.org/10.6084/m9.figshare.19525465
Data for Figure S5	Presence-absence data with bacteria as rows and genes as columns	https://doi.org/10.6084/m9.figshare.19525471
Figure S7	NMF heatmap of the metasample clustering of plasmid genes	https://doi.org/10.6084/m9.figshare.19525492
Data for Figure S7	Numerical data with 3,164 samples as rows and probabilities as columns	https://doi.org/10.6084/m9.figshare.19525480
Figure S8	NMF heatmap of the mixture coefficient for the plasmid-related genes	https://doi.org/10.6084/m9.figshare.19525591
Data for Figure S8	Numerical data with 7,126 genes as rows and probabilities as columns	https://doi.org/10.6084/m9.figshare.19525597
Figure S11	NMF heatmap of the metasample clustering of plasmid PPI rates	https://doi.org/10.6084/m9.figshare.19611276
Data for Figure S11	Numerical data with 1,713 samples as rows and probabilities as columns	https://doi.org/10.6084/m9.figshare.19611279
Figure S12	NMF heatmap of the mixture coefficient for the plasmid-related PPI rates	https://doi.org/10.6084/m9.figshare.19611282
Data for Figure S12	Numerical data with 4,032 genes as rows and probabilities as columns	https://doi.org/10.6084/m9.figshare.19611285
Data for all PPI analysis	4,377 samples' CSV files for all proteins with gene name, plasmid status and the total number of PPIs per sample	https://doi.org/10.6084/m9.figshare.19672527
Data for chrom PPI analysis	4,377 samples' CSV files for chromo-somal proteins with gene name, plasmid status & the number of PPIs per sample	shorturl.at/acuDF
Data for Figure S17	The numbers of chromosomal & plasmid genes for 4,377 samples as CSV files and as 3,008 PDF image files.	https://doi.org/10.6084/m9.figshare.19576408

Table 1. Large datasets available from this study on FigShare, showing the item, description and the corresponding DOI link. See the Supplementary Data for more. The code associated with this paper is available on Github at https://github.com/downingtim/mobilome_2022 and data files are there at https://figshare.com/projects/Bacteria_mobilome_2022/136720 "Bacteria_mobilome_2022".

Introduction

Plasmid-mediated horizontal gene transfer (HGT) underpins much of the current antimicrobial resistance (AMR) crisis (San Millan 2018, Crofts et al 2017). Plasmids are short extrachromosomal DNA elements that are typically circular and with variable copy numbers. Unlike non-conjugative plasmids, conjugative plasmids possess machinery enabling HGT between bacterial cells. This mobilises plasmid-bourne genes to move between bacterial cells in the same niche (Wein et al 2021), where these genes may allow new phenotypes (Norman et al 2009, Downing 2015, Hall et al 2017, Decano et al 2021). Plasmid-encoded genes encode a wide range of functions, including AMR, virulence, metabolism, and symbiosis (Ahmer et al 1999, Stasiak et al 2014), facilitating spread into new environmental niches (Niehus et al 2015). Importantly, they depend on the environmental context to be beneficial, such that they can be lost when no longer required (Niehus et al 2015).

Mobile genes form part of the accessory genome, distinct from chromosomal genes that typically compose the core genome (Brockhurst et al 2019), though some chromosomal genes may not be universal within a collection, and so are part of the accessory genome. Some genes can be encoded on both plasmids and chromosomes, and there are numerous instances in which beneficial genes have been mobilised to spread from plasmids to chromosomes, such as *bla*CTX-M-15 (Agyekum et al 2016, Huang et al 2017, Irrgang et al 2017, Decano et al 2019a, Decano et al 2019b, Yoon et al 2020, Ludden et al 2020, Bevan et al 2021, Shawa et al 2021). In this study, we examined the broad patterns across bacteria using the simplistic categorisation of plasmid-related versus chromosomal, which provides a foundation to examine the exact genome composition of a specific isolate from a sample.

Plasmids impose a metabolic fitness cost on their host cells (San Millan & MacLean 2017, Baltrus 2013, Harrison & Brockhurst 2012) and yet persist widely thanks to infectious spread, stability mechanisms, and chromosomal compensatory mutations (Stalder et al 2017). Plasmids are genetically very diverse, and this is paralleled by extensive variation in AMR gene profiles across bacteria (Ho et al 2020, Coelho et al 2021). The co-evolution of plasmid-related genes with chromosomal ones are consequently of keen interest. This matters because a suitable host-recipient genetic background is needed such that the transferred gene can be acquired and replicated (Forsberg et al 2014, Soucy et al 2015, Andam & Gogarten 2011). In addition, donor & recipient cells may be genetically distinct and yet have high HGT rates (Smillie et al 2011, Brito et al 2016), suggesting additional factors affecting gene and plasmid retention to quantify.

HGT depends on the physiological compatibility so that a plasmid-bourne gene can be expressed and useful in the transconjugant recipient cells by interacting with host proteins in a beneficial manner (Porse et al 2018). The complexity hypothesis (Jain et al 1999, Novick & Doolittle 2020) asserts that protein-protein interaction (PPI) network connectivity and HGT rates are negatively correlated. In line with this, previous work on small sets of protein orthologs has found that less essential genes with higher HGT rates have fewer PPIs (Nakamura et al 2004, Aris-Brosou 2005, Puigbo et al 2010). HGT rates are higher for rarer genes like those in the accessory genome on plasmids, and lower for core genes (Dewar et al 2021, Touchon et al 2020). Moreover, gene context and physical gene distance are related to network proximity (Babu et al 2006) such that accessory genes and newly transferred genes are likely to be peripheral in PPI networks, including plasmid-related ones, whereas the chromosomal proteins should be more central.

Plasmid properties vary extensively based on the host cell (De Gelder et al 2007, Dunn et al 2021, Alonso-Del Valle et al 2021, Kottara et al 2018, Sheppard et al 2020, Gama et al 2020, Alderliesten et al 2020), suggesting that the interaction between host chromosomal proteins with plasmid-encoded ones is important. The evidence collated so far implies that chromosomes adapt to plasmids more than the reverse, typically by single nucleotide polymorphisms (SNPs) (Stadler et al 2017, Harrison et al 2015, San Millan et al 2014, Hall et al 2020), but sometimes by larger gene deletions (Modi et al 1991, Porse et al 2016, Lee & Marx 2012). This co-evolution of plasmid-encoded and chromosomal proteins

means more systematic approaches to examining the extent to which these proteins interact across bacterial species could provide insights into plasmid-host compatibility.

In this study, we assessed the hypothesis that chromosomal and plasmid-related genes have different interaction levels based on the host species, PPI context and PPI connectivity. Using a core dataset of 4,419 bacterial samples, we examined PPI network connectivity using PPI data for all chromosomal proteins and for plasmid-related ones. Importantly, HGT is not strongly affected by gene sequence composition and features (Porse et al 2018), supporting our simple classification of genes as present or absent here. We examined the changes in PPI network structure based on changes in indirect connections when plasmid-related proteins enter a host's chromosomal PPI network. We identified distinctive features associated with plasmid-related proteins and their PPIs, assessed their covariation, and thus provide a pilot study of plasmid-related protein PPIs across all bacteria.

Methods

Protein-protein interaction data extraction

We retrieved PPI information for all available valid bacterial genomes (n=4,445) from STRING database v12 (Szklarczyk et al 2021) with R v4.0.1 (R Core Team 2021) via Bioconductor v3.11 (Ihaka & Gentleman 1996) and packages Rentrez v1.2.2 (Winter 2017), STRINGdb v2.0.1 (Szklarczyk et al 2018), and StringR v1.4.0 (Wickham 2019) (Figure 1). Of the initial 4,445 potential samples listed, 22 were not accessible and four had more >300,000 PPIs, suggesting potential accuracy issues and so they were excluded from analysis. The valid data (n=4,419 samples) was processed and collated using Dplyr v1.0.8 (Wickham et al 2002), Forcats v0.5.1 (Wickham 2021), Grid v4.4.1, ReadR v2.1.2 (Wickham et al 2022), Readxl v1.4.0 (Wickham & Bryan 2022), Tibble v3.1.6 (Müller & Wickham 2021), TidyR v1.2.0 (Wickham & Girlich 2022), Tidyverse v1.3.0 (Wickham et al 2019), VennDiagram v1.7.3 (Chen 2022) and Xlsx v0.6.0. The total number of unique protein names in the entire dataset was 9,551,828, though the true number of gene clusters is likely smaller due to name inaccuracy and redundancy. Using a STRINGdb score threshold of 400, we used this data to compute the numbers of unique proteins and pairwise PPIs on chromosomes: this score threshold of 400 was used throughout this study. These 4,419 post-QC samples had 3,628.5±1,710.0 (mean±SD) proteins and a mean of 57,105.8±34,698.6 PPIs each. This and subsequent data below were visualised with R packages ggplot2 v3.3.5 (Wickham 1996) and ggrepel v0.9.1 (Slowikowski 2021). All reported p values were correct for multiple testing using the Benjamini-Hochberg approach.

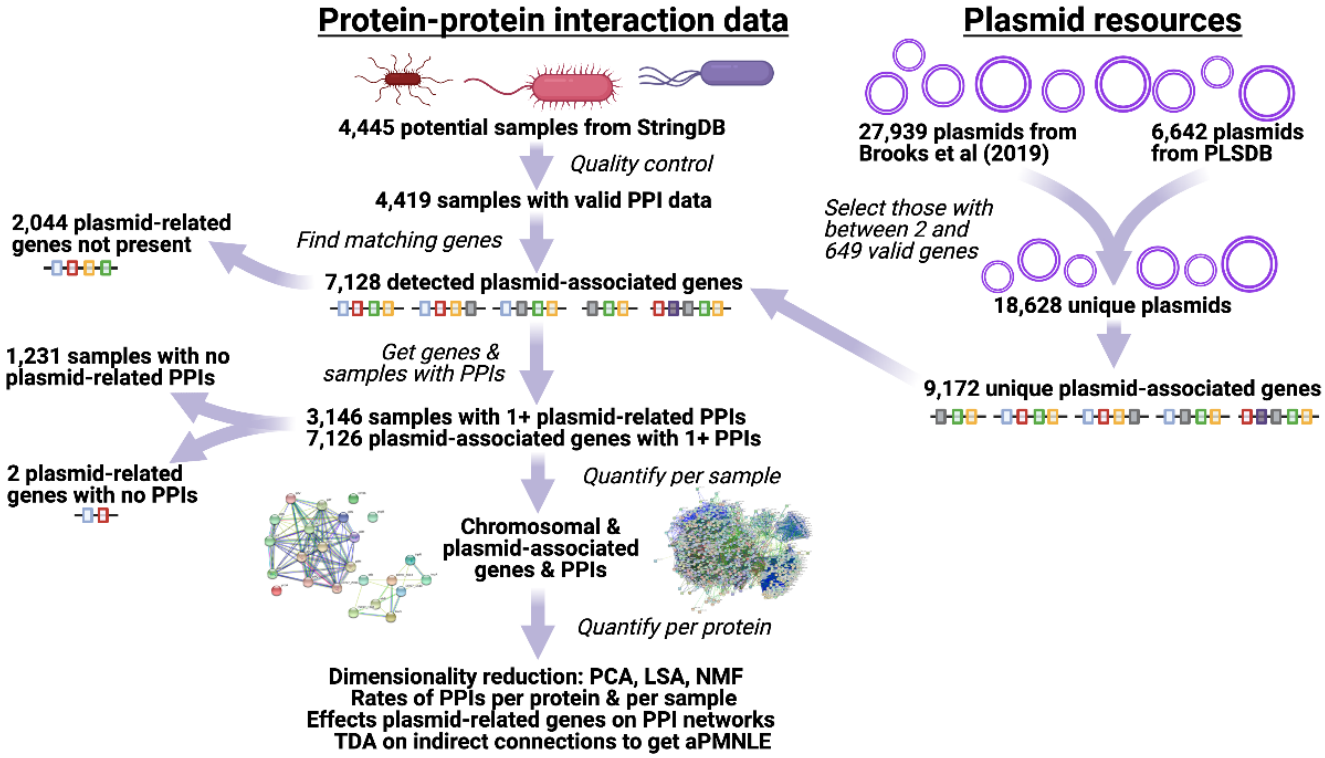


Figure 1. Summary of data sources and analytical steps. The plasmid resources (top right) and protein-protein interaction (PPI) data from StringDB were ultimately merged based on gene matches. Plasmid-related genes with PPIs were the focus, and gene prevalence and PPI level per sample were computed, as well as more general patterns across sample based on a range of complementary methods. Created with BioRender.com.

The accuracy of gene names in annotation is a pervasive issue where previous work has shown ~46±9% of genes can be functionally assigned (Zhou et al 2021). To explore that limiting factor here, we examined the frequency of four-digit (short) gene names to longer ones with five or more digits,

after removing any instances of names ending in an underscore or dash or dot followed by any digit. This showed that these samples had a median of 411±410 short names and a median of 2,931±1,744 long names each (Figure S1).

Identification of plasmid-associated genes across bacteria

To get all available plasmid-associated genes, we collated 32,839 plasmids: 27,939 from Brooks et al (2019) and 6,642 from PLSDB (v2020_11_19, Galata et al 2019) and retrieved gene annotation for each plasmid using Genbankr v1.16.0 (Figure 1). We focused on 18,628 with >1 properly annotated gene on GenBank and with <640 such genes (to avoid chromosome-related contigs). These plasmids had a median of 11 unique annotated genes per plasmid and possessed 9,172 unique plasmid-associated genes across all. The 15,531 plasmid accessions associated with these genes are available, totalling 363,425 gene-plasmid associations (Table S1, see full table FigShare doi: https://doi.org/10.6084/m9.figshare.19525630).

Identification of plasmid-associated interaction across bacteria

We identified the number of plasmid-related proteins and associated PPIs per protein per sample. A PPI was defined as plasmid-related if either protein had been found on a plasmid. The numbers of PPIs in which both proteins were exclusively chromosomal were small in the samples which contained plasmid genes. The 4,419 samples had a median of 58,293±23,905 plasmid-related PPIs and a median of 2,903±6,070 entirely chromosomal PPIs each. This meant the samples' median fraction of PPIs that were plasmid-related was 95.6%±7.2%, with the remainder chromosomal.

Screening for correlated bacterial samples and plasmid-associated genes

We examined patterns of covariance across the set of 4,419 samples and 9,172 plasmid-associated genes using three approaches. To resolve highly correlated sets of samples, the 1st assessed genes and sample-gene pairs using principal components analysis (PCA) implemented with R packages stats v3.6.2 and factoextra v1.0.7 (Kassambara & Mundt 2020). The 2nd was the distributional semantics method, latent semantic indexing (LSA): the probabilistic version has been used to examine other bacteria's drug resistance genomic data previously (Rusakovica et al 2014). Here, it applied a classical vector space model and singular value decomposition to determine the samples' plasmid-associated genes: allocating genes to samples with which they are strongly associated. LSA allocated the samples across the maximum possible 360 dimensions here, which represented the plasmid-associated genes.

The third was non-negative matrix factorisation (NMF) (Lee & Seung 1999) to examine the associations of individual plasmid-associated genes with individual bacterial samples using R packages nmf v0.23.0 (Gaujoux & Seoighe 2010) and Biobase v2.52.0 (Huber 2015). This was applied to 3,164 samples and 7,126 plasmid-associated genes as presence-absence data. The optimal rank was determined by examining changes in the cophenetic correlation coefficient across ranks 10-50 for 10 runs per rank to ensure clustering was stable and to select the maximum rank that retained a high cophenetic coefficient. This reduced the complexity of the data per sample to 27 groups (ranks) (cophenetic correlation coefficient maximised at 0.982). We also examined the covariation of samples with the PPI levels per gene for 1,713 samples and 4,032 plasmid-associated genes where the input data was the number of PPIs per protein per sample using NMF as above. The smaller dataset was due to a number of samples and genes with zero plasmid-related PPIs. The optimal rank was determined as above, and an estimate of 29 groups (ranks) was obtained (cophenetic correlation coefficient maximised at 0.911).

Topological data analysis of the bacterial protein-protein interaction networks

We developed computationally efficient metrics to investigate the topology of a PPI network, with in our focus the PPI network's indirect connectivity. Constructing the Vietoris-Rips complex (Vietoris

1927) on the PPI network allows counting "non-trivial loops": loops made of chains of PPIs (the PPIs are the edges of our Vietoris-Rips complex), such that some of the proteins involved were not directly connected to each other by a PPI. A pair of such proteins was then indirectly connected by two chains of PPIs along the loop (each obtained by starting from one protein and following the loop in one of the two directions until reaching the protein which was not directly connected to it by a PPI). For this reason, we called a non-trivial loop an "indirect connection".

The number of indirect connections and the number of PPIs are positively correlated, and previously we observed that the ratio of these two (#indirect connections/#PPIs) varies only moderately when we varied the StringDB combined score threshold, which modified the PPI network topology by reducing the number of PPIs (Decano et al 2020). If the indirect connectivity of the PPI network was defined as this ratio for one score threshold, it would lose information about the strength of the PPIs. Therefore, we introduced here a measurement which takes the PPI network topologies at all thresholds into account; we called it the Persistent Maximum of Non-trivial Loops per Edge (PMNLE). This is the maximal ratio (#indirect connections/#PPIs) which was reached or exceeded across an interval of at least 100 score thresholds. Measuring the PMNLE by calculating the indirect connections at all score thresholds was computationally inefficient, so to reduce the ecological impact by 20-fold, we used an approximate PMNLE (aPMNLE), which was the PMNLE measured only at score thresholds from 400 to 900 with a step of 20. For 16 samples, the scaled deviation between the PMNLE and aPMNLE was small at just 0.02% with a standard deviation of 0.25% (Table S2).

Processing the 4,419 PPI networks took <3 weeks on a single processor, which was completed for the PPI network with all protein and then with chromosomal proteins only. We used specialised open source software for this purpose (Rahm 2019), which constructed the two-dimensional part of the Vietoris-Rips complex by registering the trios of proteins (namely, triangles consisting of three proteins connected to one another by three PPIs) and then ran sparse matrix computations with the LinBox library (with LinBox v1.1.6) to obtain the 1st Betti number, which is the number of indirect connections, and the 0th Betti number, which is the number of connected components (subnetworks of the PPI network in which each pair of proteins is joined by a chain of PPIs - hence there is no chain of PPIs joining two distinct connected components). Additionally, we use the numbers of PPIs, trios, connected components and indirect connections computed at the combined score threshold 400 for comparison. Where unstated, the combined score threshold used was 400. In many bacteria's PPI networks, the purely chromosomal part determines the indirect connectivity, i.e., the aPMNLE measured on the full PPI network was close to the aPMNLE of the purely chromosomal part (so the influence of the plasmids on the aPMNLE was negligible in such samples). We examined the difference of the former minus the latter of these two aPMNLE values, which we scaled by the aPMNLE of the full PPI network to measure the percentage change when plasmid-linked proteins were removed.

Results

We retrieved PPI data from StringDB to compare the properties of plasmid-associated versus chromosomal proteins, with a view to the potential compatibility and signals of co-evolution for plasmid-chromosome combinations. This focused on the numbers of chromosome-linked PPIs, plasmid-associated PPIs, taxonomical classification based on plasmid gene profiles, relative rate of PPIs, indirect connections, and the effect on PPI network structure.

A positive correlation between the numbers of proteins and the numbers of interactions

We examined the association between the numbers of proteins and PPIs across representatives spanning 4,419 valid bacterial samples based on their unique annotated genes on StringDB. As expected, the numbers of proteins and PPIs was highly correlated (adjusted r^2=0.83, p=2.6e-16) (Figure 2, Table S3 – see full data at FigShare https://doi.org/10.6084/m9.figshare.19525708). This suggested that larger chromosomes have more proteins and thus PPIs. It also suggested variation in protein and PPI numbers across bacteria did not skew the association between the numbers of proteins and PPIs, enabling evaluation of features of plasmid-related genes in different samples combinations in more detail.

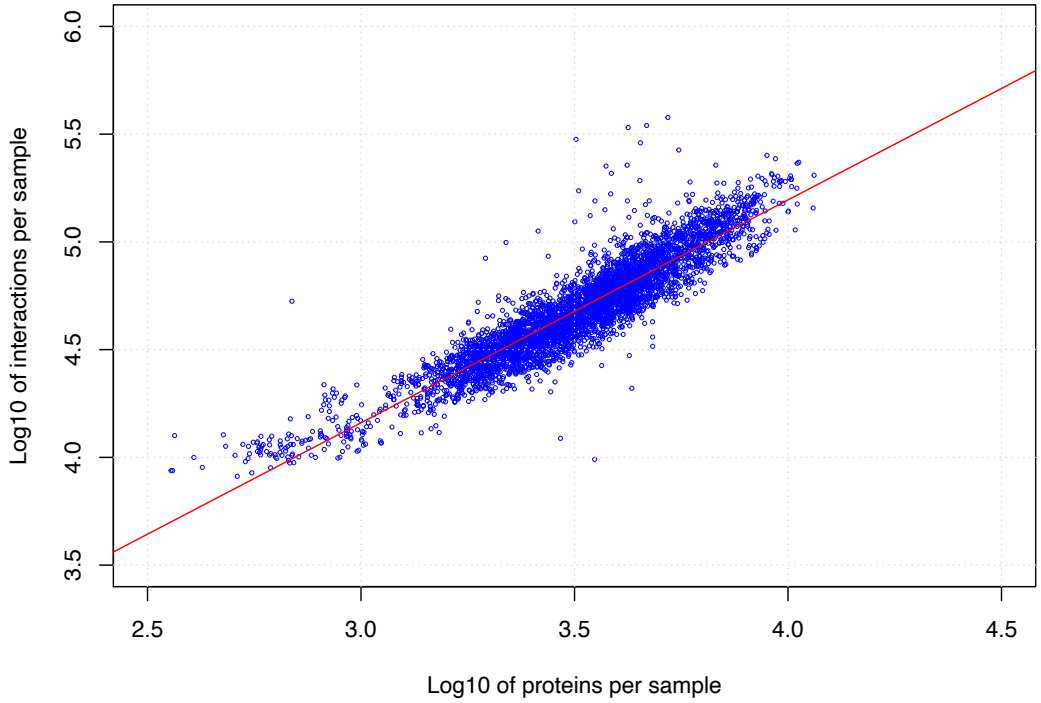


Figure 2. The number of proteins per bacterial genome (x-axis) was highly predictive of the number of PPIs (y-axis) (adjusted r^2=0.83). The data was for 4,419 bacterial samples, each of which is shown by a blue dot, with a linear model line of best fit shown in red.

Distinct patterns of plasmid-associated gene retention across bacterial samples

We compared these 4,419 samples' PPI data and the set of 9,172 unique plasmid-associated genes: 22.3% (2,044) of these genes were not detected in these samples, yielding 7,128 detected genes. The median number of samples in which a gene was found was nine with an interquartile range (IQR) of 55, indicating that plasmid-associated genes were rare (Figure 3, Table S3). The median number of plasmid-associated genes per sample was 385 with an IQR of 531, highlighting that these genes made up a small total fraction of bacterial genomes (Figure 3). This illustrated that most plasmid-associated genes were rare, with few common ones, and that the bacterial samples were so divergent that widely shared genes were not observed (Figure S2, see full figure FigShare doi: https://doi.org/10.6084/m9.figshare.19525453). Consequently, the average fraction of proteins that were plasmid-related per sample was 10.5% (380±330 out of 3,629±1,710).

Plasmid-associated proteins are rare and collectively interact with diverse chromosomal proteins

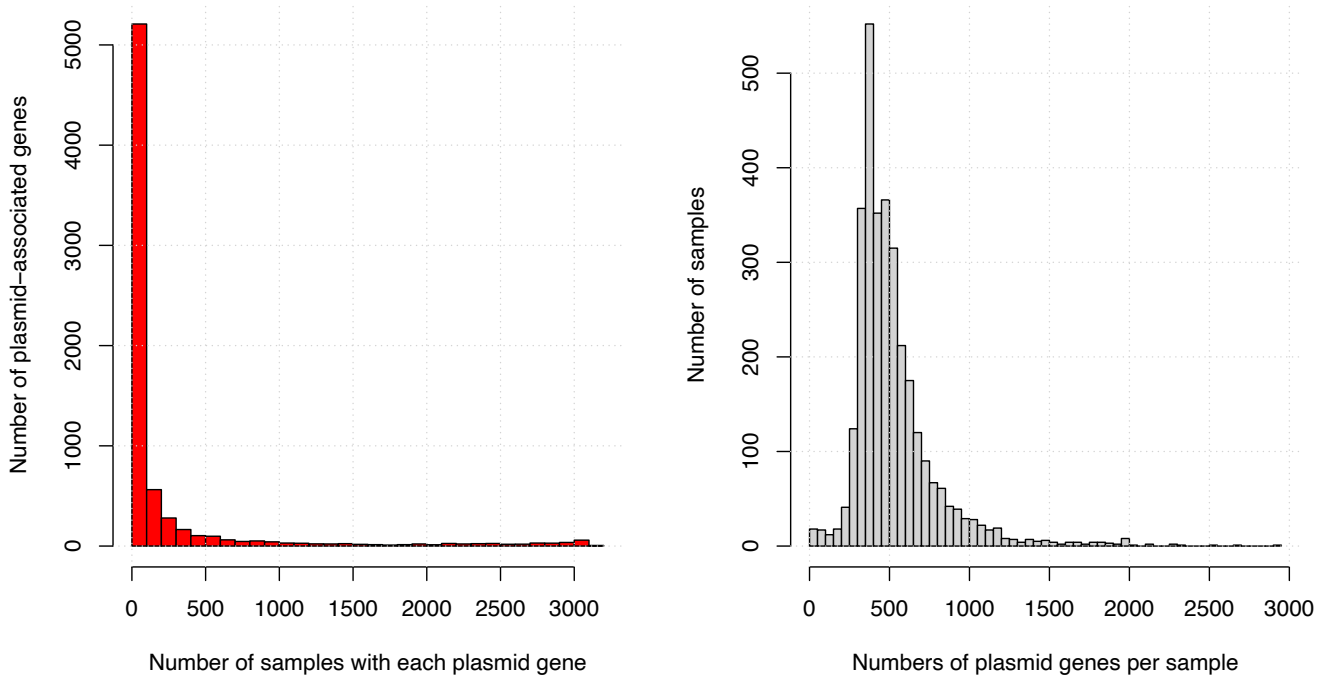


Figure 3. Left: The number of samples in which each plasmid-associated gene (red) was found was generally small. Right: The number of plasmid-associated genes per sample (grey) had a median of 384 with a skew toward numerous samples with higher rates of plasmid-linked genes. This figure excluded samples with no plasmid-associated genes.

We compared the samples' PPI data across the 7,128 unique plasmid-associated genes to identify PPIs where one or both proteins were plasmid-related (named plasmid-related PPIs here, see data at FigShare doi: https://doi.org/10.6084/m9.figshare.19672527) to distinguish them from instances where no proteins were found (chromosomal PPIs, see data per sample at shorturl.at/acuDF). 1,231 samples had no detected plasmid PPIs, 3,146 samples had >5,980 plasmid-related PPIs, and 42 samples lacked valid data. The 3,146 samples with 1+ plasmid PPIs collectively had 7,126 plasmid-associated proteins - two proteins had no PPIs. Across these 3,146 samples, there was a median of 40,368±28,941 plasmid-related PPIs and 3,244±6,240 chromosomal PPIs, with an average fraction of plasmid-related PPIs of 0.91±0.07 (Table S3). For these 3,164 samples, there was a positive association between the numbers of plasmid-related PPIs with the number of chromosomal PPIs (r^2=0.20, p=2.7e-16, Figure 4), but not with the fraction of PPIs that were plasmid-associated (r^2=0.01, p=0.71).

Plasmid-associated genes are structured and reflect taxonomic groups

Each of the 4,419 samples were clustered by similarity based on their data for the 7,128 unique plasmid-linked genes that had PPIs, where the plasmid-associated proteins per sample were considered as binary (presence-absence) data. There was moderate co-variation of plasmid-associated proteins across samples, such that evidence of taxonomic groups was apparent: for instance, the *Enterobacterales* were identified as a discrete clade (Figure S3, see full figure FigShare doi: https://doi.org/10.6084/m9.figshare.19525435). Evidence for this covariation from PCA indicated the samples' PC1 had 29.2% of variation, PC2 had 7.7%, and lower PCs minor effects (Figure S4A). The plasmid-related genes had a higher rate of covariation: 52.9% of variation was allocated to PC1, and just 4.6% to PC2 (Figure S4B, see full Figure S4 at FigShare doi: https://doi.org/10.6084/m9.figshare.19525465, and full presence-absence data at FigShare doi: https://doi.org/10.6084/m9.figshare.19525471). These patterns were largely recapitulated by LSA, which showed limited variation across the 1st (7.3% of variation) and 2nd (2.3% of variation) dimensions (Figure S6).

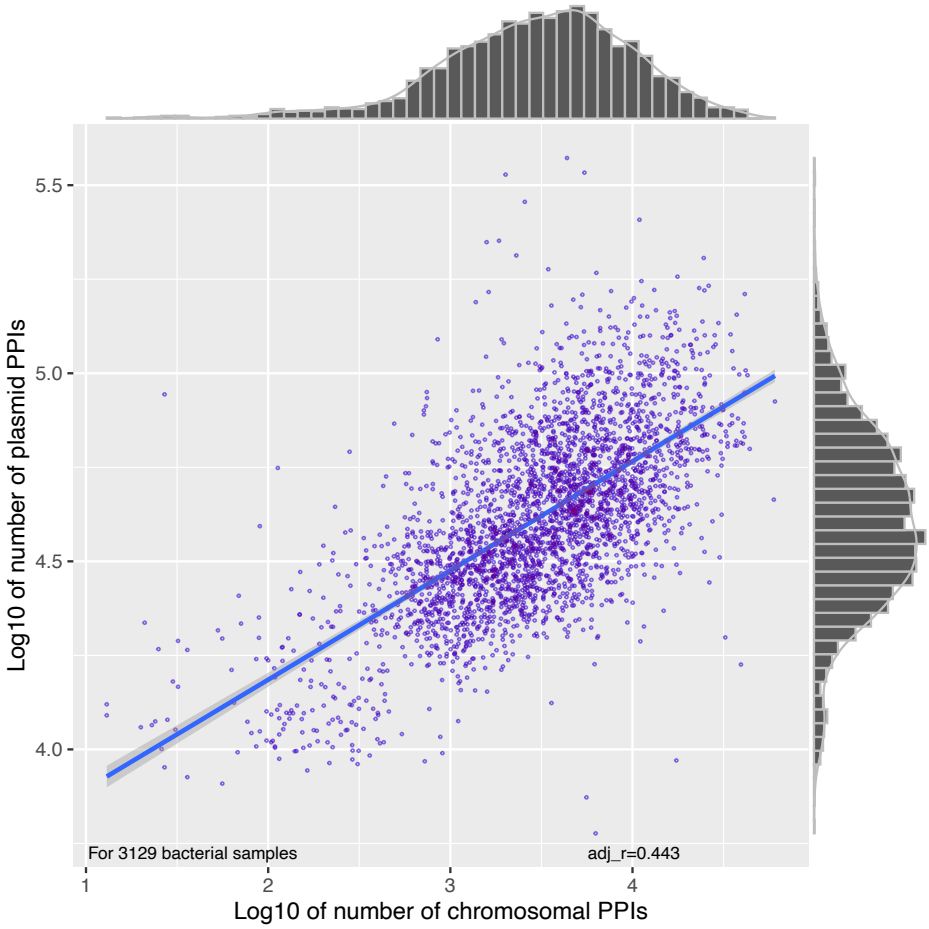


Figure 4. A positive correlation between the log10-scaled number of plasmid- (y-axis) and chromosomal-related (x-axis) PPIs among all 3,129 samples (purple points) with at least one plasmid-related PPI and ten chromosomal PPIs. The line of best fit (blue) shows a linear correlation of the log10 of the number of plasmid- and chromosomal-related (x-axis) PPIs ($r=0.443$, $r^2=0.20$, $p=2.7e-16$). The histograms indicate the marginal densities for each axis.

We examined the plasmid gene associations across samples that had at least one of the 7,126 plasmid-associated proteins. This allocated these samples and genes to 27 distinct groups based on their pattern of plasmid-associated protein sharing (Figure S7, see full figure at FigShare doi: https://doi.org/10.6084/m9.figshare.19525492 and data at FigShare doi: https://doi.org/10.6084/m9.figshare.19525480), and the proteins to these same 27 groups as mixture coefficients based on their prevalence in the samples (Figure S8, see full figure at FigShare doi: https://doi.org/10.6084/m9.figshare.19525591 and data at FigShare doi: https://doi.org/10.6084/m9.figshare.19525597). Most *Enterobacterales* were among the 200 samples with a factor from the same rank (6) that corresponded to 4,004 genes as mixture coefficients (Figure S9), which was far higher than the average number of genes associated with each rank (1,202). The *Enterobacterales*-linked genes included ones associated with novel phenotypes, such as the possession of *fim* (FimA-I/Z), *pap* (PapC/D/X) and *pil* (PilA/J/L/Q/S/T) operon proteins. 130 samples including many *Enterobacterales* were also allocated to rank 14 with 3,639 genes as mixture coefficients, of which 2,075 were also in rank 6 (Jaccard index 0.65, Figure S10).

We also investigated 4,032 plasmid-related proteins' associations based on PPI rates across 1,713 samples. These subsets of proteins and samples were required to have more than one PPI to provide valid input data. This allocated these samples and genes to 29 distinct groups based on their pattern of PPI numbers (Figure S11, see full figure at FigShare doi: https://doi.org/10.6084/m9.figshare.19611276 and data at FigShare doi: https://doi.org/10.6084/m9.figshare.19611279), and these same proteins to the 29 groups as mixture coefficients based on their PPI numbers for each sample (Figure S12, see full figure at FigShare doi: https://doi.org/10.6084/m9.figshare.19611282 and data at FigShare doi: https://doi.org/10.6084/m9.figshare.19611285). Some *Enterobacterales* were among the 45 samples with a factor from the same rank (6) that corresponded to 1,445 genes as mixture coefficients (Figure S13), which again was far higher than the average number of genes associated with each rank (666.2), but other *Enterobacterales* were allocated with a wide range of ranks and gene PPI rates (Figure S14).

Plasmid-associated proteins have much higher rates of protein-protein interactions

Unexpectedly, chromosomal proteins had a much lower rate of average PPIs per sample per protein compared to plasmid-associated proteins (2.8±1.5 vs 56.8±13.6) for proteins that had at least one PPI across the 3,146 samples possessing plasmid-associated proteins (Figure 5) (Table S4, see data at FigShare doi: https://doi.org/10.6084/m9.figshare.19575820). This corresponded to a 20-fold difference in PPI rates, even though these proteins were from the same samples (Figure S15). This pattern was consistent across all samples based on pairwise t-tests of this chromosomal-plasmid rate difference for each individual sample (Figure S16), such as all in *E. coli* samples examined here (Figure S17, see data at FigShare doi: https://doi.org/10.6084/m9.figshare.1111). As an illustration, a PPI network centred on the SfmC protein in *E. coli* K12 MG1655 showed high levels of connectivity across a PPI network of 20 other proteins (Figure S18, Table S5).

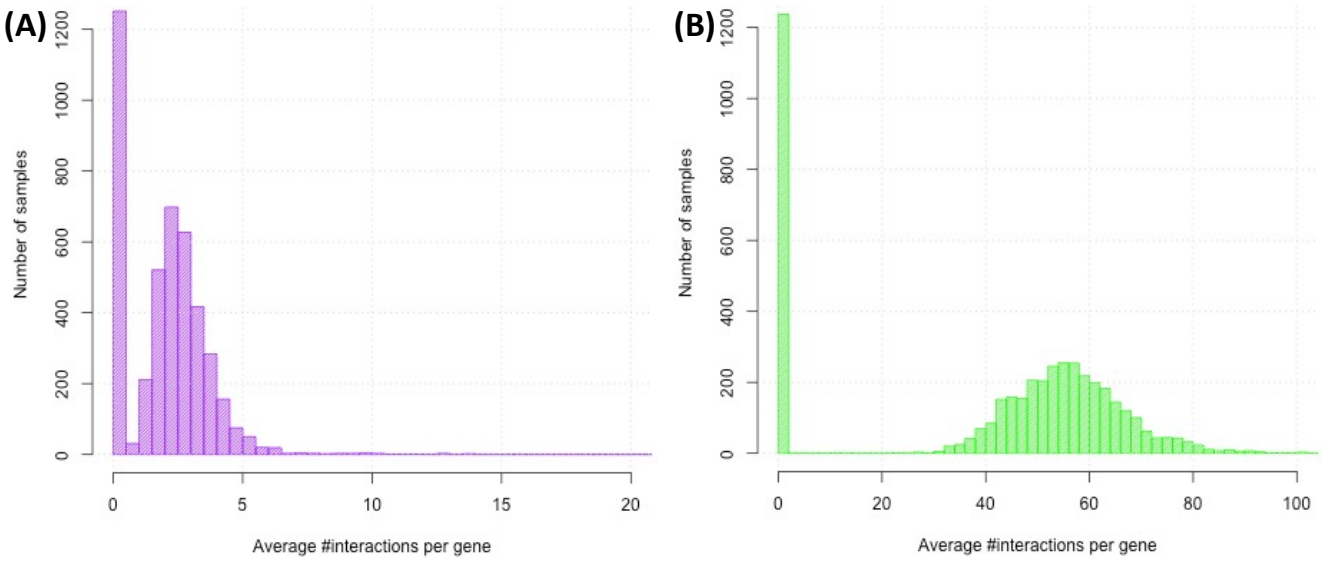


Figure 5. The variation in the mean number of PPIs per protein across 3,146 samples with plasmid-associated proteins samples showed (A) far lower PPI rates for chromosomal proteins (purple) than (B) plasmid-associated ones (green). 1,240 samples had low levels of detected PPIs (mean < 0.1). Among the samples with higher rates of detected PPIs (mean > 0.1), the mean (±SD) for chromosomal proteins was much lower than that for plasmid-associated ones (2.8±1.5 vs 56.8±13.6). Note the x-axes for each plot differ to visualise the numbers per group more clearly.

Protein-protein interaction network topology is generally not affected by plasmid proteins

To examine the effect of plasmid-related proteins on PPI network topology, we examined the aPMNLE for all proteins and for chromosomal proteins only, such that the difference between these aPMNLE values indicated the plasmid-linked effects (Figure S19). We hypothesised that if plasmid and chromosomal proteins were equally spread across the PPI network, then the indirect connectivity (as measured by the aPMNLE) would increase due to the loss of connections. On the other hand, if plasmid-associated proteins were on average more peripheral in the PPI network, then their loss may not change the general indirect connectivity (and aPMNLE). In such a structured network, if the plasmid-linked proteins had relatively more indirect connections, then the indirect connectivity (aPMNLE) would decrease if the plasmid-related proteins were removed (and *vice versa* for an increase).

A majority of samples (94.6%, 2,777 out of 2,934) showed no substantive difference in the aPMNLE values between all and chromosomal proteins, indicating plasmid-related proteins had no sizable

effects on the PPI network structure. These 2,934 samples examined had at least one plasmid-related PPI and valid aPMNLE data for both all proteins and the chromosomal proteins. This was consistent with the idea that plasmid-related proteins were somewhat separate from chromosomal ones, perhaps at the PPI network periphery, and that plasmid-related and chromosomal proteins had fewer connections with each group.

154 samples (5.2%) had a chromosomal aPMNLE that was >14% (two standard deviations) lower than the rate for all proteins (Table S6), and this group of 153 had proportionally lower chromosomal aPMNLE values when compared to the other 2,777 aPMNLE-stable samples (0.410 vs 0.534, t-test p=1.8e-57) (Figure S20). Furthermore, if the plasmid-related genes were excluded across all samples, the number of PPI network connected components grew by 92% for the 2,777 aPMNLE-stable samples, whereas for these 155 aPMNLE-unstable samples it grew by 330%, indicating larger changes in PPI network structures. This indicated that plasmid-related proteins created more indirect connections in these samples, which consistent with a model where the plasmid-related proteins connected distinct PPI network components that had few PPIs, and so created more indirect connections. It was possible that plasmid-related proteins may possess important PPI signalling roles in these samples.

Three samples (0.1%, Table S7) had lower aPMNLE rates for all proteins than the chromosomal ones alone by two standard deviations, indicating the plasmid-related proteins reduced the numbers of indirect connections, perhaps by bridging gaps in PPI paths. This was consistent with a model where the plasmid- and chromosome-related proteins were mixed together. These samples had limited plasmid-related proteins and yet higher rates of plasmid-related PPIs, indicating that these plasmid proteins may be tightly integrated into the PPI network.

Plasmid-related proteins add both direct and indirect connections

Extending our observations above that plasmid-related proteins generally tend to be peripheral to the PPI network backbone that has more chromosomal proteins, we examined the numbers of indirect connections per protein using the above 2,934 samples. This showed that the samples had 1.7-fold more indirect connections if plasmid-related proteins were included (average 8.09±2.32 indirect connections per protein vs 4.68±2.18, Figure 6). It was possible that plasmid-related proteins were inserted (in part) among modular parts of the chromosomal PPI network backbone. In support of this, the number of proteins per PPI network connected component was 2.3-fold higher when plasmid-related proteins were included (average 179.1±160.9 proteins per connected component vs 78.4±49.6, Figure 6).

We assessed the numbers of indirect connections per protein and the numbers of proteins per connected component for all proteins versus chromosomal ones. This showed that this rate of indirect connections was more strongly associated with the chromosomal pattern than that for all proteins in relation to the numbers of: PPIs (r=0.75 vs 0.43), plasmid-related PPIs (r=0.73 vs 0.48), chromosome-related PPIs (r=0.57 vs 0.08), and proteins per sample (r=0.55 vs 0.06) (Figure S21). Likewise, the numbers of proteins per connected component was more strongly with the chromosomal set than for all proteins in relation to the numbers of: PPIs (r=0.52 vs 0.16), plasmid-related PPIs (r=0.50 vs 0.19), chromosome-related PPIs (r=0.54 vs 0.07), and proteins per sample (r=0.42 vs 0.00) (Figure S22). This highlighted that when there were more PPIs and proteins present, for chromosomal proteins the number of indirect connections increased and the number of connected components in the network decreased. In comparison, for all proteins the relative increase was smaller, perhaps because the initial rate of indirect connections was already high and *vice versa* the starting number of connected components was low. This meant that that addition of a plasmid with its proteins and PPIs to an existing chromosomal PPI network would increase the number of indirect connections and decrease the number of disconnected PPI network components, and that this would scale with the number of plasmid-related PPIs, independent of the total number of proteins involved and the initial connectivity of the chromosomal proteins.

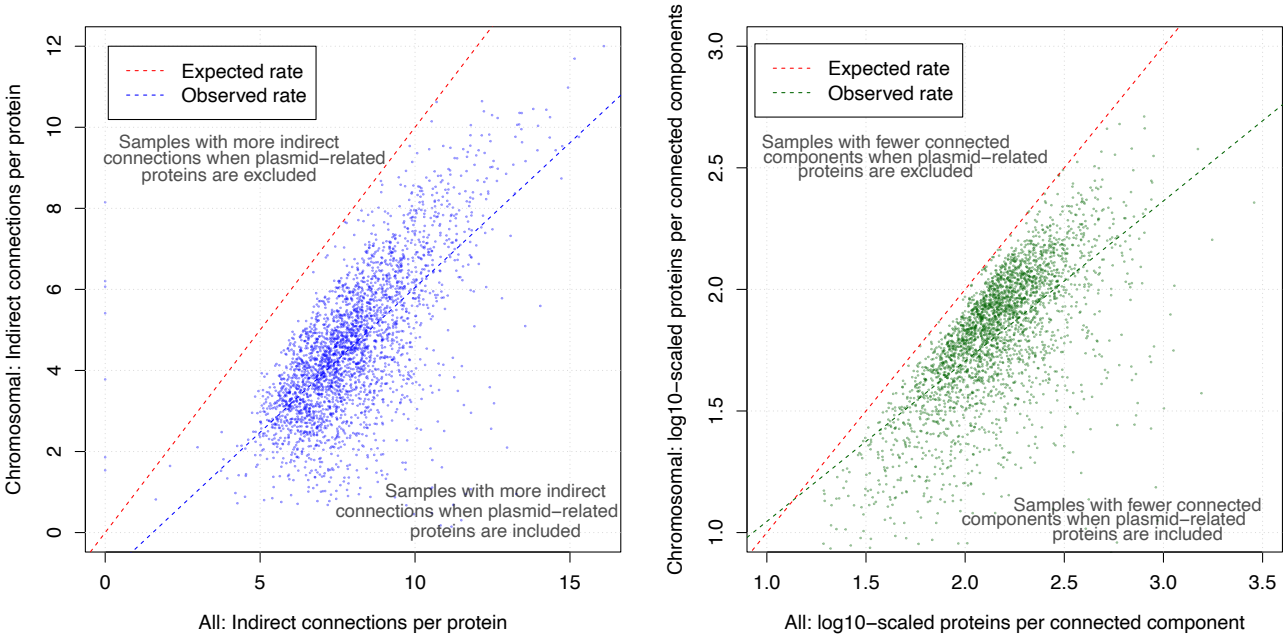


Figure 6. The relative rate for all proteins (x-axis) and for chromosomal proteins (y-axis) for the (left) numbers of indirect connections per proteins and (right) the log10-scaled proteins per connected component. If the numbers of indirect connections per protein was independent of the presence or absence of plasmid-related proteins, the 2,934 samples (blue points) would show a positive linear trend (red line), but instead showed a less positive pattern (blue line). Similarly, if the numbers of proteins per connected component in the PPI network was independent of the presence or absence of plasmid-related proteins, the 2,934 samples (green points) would show a positive linear trend (red line), but instead showed a less positive pattern (green line). This indicated that plasmid-related proteins increased the number of indirect connections and decreased the number of connected components.

Discussion

AMR is a dominant global health concern (Antimicrobial Resistance Collaborators 2022) that is caused by antibiotics in the environment and exacerbated by plasmid-driven AMR gene HGT (Crofts et al 2017, Yao et al 2022). HGT has been the major process shaping bacterial genome structure in *E. coli*, more so than gene duplication (Lercher & Pál 2008, Price et al 2008). Mobile plasmid-encoded genes can endow host cells with new phenotypes, including those linked to changed adhesion, AMR and virulence (Lipworth et al 2022), and their fitness effects are associated with specific genetic, regulatory and protein interactions (Hall et al 2017). To further our understanding of how new plasmid-related proteins can affect the structure and signalling bacterial PPI networks, we compared the PPI profiles of plasmid-linked and chromosomal proteins across 4,419 representative samples of all well-characterised bacteria using data from StringDB.

We found that the number of proteins per sample was correlated with the number of PPIs, as anticipated. This allowed further exploration to compare plasmid-related proteins with chromosomal ones. Most plasmid-related proteins were rare across bacteria, and they composed an average of 10.5% of the total number of proteins per sample. We defined plasmid-related PPIs as PPIs where at least one of the interacting proteins had previously been observed on a contig labelled as a plasmid, separate from chromosomal PPIs where neither protein was plasmid-linked. 28% of samples had zero plasmid-related PPIs. In spite of the limited overall number of plasmid-related proteins, most (>90%) PPIs involved one plasmid-related protein in the samples that had plasmid-related PPIs: 6% of PPIs had only two chromosomal proteins, and <3% had only two plasmid-related proteins.

The plasmid-related genes covaried with one another depending on the genetic background of the host cell such that the clustering the samples based on their plasmid-related gene similarity identified 27

different groups of covarying plasmid-related genes and bacterial samples. This that reflected known taxonomical differences, consistent with the assertion that genetically similar hosts are more able to tolerate HGT from one another (Wellner et al 2007, Price et al 2008, Cohen et al 2011, Andam & Gogarten 2011, Forsberg et al 2014, Soucy et al 2015, Lipworth et al 2022). This matters for the control of AMR gene spread because chromosomal changes associated with one plasmid may endow cells with the capacity to cooperate with other related plasmids more effectively (Loftie-Eaton et al 2017, Jordt et al 2020). In contrast, the PPI rates per protein per sample showed less taxonomical resolution power, positing that the numbers of interactions per protein is typically a less conserved, more dynamical feature.

Plasmid-related proteins tend to be at the periphery of a PPI network (Lercher & Pál 2008, Price et al 2008) because plasmids seldom encode essential genes (Wein et al 2021), though gene essentiality depends on the host (Rousset et al 2021). Using the aPMNLE as a measure of indirect connections, we found most samples (>94%) were PMNLE-stable such that the presence of plasmid-related proteins and their associated plasmid-related PPIs had no substantial changes on PPI networks. In contrast, 5% of samples were PMNLE-unstable: the absence of plasmid-related proteins meant their aPMNLE values dropped considerably. For these PMNLE-unstable samples, excluding the plasmid-related proteins increased the number of connected components by 329%, far above the average increase of 92%, implying that their plasmid-related proteins connected discrete PPI networks hubs of chromosomal proteins. Thus, plasmid-related proteins in these samples may have key signalling roles connecting modular hubs of chromosomal proteins, and potentially may be more likely to be transposed to the chromosome as a consequence of their central PPI network context.

The complexity hypothesis proposed by Jain et al 1999 asserts that HGT-related proteins have fewer PPIs and to have operational cellular processes and not informational ones like transcription or translation (Rivera et al 1998, Jain et al 1999, Jain et al 2002). Although it has been asserted that proteins with many PPIs are less likely to undergo HGT (Wellner et al 2007, Price et al 2008, Cohen et al 2011), we found that plasmid-linked proteins had more PPIs, with the prior assumption that plasmid-encoded genes undergo more HGT. This was consistent with our observation that the vast majority of PPIs were plasmid-related, even though an average of 10.5% of proteins per sample were plasmid-related. Nuances to the complexity hypothesis have been found before: firstly, the independence of HGT rates with the tendency to function in a protein complex (Wellner et al 2007); secondly, the effect of cellular component state, function complexity and function conservation on adaptive evolution potential (Aris-Brosou 2005); thirdly, HGT-linked genes have higher numbers of regulatory factors (Price et al 2008); and fourthly, PPI connectivity has a large effect for proteins with a similar function (Cohen et al 2011). Notably, many previous studies have focused on conserved orthologs, whose genomic contexts and features may differ from genes that are more mobile. Consequently, operational versus informational processes, complex protein functions, an intracellular location, an ancient essential function, and host-recipient genetic distance may be more dominant factors in determining PPI rates than HGT opportunities. An important caveat is that the general trends we observed likely do not hold for all genera, families or orders.

Our findings were limited by a number of issues. Firstly, our analysis was based on gene name similarity with the assumption that the redundancy in the annotated plasmid gene names would help alleviate the imprecision in gene naming systems, but this did not mitigate fully the limitation that some genes were likely misclassified as chromosomal or plasmid-encoded. Secondly, many plasmids and bacterial samples are not annotated consistently, so many gene name matches were undoubtedly missed. Thirdly, we assumed that previous studies' labelling of large contigs as plasmids was correct: we made this (and all other) data accessible so that revisions to plasmid annotation could be made in light of the high number of genes (>500) in certain plasmids (see Table S1): sequence-based plasmid detection tools could help with this task. Fourthly, we obtained gene name data per sample from StringDB (Szklarczyk et al 2018, Szklarczyk et al 2021), and yet out of 9,172 unique plasmid-associated genes we only detected 7,128 genes, which highlighted imprecision either in the plasmid annotation, the sample annotation or (most likely) both. Fifthly, we did not perform sequence

alignment nor read mapping to confirm at the sequence level gene matches, which was not computationally tractable for this entire dataset: this bioresource acts as a step towards that goal. Future work is needed to improve plasmid annotation, genome annotation, contig classification as plasmid or chromosomal and conduct gene comparisons at the sequence level. Future investigations of the effects of plasmid-related and chromosomal proteins on PPI networks should focus on specific samples that have known plasmid compositions.

Author Contributions

TD – Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualisation, Project administration. AR – Methodology, Software, Formal analysis, Writing - Review & Editing.

Acknowledgements

The authors thanks Min Jie Lee for certain R code contributions.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

Agyekum A, Fajardo-Lubián A, Ansong D, Partridge SR, Agbenyega T, Iredell JR. 2016. blaCTX-M-15 carried by IncF-type plasmids is the dominant ESBL gene in Escherichia coli and Klebsiella pneumoniae at a hospital in Ghana. Diagn Microbiol Infect Dis. 84(4):328-33. doi: 10.1016/j.diagmicrobio.2015.12.010

Ahmer BM, et al. 1999. The virulence plasmid of Salmonella typhimurium is self-transmissible. J. Bacteriol. 181, 1364–1368

Alderliesten JB, Duxbury SJN, Zwart MP, de Visser JAGM, Stegeman A, Fischer EAJ. 2020. Effect of donor-recipient relatedness on the plasmid conjugation frequency: a meta-analysis. BMC Microbiol. 20(1):135. doi: 10.1186/s12866-020-01825-4

Alonso-Del Valle A, et al. 2021. Variability of plasmid fitness effects contributes to plasmid persistence in bacterial communities. Nat Commun. 12(1):2653. doi: 10.1038/s41467-021-22849-y

Anda, M. et al. 2015. Bacterial clade with the ribosomal RNA operon on a small plasmid rather than the chromosome. Proc. Natl Acad. Sci. USA 112, 14343–14347.

Andam CP, Gogarten JP. 2011. Biased gene transfer in microbial evolution. Nat. Rev. Microbiol. 9, 543–555

Antimicrobial Resistance Collaborators. 2022. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet doi: 10.1016/S0140-6736(21)02724-0

Aris-Brosou S. 2005. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. Mol Biol Evol. 22(2):200-9. doi: 10.1093/molbev/msi00

Babu M, et al. 2006. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. Journal of molecular biology 358:614-633

Baltrus DA. 2013. Exploring the costs of horizontal gene transfer. Trends Ecol Evol. 28(8):489-95. doi: 10.1016/j.tree.2013.04.002

Bevan ER, Powell MJ, Toleman MA, Thomas CM, Piddock LJV, Hawkey PM. 2021. Molecular characterization of plasmids encoding blaCTX-M from faecal Escherichia coli in travellers returning to the UK from South Asia. J Hosp Infect. 114:134-143. doi: 10.1016/j.jhin.2021.03.030

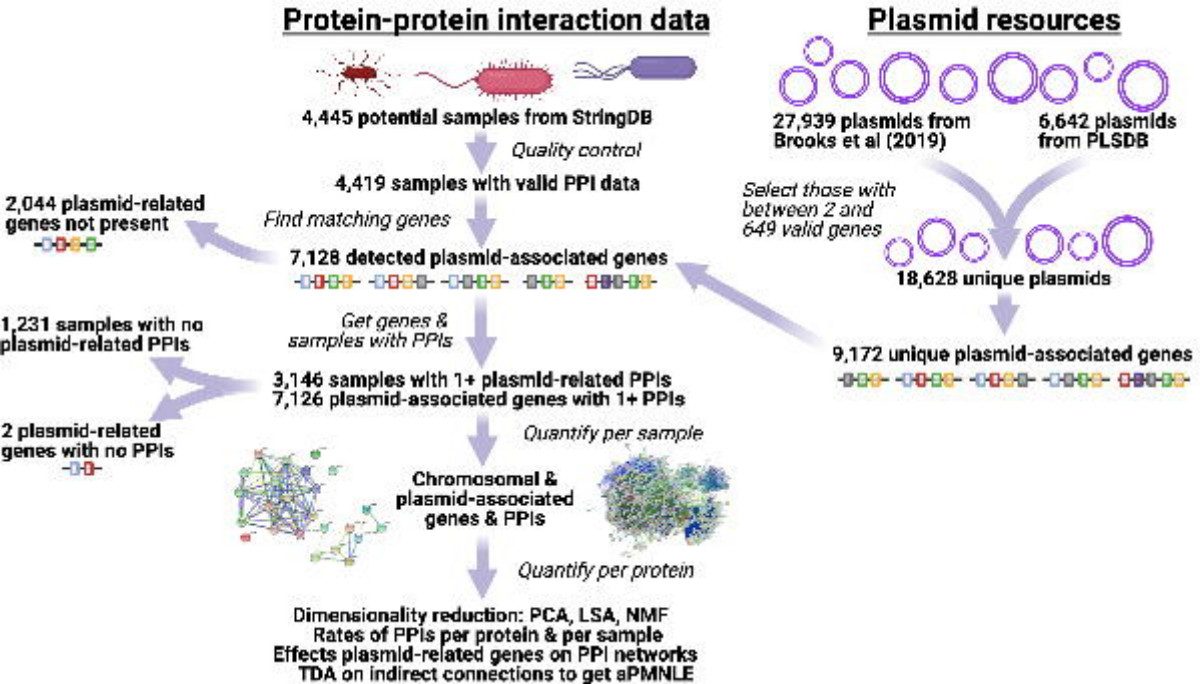
Brinkmann H, et al. 2018. Horizontal operon transfer, plasmids, and the evolution of photosynthesis in Rhodobacteraceae. ISME J. 12, 1994–2010.

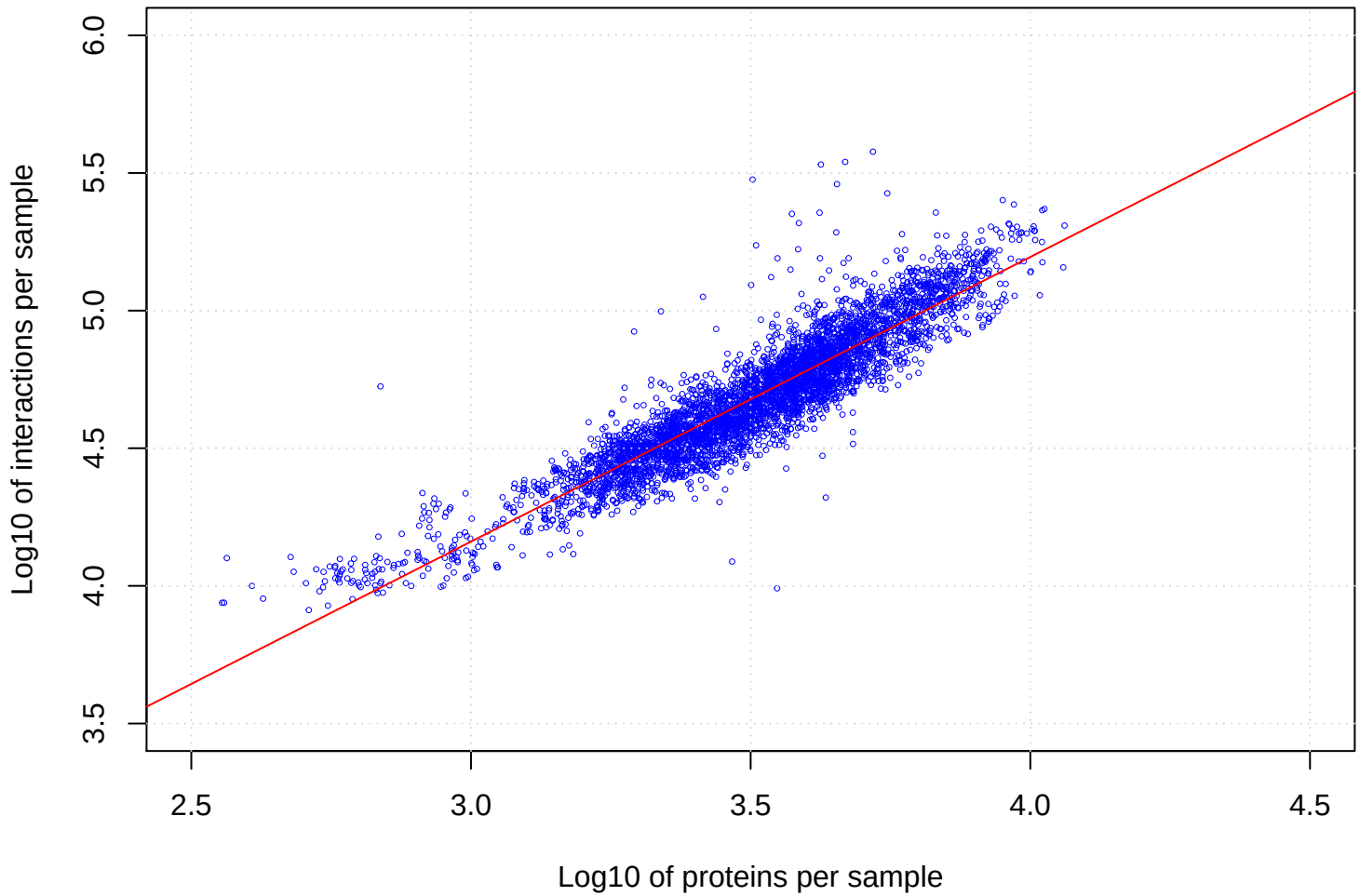
Brito IL, et al. 2016. Mobile genes in the human microbiome are structured from global to individual scales. Nature. 2016 535(7612):435–439
Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. 2019. The Ecology and Evolution of Pangenomes. Curr Biol. 29(20):R1094-R1103. doi: 10.1016/j.cub.2019.08.012
Carattoli, A. 2009. Resistance plasmid families in Enterobacteriaceae. Antimicrob. Agents Chemother. 53, 2227–2238.
Chen H. 2022. VennDiagram: Generate High-Resolution Venn and Euler Plots. R package version 1.7.3. https://CRAN.R-project.org/package=VennDiagram
Coelho LP, et al. 2021. Towards the biogeography of prokaryotic genes. Nature doi: https://doi.org/10.1038/s41586-021-04233-4
Cohen O, et al. 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol. Biol. Evol. 28, 1481–1489
Cohen O, Gophna U, Pupko T. 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol. 28(4):1481-9. doi: 10.1093/molbev/msq333
Crofts TS, Gasparrini AJ, Dantas G. 2017. Next-generation approaches to understand and combat the antibiotic resistome. Nat Rev Microbiol. 15(7):422-434. doi: 10.1038/nrmicro.2017.28
De Gelder L, Ponciano JM, Joyce P, Top EM. 2007. Stability of a promiscuous plasmid in different hosts: no guarantee for a long-term relationship. Microbiology (Reading). 153(Pt 2):452-463. doi: 10.1099/mic.0.2006/001784-0
Decano AG, Downing T. 2019. An Escherichia coli ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. Scientific Reports doi: 10.1038/s41598-019-54004-5
Decano AG, et al. 2019. Complete assembly of Escherichia coli ST131 genomes using long reads demonstrates antibiotic resistance gene variation within diverse plasmid and chromosomal contexts. mSphere doi: 10.1128/mSphere.00130-19
Decano AG, et al. 2020. Plasmids shape the diverse accessory resistomes of Escherichia coli ST131. Access Microbiology doi: 10.1099/acmi.0.000179
Dewar AE, et al. 2021. Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range. Nat Ecol Evol. 5(12):1624-1636. doi: 10.1038/s41559-021-01573-2
Downing T. 2015. Tackling drug resistant infection outbreaks of global pandemic Escherichia coli ST131 using evolutionary and epidemiological genomics. Microorganisms 3(2):236-267.
Dunn S, Carrilero L, Brockhurst M, McNally A. 2021. Limited and Strain-Specific Transcriptional and Growth Responses to Acquisition of a Multidrug Resistance Plasmid in Genetically Diverse Escherichia coli Lineages. mSystems. 6(2):e00083-21. doi: 10.1128/mSystems.00083-21
Forsberg KJ, et al. 2014. Bacterial phylogeny structures soil resistomes across habitats. Nature 509:612-616
Gama JA, Kloos J, Johnsen PJ, Samuelsen Ø. 2020. Host dependent maintenance of a blaNDM-1-encoding plasmid in clinical Escherichia coli isolates. Sci Rep. 10(1):9332. doi: 10.1038/s41598-020-66239-8
Gaujoux R, Seoighe C. 2010. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 11:367
Hall JPJ, Brockhurst MA, Harrison E. 2017. Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. Philos Trans R Soc Lond B Biol Sci. 372(1735):20160424. doi: 10.1098/rstb.2016.0424
Hall JPJ, Wright RCT, Guymer D, Harrison E, Brockhurst MA. 2020. Extremely fast amelioration of plasmid fitness costs by multiple functionally diverse pathways. Microbiology (Reading). 166(1):56-62. doi: 10.1099/mic.0.000862
Harrison E, Brockhurst MA. 2012. Plasmid-mediated horizontal gene transfer is a coevolutionary process. Trends Microbiol. 2012 20(6):262-7. doi: 10.1016/j.tim.2012.04.003
Harrison E, Guymer D, Spiers AJ, Paterson S, Brockhurst MA. 2015. Parallel compensatory evolution stabilizes plasmids across the parasitism-mutualism continuum. Curr Biol. 25(15):2034-9. doi: 10.1016/j.cub.2015.06.024

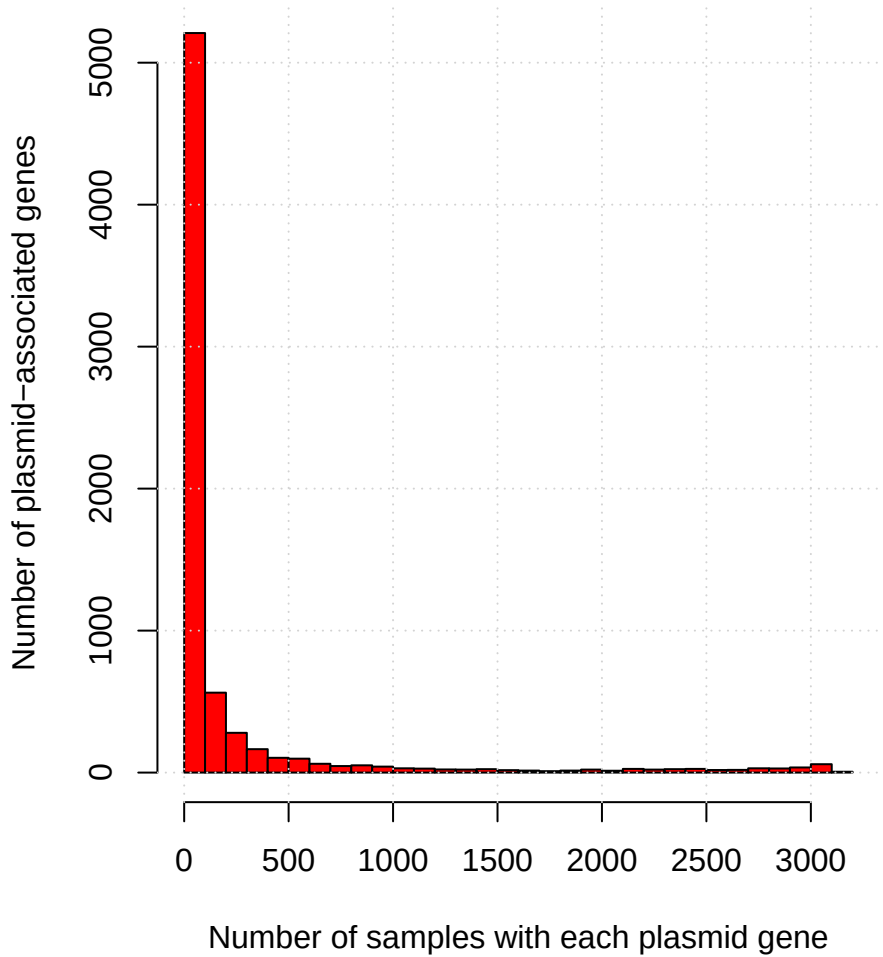
Ho J, et al. 2020. Systematic review of human gut resistome studies revealed variable definitions and approaches. Gut Microbes. 12(1):1700755. doi: 10.1080/19490976.2019.1700755
Huang W, Wang G, Sebra R, Zhuge J, Yin C, Aguero-Rosenfeld ME, Schuetz AN, Dimitrova N, Fallon JT. 2017. Emergence and Evolution of Multidrug-Resistant Klebsiella pneumoniae with both blaKPC and blaCTX-M Integrated in the Chromosome. Antimicrob Agents Chemother. 61(7):e00076-17. doi: 10.1128/AAC.00076-17
Huber VJ, et al. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. W. Huber, V.J. Nature Methods 12:115.
Ihaka R, Gentleman R. 1996. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics. 1996 (5)3:299-314.
Irrgang A, et al. 2017. CTX-M-15-Producing E. coli Isolates from Food Products in Germany Are Mainly Associated with an IncF-Type Plasmid and Belong to Two Predominant Clonal E. coli Lineages. Front Microbiol. 8:2318. doi: 10.3389/fmicb.2017.02318
Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A. 96(7):3801–3806.
Jain R, Rivera MC, Moore JE, Lake JA. 2002. Horizontal gene transfer in microbial genome evolution. Theor Popul Biol. 61(4):489-95. doi: 10.1006/tpbi.2002.1596
Jordt H, Stalder T, Kosterlitz O, Ponciano JM, Top EM, Kerr B. 2020. Coevolution of host-plasmid pairs facilitates the emergence of novel multidrug resistance. Nat Ecol Evol. 4(6):863-869. doi: 10.1038/s41559-020-1170-1
Kamil Slowikowski. 2021. ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.9.1. https://CRAN.R-project.org/package=ggrepel
Kassambara A, Mundt F. 2020. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra
Kottara A, Hall JPJ, Harrison E, Brockhurst MA. 2018. Variable plasmid fitness effects and mobile genetic element dynamics across Pseudomonas species. FEMS Microbiol Ecol. 94(1):fix172. doi: 10.1093/femsec/fix172
Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791.
Lee MC, Marx CJ. 2012. Repeated, selection-driven genome reduction of accessory genes in experimental populations. PLoS Genet. 8(5):e1002651. doi: 10.1371/journal.pgen.1002651
Lehtinen S, et al. 2021. Evolutionary mechanisms that determine which bacterial genes are carried on plasmids. Evol. Lett. 5, 290–301
Lercher MJ, Pál C. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol. Biol. Evol. 25:559
Lipworth S, et al. 2022. The mobilome associated with Gram-negative bloodstream infections: A large-scale observational hybrid sequencing based study. MedRxiv doi: https://doi.org/10.1101/2022.04.03.22273290
Loftie-Eaton W, Bashford K, Quinn H, Dong K, Millstein J, Hunter S, Thomason MK, Merrikh H, Ponciano JM, Top EM. 2017. Compensatory mutations improve general permissiveness to antibiotic resistance plasmids. Nat Ecol Evol. 1(9):1354-1363. doi: 10.1038/s41559-017-0243-2
Ludden C, Decano AG, et al. 2020. Genomic surveillance of Escherichia coli ST131 identifies local expansion and serial replacement of subclones. Microbial Genomics doi: https://doi.org/10.1099/mgen.0.000352
Manzano-Marin A, et al. 2020. Serial horizontal transfer of vitamin-biosynthetic genes enables the establishment of new nutritional symbionts in aphids' di-symbiotic systems. ISME J. 14, 259–273.
Modi RI, Wilke CM, Rosenzweig RF, Adams J. 1991. Plasmid macro-evolution: selection of deletions during adaptation in a nutrient-limited environment. Genetica. 84(3):195-202. doi: 10.1007/BF00127247
Müller K, Wickham H. 2021. tibble: Simple Data Frames. R package version 3.1.6. https://CRAN.R-project.org/package=tibble

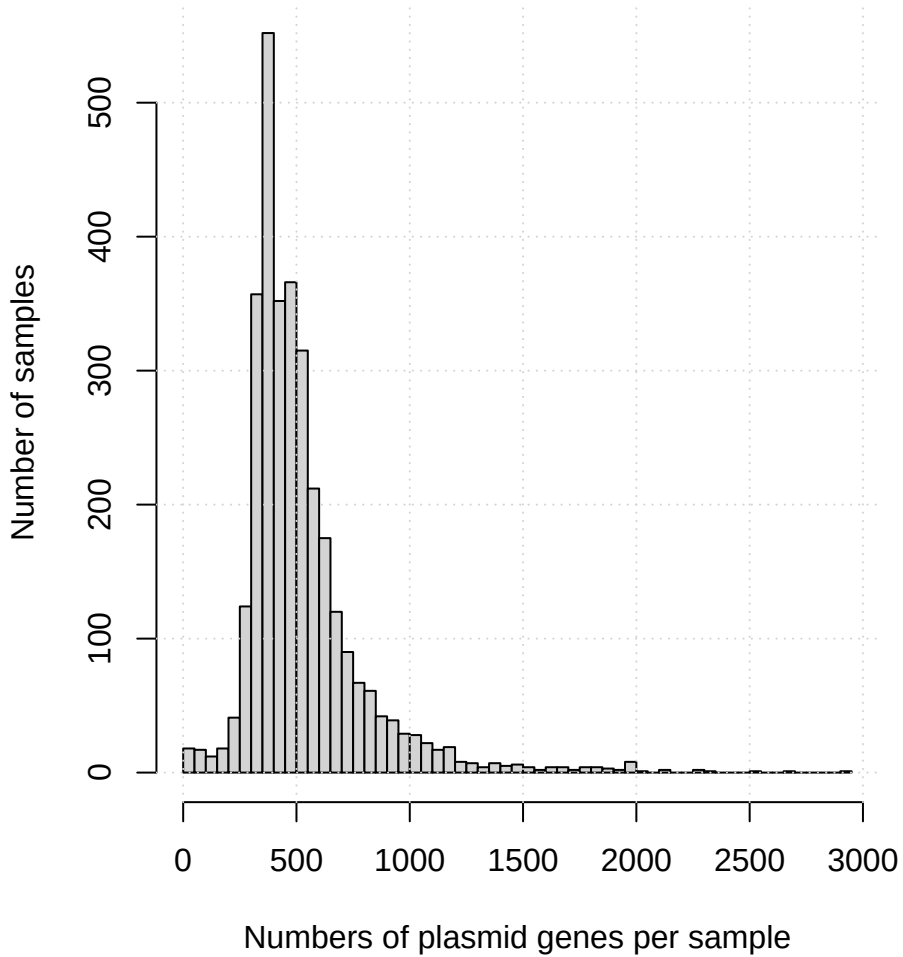
Nakamura Y, et al. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat. Genet. 36:760–76
Niehus R, et al. 2015. Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nat. Commun. 6, 8924
Norman A, Hansen LH, Sørensen SJ. 2009. Conjugative plasmids: vessels of the communal gene pool. Philos Trans R Soc Lond B Biol Sci. 364(1527):2275-89. doi: 10.1098/rstb.2009.0037
Novick A, Doolittle WF. 2020. Horizontal persistence and the complexity hypothesis. Biology & Philosophy 35:2.
Pilla G, Tang CM. 2018. Going around in circles: Virulence plasmids in enteric pathogens. Nat. Rev. Microbiol. 16, 484–495.
Porse A, et al. 2018. Biochemical mechanisms determine the functional compatibility of heterologous genes. Nat Commun. 9(1):522. doi:10.1038/s41467-018-02944-3
Porse A, Schønning K, Munck C, Sommer MO. 2016. Survival and Evolution of a Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts. Mol Biol Evol. 33(11):2860-2873. doi: 10.1093/molbev/msw163
Price MN, Dehal PS, Arkin AP. 2008. Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. Genome Biol. 9(1):R4. doi: 10.1186/gb-2008-9-1-r4
Puigbo P, Wolf YI, Koonin EV. 2010. The tree and net components of prokaryote evolution. Genome Biol. Evol. 2, 745–756
R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
Rahm A. HomologyLive; 2019
Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes, Proc Natl Acad Sci U S A, vol. 95 (pg. 6239-6244)
Rousset F, et al. 2021. The impact of genetic diversity on gene essentiality within the Escherichia coli species. Nat Microbiol. 6(3):301-312. doi: 10.1038/s41564-020-00839-y
Rusakovica J, et al. 2014. Probabilistic latent semantic analysis applied to whole bacterial genomes identifies common genomic features. J Integr Bioinform. 11(2):243. doi: 10.2390/biecoll-jib-2014-243
San Millan A, MacLean RC. Fitness Costs of Plasmids: a Limit to Plasmid Transmission. Microbiol Spectr. 2017 5(5). doi: 10.1128/microbiolspec.MTBP-0016-2017
San Millan A, Peña-Miller R, Toll-Riera M, Halbert ZV, McLean AR, Cooper BS, MacLean RC. 2014. Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. Nat Commun. 5:5208. doi: 10.1038/ncomms6208
San Millan A. 2018. Evolution of plasmid-mediated antibiotic resistance in the clinical context. Trends Microbiol. 26, 978–985
Shawa M, Furuta Y, Mulenga G, Mubanga M, Mulenga E, Zorigt T, Kaile C, Simbotwe M, Paudel A, Hang'ombe B, Higashi H. 2021. Novel chromosomal insertions of ISEcp1-blaCTX-M-15 and diverse antimicrobial resistance genes in Zambian clinical isolates of Enterobacter cloacae and Escherichia coli. Antimicrob Resist Infect Control. 10(1):79. doi: 10.1186/s13756-021-00941-8
Sheppard RJ, Beddis AE, Barraclough TG. 2020. The role of hosts, plasmids and environment in determining plasmid transfer rates: A meta-analysis. Plasmid. 108:102489. doi: 10.1016/j.plasmid.2020.102489
Smillie CS, et al. Ecology drives a global network of gene exchange connecting the human microbiome. Nature. 2011 480(7376):241–4
Soucy SM, et al. 2015. Horizontal gene transfer: building the web of life. Nat. Rev. Genet. 16, 472–482
Stalder T, Rogers LM, Renfrow C, Yano H, Smith Z, Top EM. Emerging patterns of plasmid-host coevolution that stabilize antibiotic resistance. Sci Rep. 2017 7(1):4853. doi: 10.1038/s41598-017-04662-0
Stasiak G. et al. 2014. Functional relationships between plasmids and their significance for metabolism and symbiotic performance of Rhizobium leguminosarum bv. trifolii. J. Appl. Genet. 55, 515–527

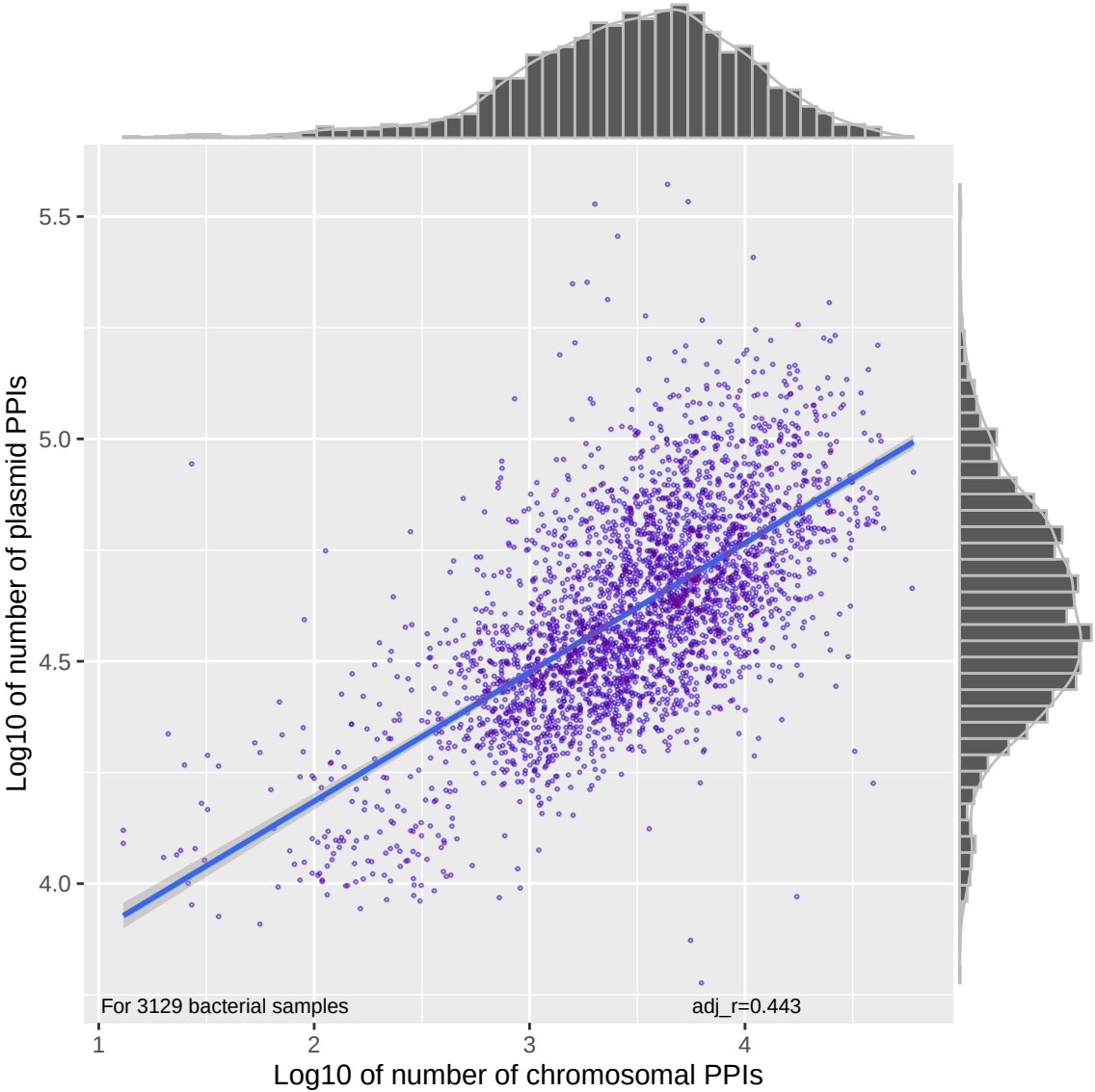
Szklarczyk D, et al. 2018. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research 47.D1: D607-D613.
Szklarczyk D, et al. 2021. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 49(D1):D605-D612. doi: 10.1093/nar/gkaa1074
The LinBox Group Exact linear algebra over the Integers and finite rings, version 1.1.6; 2008
Touchon M, et al. 2020. Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. PLoS Genet. 16(6):e1008866. doi: 10.1371/journal.pgen.1008866
Vietoris L. Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. Mathematische Annalen 1927; 97:454–472
Wang, X. et al. 2018. Comparative symbiotic plasmid analysis indicates that symbiosis gene ancestor type affects plasmid genetic evolution. Lett. Appl. Microbiol. 67, 22–31.
Wein T, et al. Essential gene acquisition destabilizes plasmid inheritance. PLoS Genet. 2021 17(7):e1009656. doi: 10.1371/journal.pgen.1009656
Wellner A, Lurie MN, Gophna U. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. Genome Biol. 2007;8(8):R156. doi: 10.1186/gb-2007-8-8-r156
Wickham H, Bryan J. 2022. readxl: Read Excel Files. R package version 1.4.0. https://CRAN.R-project.org/package=readxl
Wickham H, et al. 2019. Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
Wickham H, François R, Henry L, Müller K. 2022. dplyr: A Grammar of Data Manipulation. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr
Wickham H, Girlich M. 2022. tidyr: Tidy Messy Data. R package version 1.2.0. https://CRAN.R-project.org/package=tidyr
Wickham H, Hester J, Bryan J. 2022. readr: Read Rectangular Text Data. R package version 2.1.2. https://CRAN.R-project.org/package=readr
Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
Wickham H. 2021. forcats: Tools for Working with Categorical Variables (Factors). R package version 0.5.1. https://CRAN.R-project.org/package=forcats
Wickham H. 2019. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr
Winter DJ. 2017. rentrez: an R package for the NCBI eUtils API The R Journal 9(2):520-526
Yao Y, et al. Intra- and interpopulation transposition of mobile genetic elements driven by antibiotic selection. Nat Ecol Evol. 2022 doi: 10.1038/s41559-022-01705-2
Yoon EJ, Gwon B, Liu C, Kim D, Won D, Park SG, Choi JR, Jeong SH. 2020. Beneficial Chromosomal Integration of the Genes for CTX-M Extended-Spectrum β-Lactamase in Klebsiella pneumoniae for Stable Propagation. mSystems. 5(5):e00459-20. doi: 10.1128/mSystems.00459-20
Zhou H, et al. 2021. Functions predict horizontal gene transfer and the emergence of antibiotic resistance. Sci Adv. 7(43):eabj5056. doi: 10.1126/sciadv.abj5056.

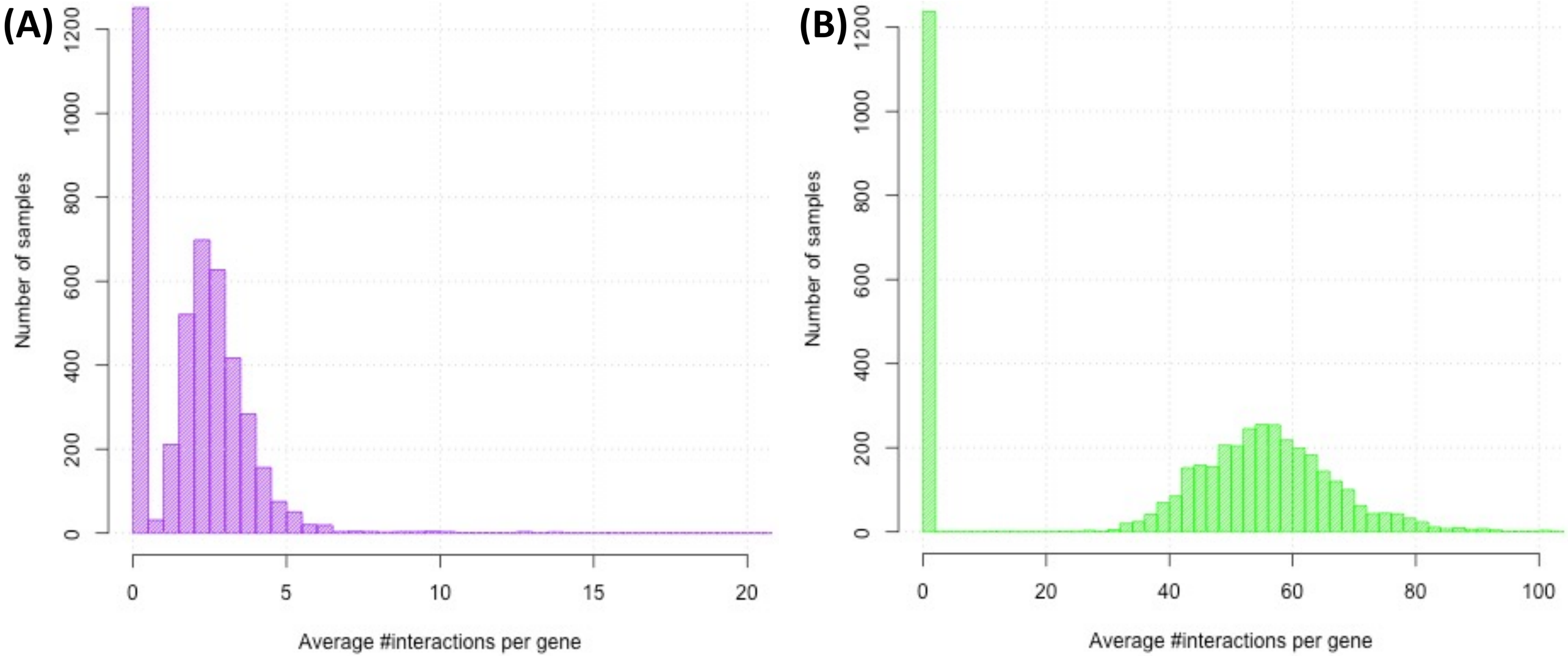

Protein-protein interaction data
Plasmid resources
4,445 potential samples from StringDB
Quality control
4,419 samples with valid PPI data
Find matching genes
2,044 plasmid-related genes not present
7,128 detected plasmid-associated genes
Get genes & samples with PPIs
1,231 samples with no plasmid-related PPIs
3,146 samples with 1+ plasmid-related PPIs
7,126 plasmid-associated genes with 1+ PPIs
2 plasmid-related genes with no PPIs
Quantify per sample
Chromosomal & plasmid-associated genes & PPIs
Quantify per protein
Dimensionality reduction: PCA, LSA, NMF
Rates of PPIs per protein & per sample
Effects plasmid-related genes on PPI networks
TDA on indirect connections to get aPMNLE
27,939 plasmids from Brooks et al (2019)
6,642 plasmids from PLSDB
Select those with between 2 and 649 valid genes
18,628 unique plasmids
9,172 unique plasmid-associated genes

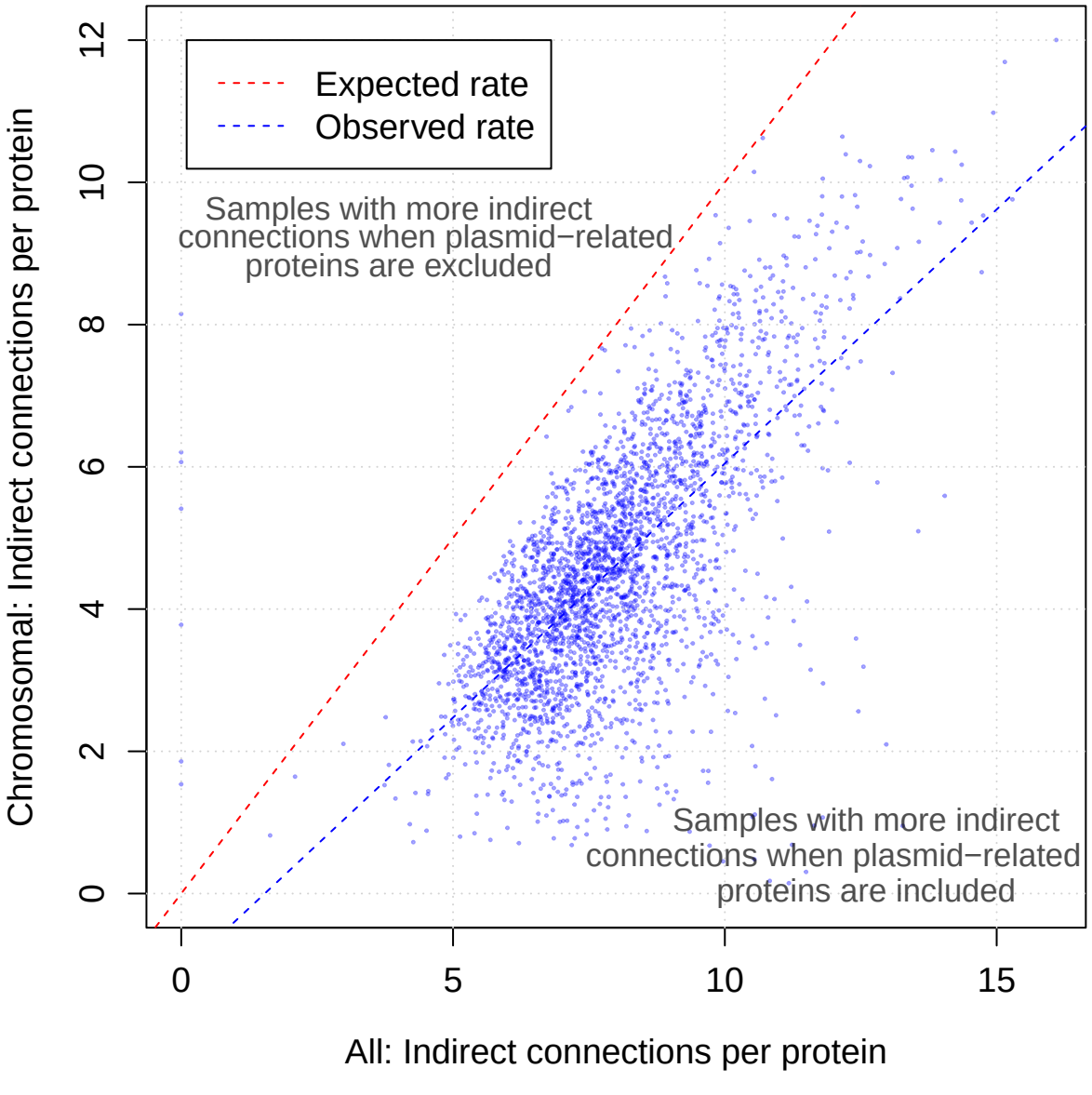

Log10 of interactions per sample
6.0
5.5
5.0
4.5
4.0
3.5
2.5
3.0
3.5
4.0
4.5
Log10 of proteins per sample

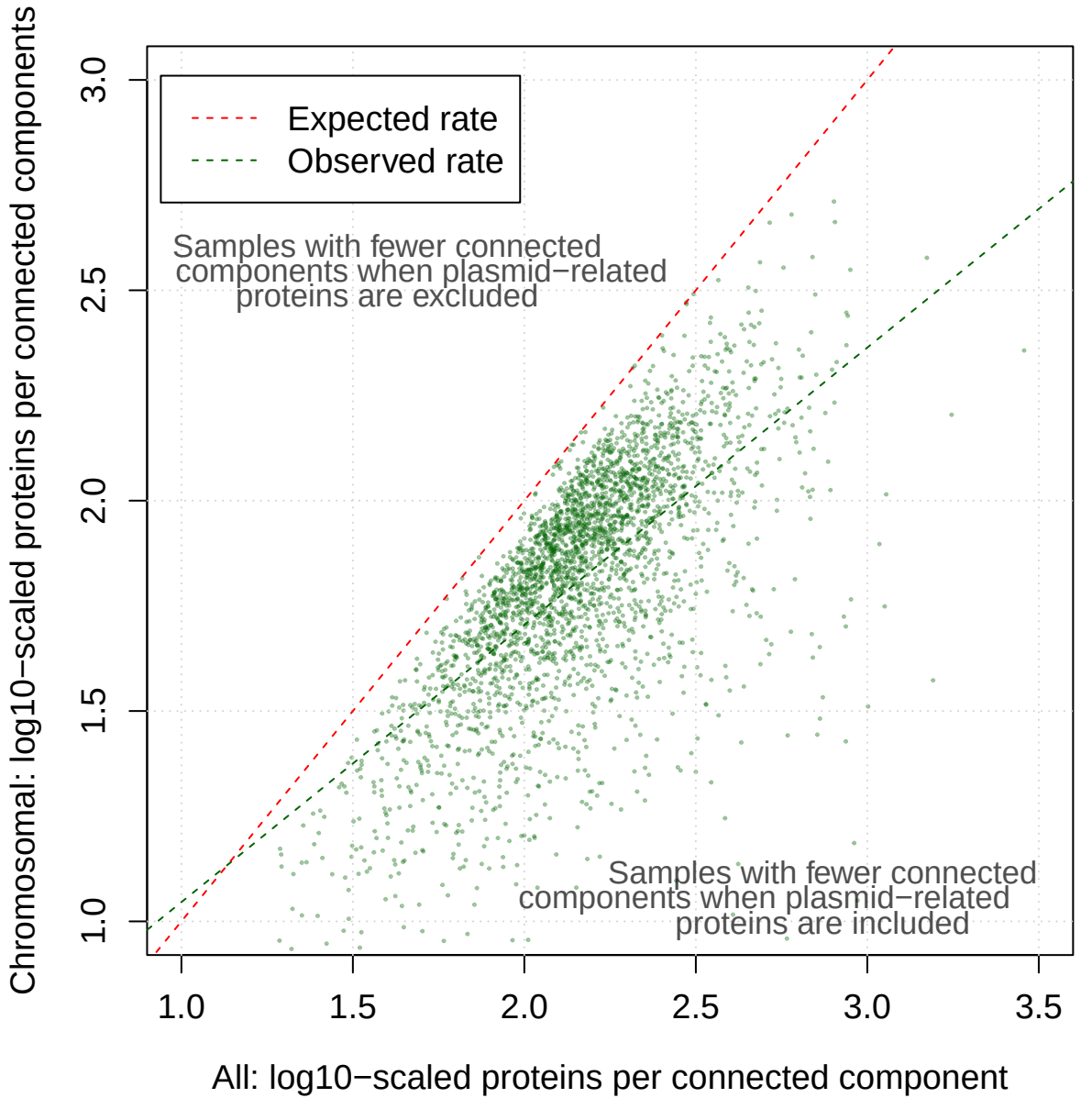

Number of plasmid−associated genes
0
1000
2000
3000
4000
5000
0
500
1000
1500
2000
2500
3000
Number of samples with each plasmid gene


Number of samples
0
100
200
300
400
500
0
500
1000
1500
2000
2500
3000
Numbers of plasmid genes per sample


5.5
5.0
4.5
4.0
Log10 of number of plasmid PPIs
For 3129 bacterial samples
adj_r=0.443
1
2
3
4
Log10 of number of chromosomal PPIs


(A)
Number of samples
0
200
400
600
800
1000
1200
0
5
10
15
20
Average #interactions per gene
(B)
Number of samples
0
200
400
600
800
1000
1200
0
20
40
60
80
100
Average #interactions per gene


Expected rate
Observed rate
Samples with more indirect connections when plasmid−related proteins are excluded
Samples with more indirect connections when plasmid−related proteins are included
Chromosomal: Indirect connections per protein
All: Indirect connections per protein
0
2
4
6
8
10
12
0
5
10
15


Expected rate
Observed rate
Samples with fewer connected components when plasmid−related proteins are excluded
Samples with fewer connected components when plasmid−related proteins are included
Chromosomal: log10−scaled proteins per connected components
All: log10−scaled proteins per connected component
1.0
1.5
2.0
2.5
3.0
1.0
1.5
2.0
2.5
3.0
3.5