scientific reports

OPEN

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

Tim Downing[1,3] ✉ & Alexander Rahm[2]

Plasmids facilitate horizontal gene transfer, which enables the diversification of pathogens into new anatomical and environmental niches, implying that plasmid-encoded genes can cooperate well with chromosomal genes. We hypothesise that such mobile genes are functionally different to chromosomal ones due to this ability to encode proteins performing non-essential functions like antimicrobial resistance and traverse distinct host cells. The effect of plasmid-driven 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-encoded proteins across representative specimens from most bacteria using all available plasmids. We found that plasmid-encoded genes constitute ~ 0.65% of the total number of genes per bacterial sample, and that plasmid genes are preferentially associated with different species but had limited taxonomical power beyond this. Surprisingly, plasmid-encoded proteins had both more 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-encoded proteins had limited overall impact in > 96% of samples. This paper assembled extensive data on plasmid-encoded proteins, their interactions and associations with diverse bacterial specimens that is available for the community to investigate in more detail.

Plasmids are short extrachromosomal DNA elements that are typically circular and with variable copy numbers. Conjugative plasmids possess machinery enabling horizontal gene transfer (HGT) between bacterial cells, a key mechanism of bacterial evolution. Plasmid DNA may also transfer infrequently between cells by transformation, vesicles and phages[1–4]. This allows plasmid-encoded genes to move between bacterial cells in the same niche[5], where these genes may allow new phenotypes[6–9]. Plasmid-encoded genes encode proteins performing a wide range of functions, including antimicrobial resistance (AMR), virulence, metabolism, and symbiosis[10–12], facilitating spread into new environmental niches[13]. Importantly, they depend on the environmental context to be beneficial, such that they can be lost when no longer required[13].

Mobile genes form part of the accessory genome, distinct from chromosomal genes that typically compose the core genome[14], 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[15–23]. In this study, we examined the broad patterns of compatibility across bacteria using the categorisation of plasmid-encoded versus chromosomal.

Plasmids impose a metabolic fitness cost on their host cells[24–26] and yet persist widely thanks to infectious spread, stability mechanisms, and chromosomal compensatory mutations[27]. Plasmids are genetically very diverse, and their conjugation generates additional variation in bacterial AMR gene profiles[28,29]. This high plasmid mobility creates opportunities for plasmid-host co-evolution in permissive environments that allow plasmids to exist in hosts cells for long periods of time[30,31]. Gene encoding products that directly interact with host molecules

[1]School of Biotechnology, Dublin City University, Dublin, Ireland. [2]GAATI Lab, University of French Polynesia, Tahiti, French Polynesia. [3]Present address: The Pirbright Institute, Pirbright, UK. ✉email: tim.downing@dcu.ie

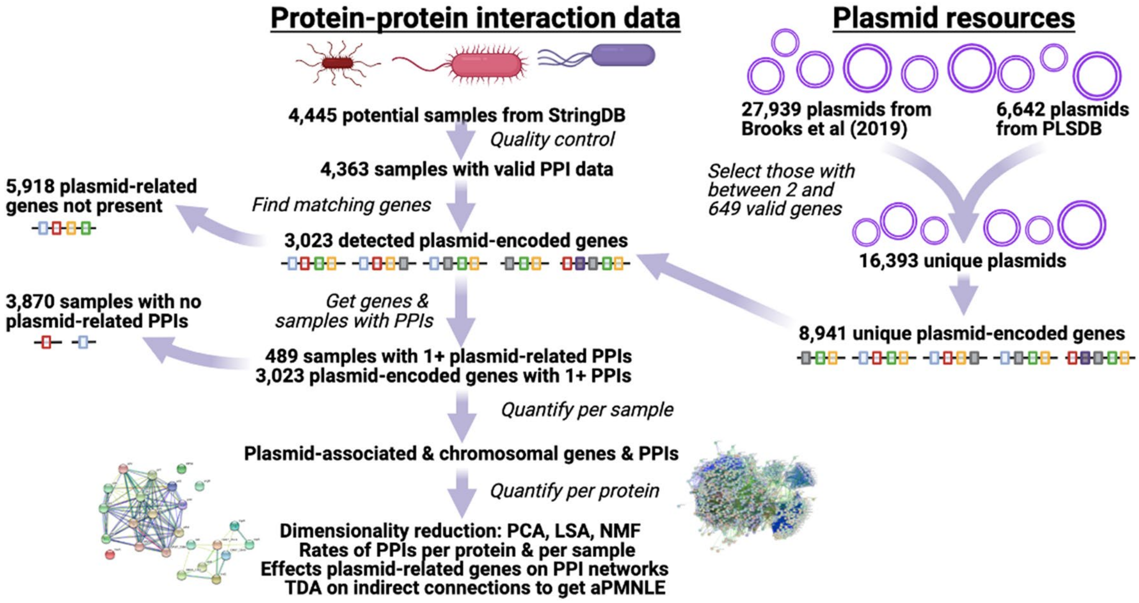
Protein-protein interaction data
Plasmid resources
4,445 potential samples from StringDB
Quality control
4,363 samples with valid PPI data
27,939 plasmids from Brooks et al (2019)
6,642 plasmids from PLSDB
Select those with between 2 and 649 valid genes
16,393 unique plasmids
8,941 unique plasmid-encoded genes
5,918 plasmid-related genes not present
Find matching genes
3,023 detected plasmid-encoded genes
Get genes & samples with PPIs
3,870 samples with no plasmid-related PPIs
489 samples with 1+ plasmid-related PPIs
3,023 plasmid-encoded genes with 1+ PPIs
Quantify per sample
Plasmid-associated & chromosomal 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

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-encoded proteins 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.

may adapt more rapidly in such a context. This matters because a suitable host-recipient genetic background is needed such that the transferred gene can be acquired, integrated, and expressed[32–34]. In addition, donor & recipient cells may be genetically distinct and yet have high HGT rates[35,36], suggesting that additional factors affecting gene and plasmid retention need to be quantified.

HGT depends on the physiological compatibility between the donor and recipient so that a plasmid-borne gene can be expressed and provide a fitness benefit to the recipient cell by effectively interacting with the recipient proteins[6,37]. The complexity hypothesis[38,39] 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[40–42]. HGT rates are higher for rarer genes like those in the accessory genome on plasmids, and lower for core genes[43,44]. Moreover, gene context and physical gene distance are related to network proximity[45] such that accessory genes and newly transferred proteins are likely to be peripheral in PPI networks, including plasmid-encoded ones, whereas the chromosomal proteins should be more central.

Plasmid properties vary extensively based on the host cell[46–52] suggesting that the interaction between the host chromosomal proteins and plasmid-encoded proteins is important. The evidence collated so far implies that chromosomes adapt more readily to plasmids as opposed to vice-versa, usually by single nucleotide polymorphisms (SNPs)[27,53–55], but sometimes also by larger gene deletions[56–58]. This co-evolution of plasmid-encoded and chromosomal proteins suggests that a more systematic approach is required to examine the extent to which these proteins interact across bacterial species, which could potentially provide insights into plasmid-host compatibility.

In this study, we assessed the hypothesis that chromosomal and plasmid-encoded genes have different interaction levels based on the host species, PPI context and PPI connectivity. Using a core dataset of 4363 bacterial samples, we examined PPI network connectivity using PPI data for all chromosomal proteins and for plasmid-encoded ones. Importantly, HGT is not strongly affected by gene sequence composition and features[37], supporting our simple classification of genes as present or absent here. We examined the changes in PPI network structure based on changes in direct and indirect connections when plasmid-encoded proteins enter a host's chromosomal PPI network. We identified distinctive features associated with plasmid-encoded proteins and their PPIs, assessed their covariation, and thus provide a pilot study of plasmid-related PPIs across all bacteria.

Methods

Protein–protein interaction data extraction.

We retrieved PPI information for all available valid bacterial genomes (n = 4445) from STRING database v12[59,60] with R v4.0.1[61] and RStudio v2022.2.3.492[61] via Bioconductor v3.11[62] and packages Rentrez v1.2.2[63], STRINGdb v2.0.1[59,60], and StringR v1.4.0[64] (Fig. 1). These genomes were selected to encompass all available bacterial genomes with possible PPI data. Of the initial 4445 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 = 4419 samples) was processed and collated using Dplyr v1.0.8[65], Forcats v0.5.1[66], Grid v4.4.1, ReadR v2.1.2[67], Readxl v1.4.0[68], Tibble v3.1.6[69], TidyR v1.2.0[70], Tidyverse v1.3.0[71], VennDiagram v1.7.3[72] 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 4419 post-QC samples had 3628.5 ± 1710.0 (mean ± SD) proteins and a mean of 57,105.8 ± 34,698.6 PPIs each. Not all data was accessible for 56 samples among these 4419, leaving 4363 samples. This and subsequent data below were visualised with R packages ggplot2 v3.3.5[73] and ggrepel v0.9.1[74]. All reported p values were corrected for multiple testing using the Benjamini–Hochberg approach in R.

The accuracy of gene names in annotation is a pervasive issue where previous work has shown ~ 46 ± 9% of genes can be functionally assigned[75]. 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 2931 ± 1744 long names each (Figure S1).

Identification of plasmid-encoded genes and interactions across bacteria.

To get all available plasmid-encoded genes, we collated 32,839 plasmids: 27,939 from[76] and 6642 from PLSDB (v2020_11_19)[77] and retrieved gene annotation for each plasmid using Genbankr v1.16.0 (Fig. 1). We focused on 16,383 with between two and 640 annotated genes (< 640 to avoid chromosome-related contigs). We verified that the plasmids were annotated as plasmids, including 12 with lengths > 2 Mb for *Ralstonia solanacearum* and *Rhizobium gallicum*. A gene was defined as plasmid-encoded if it had been found on a plasmid originating in the same species, yielding two categories for genes: plasmid-encoded or chromosomal. These plasmids had a median of 12 unique annotated genes per plasmid and had 8941 unique plasmid-encoded genes across 98,534 unique sample-gene associations (Table S1, see full table FigShare https://doi.org/10.6084/m9.figshare.19525630). Analyses focused on the 288 samples with at least one plasmid gene, and the 5538 plasmid-encoded genes detected in at least one sample here.

Next, we identified the numbers of PPIs per protein per sample. A PPI was defined as plasmid-related if either (or both) protein(s) were found on a plasmid, yielding three categories for proteins: chromosomal, plasmid-related with one plasmid-linked protein in the pair, or exclusively plasmid-related where both protein pairs were plasmid-encoded—in practise the latter category was rare and was not investigated in detail here. The numbers of PPIs in which both proteins were exclusively chromosomal was large (286,364,425) compared to the totals for plasmid-related PPIs (390,592) and PPIs exclusively on plasmid-encoded proteins (46,772). This meant the fraction of PPIs that were plasmid-related was 0.136%, and 0.016% of all PPIs were exclusively plasmid-related, and the remainder chromosomal (99.86%). In the samples with at least one plasmid-related PPI, we used the probabilities of having a plasmid-related or chromosomal protein like allele frequencies to compare to the observed frequencies of plasmid-restricted, plasmid-chromosome and chromosome-restricted PPIs per sample in a manner analogous to Wright's F-statistic[78] as $F = 1 - obs/exp$ where obs was the fraction of plasmid-chromosome PPIs, and exp was twice the product of the frequencies of plasmid-related and chromosomal proteins. This compared the expected probabilities of plasmid-restricted, plasmid-chromosome and chromosome-restricted PPIs against the observed rates to test for structure (or absence thereof) between the plasmid-related and chromosomal proteins.

We examined the taxonomic classifications of the samples across families, orders, classes and phyla using R package taxize v0.9.98[79] to examine their scaled pairwise Euclidean distances of the plasmid presence-absence data in a dendrogram from hierarchical clustering from R packages stats v3.6.2 and dendextend v1.15.2[80]. This identified totals of 100 families, 46 orders, 20 classes and 10 phyla.

Screening for correlated bacterial samples and plasmid-encoded genes.

We examined patterns of covariance across the set of the samples and 3023 plasmid-encoded genes using three approaches. To resolve highly correlated sets of samples, the first approach was to assess genes and sample-gene pairs using principal components analysis (PCA) implemented with R packages stats v3.6.2 and factoextra v1.0.7[81]. The second 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[82]. Here, it applied a classical vector space model and singular value decomposition to determine the samples' plasmid-encoded 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-encoded genes.

The third approach was non-negative matrix factorisation (NMF)[83] to examine the associations of individual plasmid-encoded genes with individual bacterial samples using R packages nmf v0.23.0[84] and Biobase v2.52.0[85]. This was applied to 331 samples with at least three plasmid-encoded genes, and to 1733 plasmid-encoded genes found in at three samples. This data was analysed as binary presence-absence data. The optimal rank was determined by examining changes in the cophenetic correlation coefficient across ranks 4–26 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 six groups (ranks) (cophenetic correlation coefficient maximised at 0.971). We also examined the covariation of samples with the PPI levels per gene for 481 samples and 2363 plasmid-encoded genes where the input data was the number of PPIs per protein per sample using the NMF approach (as outlined above) where each protein had at least 30 PPIs across the sample and each sample had at least ten PPIs in total. 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 seven groups (ranks) was obtained (cophenetic correlation coefficient maximised at 0.9856).

Topological data analysis of the bacterial protein–protein interaction networks.

We developed computationally efficient metrics to investigate the topology of a PPI network, with a focus on the PPI network's indirect connectivity as well as direct PPIs. Constructing the Vietoris-Rips complex[86] 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 (number of indirect connections divided by number of PPIs) varies only moderately when we varied the StringDB combined score threshold, which modified the PPI network topology by reducing the number of PPIs[9]. 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 (number of indirect connections divided by number of 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 and computation time 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).

The PPI networks of the 491 samples with plasmid-encoded proteins were analysed with all proteins, and then with chromosomal ones only. This used a specialised open-source tool[87] that 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[88]) 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 bacterial PPI networks, the purely chromosomal part determined the indirect connectivity, i.e., if the aPMNLE for the full PPI network was similar to the aPMNLE of the chromosomal PPI network, then the influence of the plasmids on the aPMNLE was negligible. 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. The aPMNLE was associated with the numbers of PPIs per sample ($r = 0.42$, $p = 8e{-}182$).

Results

We retrieved PPI data from StringDB to compare the properties of plasmid-encoded versus chromosomal proteins, with a view to understand 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 4363 valid bacterial samples based on their unique annotated proteins on StringDB. As expected, the numbers of proteins and PPIs was highly correlated (adjusted $r = 0.914$, $p = 5.2e{-}15$) (Fig. 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 that the 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-encoded proteins in different samples combinations in more detail.

Distinct patterns of plasmid-encoded gene retention across bacterial samples. We compared these 4363 samples' PPI data with the set of 8941 unique plasmid-linked genes where a gene was defined as plasmid-encoded if it was detected on a plasmid from the same species. 66.2% (5918) of these genes were not detected, yielding 3023 detected genes. 491 samples (11.2%) had plasmid-encoded genes, leaving 88.8% (3872) samples with no such genes detected. In this set of 491, the median number of samples in which a plasmid-encoded gene was found was three (interquartile range four, Fig. 3A), and the median number of plasmid-encoded genes per sample was five, with an interquartile range of 12 (Fig. 3B). The fraction of plasmid-encoded genes per sample was 0.65 ± 2.5% (27 ± 112 out of 3927 ± 1367, mean ± SD) in these 491 (Fig. 3). Six *E. coli* genomes had a median of 1035 plasmid-encoded genes per sample and were the only samples with over 400 plasmid-encoded genes. These results illustrated that most plasmid-encoded genes were rare, with few common ones, and that widely shared plasmid-encoded genes were unusual (Figure S2, see full figure FigShare https://doi.org/10.6084/m9.figshare.19525453).

Plasmid-encoded genes are structured and reflect taxonomic groups. The samples were clustered by similarity based on their binary 3023 unique plasmid-encoded gene data (presence-absence). Taxonomic classifications across the 100 families, 46 orders, 20 classes and 10 phyla in these 491 samples showed that the *Enterobacteriaceae* were the most diverse family ($n = 24$ samples, $Z = 1.49$), and that the *Enterobacterales*

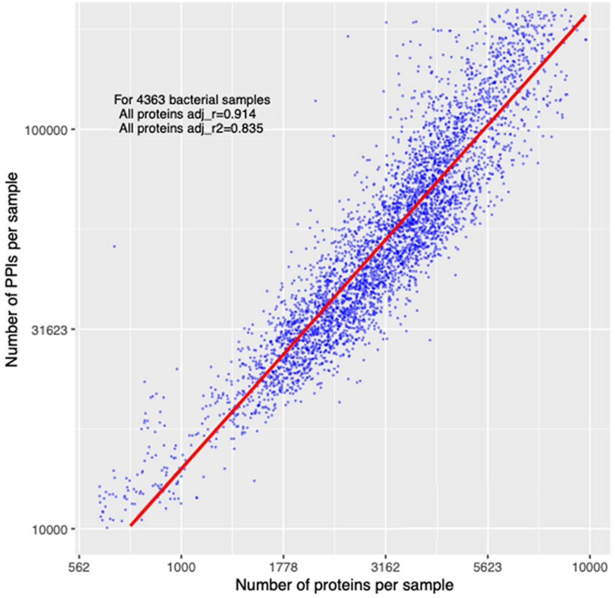

Figure 2. The number of proteins per bacterial genome (x-axis) was highly predictive of the number of PPIs (y-axis) (adjusted $r^2 = 0.835$). The data was for 4363 bacterial samples, each of which is shown by a blue dot, with a linear model line of best fit shown in red.

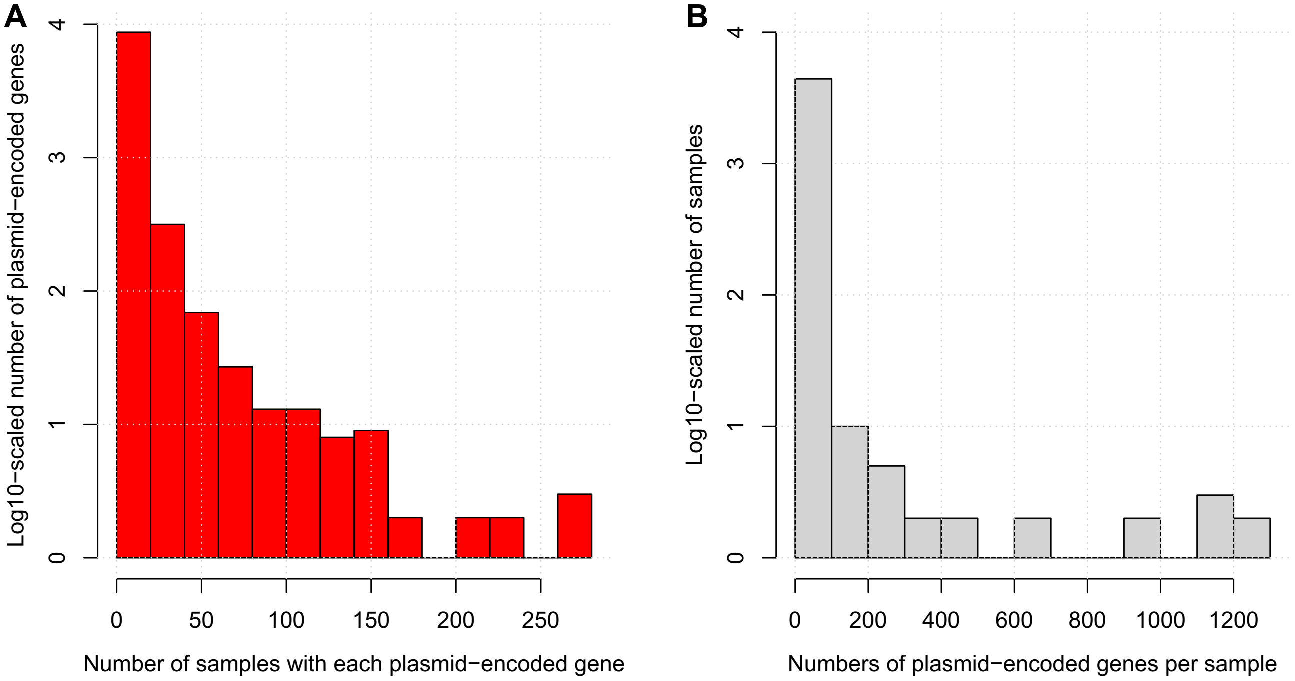

Figure 3. (**A**) The number of samples in which each plasmid-encoded gene (red) was found was generally small with a median of five. (**B**) The number of plasmid-encoded genes per sample (grey) had a median of three, including six *E. coli* samples with a median of 1035 plasmid-encoded genes per sample that represent all samples with > 400 genes. These figures excluded samples with no plasmid-encoded genes.

(n = 57, Z = 0.77) was the most variable order, with limited changes at the class and order levels. Overall, the plasmid-encoded gene profiles per taxonomic group did not yield an informative taxonomic resolution (Figure S3, see full figure FigShare https://doi.org/10.6084/m9.figshare.19525435).

Evidence for this low overall covariation came from PCA of this data where six *E. coli* and three *Ralstonia solanacearum* samples explained most variation across PC1 (34.5%) and PC2 (5.6%) (Figure S4A). These six *E. coli* had far more plasmid-encoded genes compared to the sample with the next highest number (*Ralstonia solanacearum* PSI07 with 347), which in turn exceeded the other samples substantially. Excluding these nine divergent *E. coli* and *R. solanacearum* samples, the remaining 482 samples' PC1 had 7.0% of variation and PC2 had 5.0% (Figure S4C). The plasmid-encoded proteins had a similar rate of covariation across all 491 samples where 16.8% of variation was allocated to PC1, and 14.9% to PC2 (Figure S4B), with higher heterogeneity in the 482 samples (6.8% to PC1, 4.9% to PC2, Figure S4D, see full figure at FigShare https://doi.org/10.6084/m9.figshare.19525465, and data at FigShare https://doi.org/10.6084/m9.figshare.19525471). These patterns were largely recapitulated by LSA of all samples, which showed limited variation across the 1st (10.0% of variation) and 2nd (4.1% of variation) dimensions (Figure S5).

Plasmid-encoded proteins are rare and collectively interact with diverse chromosomal proteins. We examined the samples' 3023 unique plasmid-encoded proteins to identify PPIs where one or both proteins were plasmid-related (termed plasmid-related PPIs here, see data at FigShare https://doi.org/10.6084/m9.figshare.19672527) to distinguish them from chromosomal PPIs where no plasmid-encoded proteins were found, and we also tracked the rarer category of plasmid-related PPIs where both protein pairs were plasmid-encoded (see data per sample at shorturl.at/mtFNZ). Overall, as a percentage of all PPIs, plasmid-related PPIs constituted a median of 0.35 ± 4.76% (mean ± SD) PPIs per sample, including 0.02% of all PPIs exclusively involving plasmid-encoded proteins. 3874 samples had zero detected plasmid-related PPIs, 489 samples had at least one plasmid-related PPI (two samples had one plasmid-encoded protein with no PPIs). These 489 samples had a median of 169 PPIs where at least one protein was plasmid-related and a median of 53,947 ± 27,590 chromosomal PPIs per sample, which was comparable to the rate for all samples (48,834 ± 46,615, Table S3). Their total number of plasmid-related PPIs had no clear association with the numbers of proteins per sample ($r = 0.05$), but did with the numbers of chromosomal PPIs per sample ($r = 0.27$, Figure S6). Of these 489 samples, 283 had at least one PPI involving only plasmid-encoded proteins, with a median of five such PPIs.

We examined the plasmid-gene associations across 331 samples that had at least three of the 1733 plasmid-encoded proteins detected in at least three samples using NMF (Figure S7). This allocated these samples and proteins to six distinct groups based on their pattern of plasmid-encoded protein sharing (Figure S8, see full figure at FigShare https://doi.org/10.6084/m9.figshare.19525492), and the proteins to these same six groups as mixture coefficients based on their prevalence in the samples (Figure S9, see full figure at FigShare https://doi.org/10.6084/m9.figshare.1952559). There were just five samples in rank 3 that were five *E. coli* (all but *E. coli* 536), which was far below the mean number of samples per rank (77.3, Figure S10). These were associated with 1065 plasmid-encoded genes, far above the average number of proteins per rank (526.2, Figure S11).

Similarly, we investigated 2363 plasmid-encoded proteins' associations based on PPI rates across the same 481 samples where each protein had > 30 PPIs and each sample had > 10 PPIs. This allocated these samples and proteins to seven distinct groups based on their pattern of PPI numbers (Figure S12, see full figure at FigShare https://doi.org/10.6084/m9.figshare.19611276), and these same proteins to the seven groups as mixture coefficients based on their PPI numbers for each sample (Figure S13, see full figure at FigShare https://doi.org/10.6084/m9.figshare.19611282). Like above, all six *E. coli* (including *E. coli* 536) were the sole members of a rank (five), and corresponded to 238 proteins' PPI rates as mixture coefficients (Figure S14), which again was more than the average number of PPI rates associated with each rank (748, Figure S15).

Plasmid-encoded proteins have two-fold higher rates of protein–protein interactions. The mean number of PPIs per protein was 2.2-fold higher for plasmid-related proteins compared to chromosomal ones using data for all samples, though with large variations (n = 13,348 plasmid-related proteins in 491 samples: 39.0 ± 35.0 vs n = 15,450,151 chromosomal proteins in 4363 samples: 17.6 ± 5.7, mean ± SD, t-test $p = 4.8e{-}16$, Fig. 4). This was not due to artefacts in the 489 samples with plasmid-related PPIs: their rate of chromosomal PPIs at 14.8 ± 3.7 PPIs per protein was similar to the other samples (Table S3, see data at FigShare https://doi.org/10.6084/m9.figshare.19575820). However, this effect was protein-specific and not consistently discriminatory because these different groups overlapped considerably (Fig. 4C), implying that genomic context was not a major determinant of PPI rates, and plasmid-related proteins simply had a wider range of PPI levels. To control for the high variation in PPI rates per plasmid-related protein, we found that 19% (65 out of 345) of samples with sufficient numbers of plasmid-related proteins and PPIs to test quantitatively had plasmid-encoded proteins with much higher PPI rates per sample (t-test $p < 0.05$), including all six *E. coli* (Figure S16, see data at FigShare https://doi.org/10.6084/m9.figshare.20230299). As an illustration of PPI network structure, a network centred on the SfmC protein in *E. coli* K12 MG1655 showed high levels of connectivity amongst 20 other proteins (Figure S17, Table S4).

Nearly all (> 99%) plasmid-encoded proteins had a PPI with a chromosomal protein, whereas only 7% of plasmid-encoded proteins had a PPI with another plasmid protein in the 489 samples with at least one plasmid-related PPI. Using the plasmid protein frequency information above, we tested for structure between plasmid-related proteins and chromosomal ones based on the rates of intergroup PPIs in a manner analogous to Wright's F-statistic to quantify the difference between the observed and expected levels of plasmid-chromosome and plasmid-specific PPIs per sample[78]. We found a relatively higher rate of plasmid-plasmid PPIs compared to plasmid-chromosome PPIs, which in turn were much higher than chromosomal PPIs (Figure S18). This suggested that plasmid-encoded proteins tended to interact with other plasmid proteins more often than would be expected by chance, indicating PPI network separation of plasmid-encoded and chromosomal proteins.

Protein–protein interaction networks are generally robust to the loss of plasmid proteins. To examine the effect of plasmid-encoded 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 plasmid-driven effects (Figure S19). The aPMNLE reflects indirect (or secondary) connections, which may be

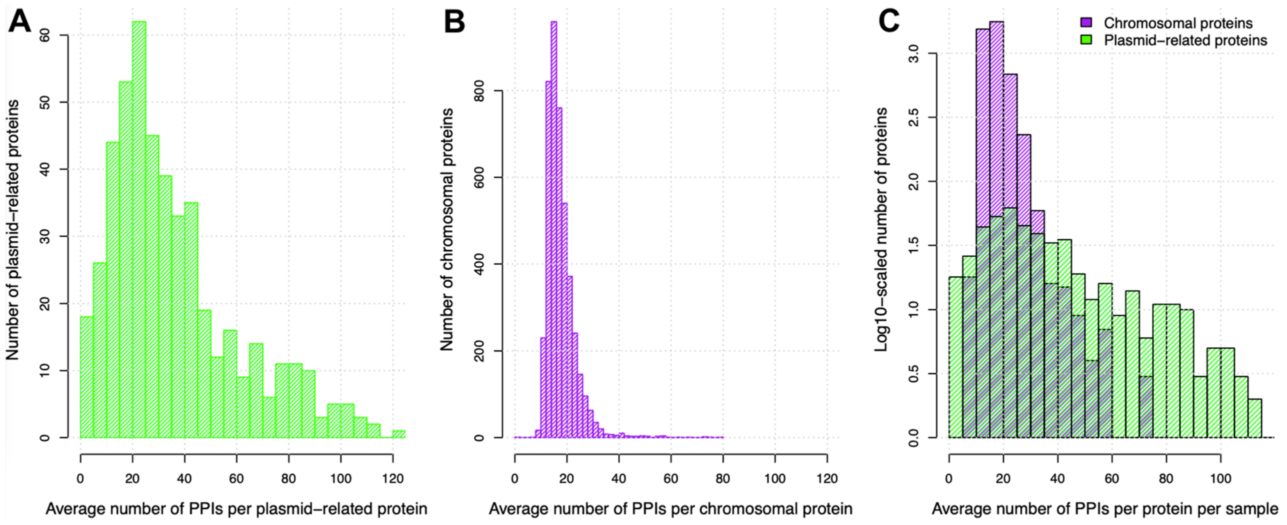

Figure 4. The average number of PPIs per (**A**) plasmid-related protein (green, 39.0 ± 35.0, mean ± SD) was higher than that for (**B**) chromosomal proteins (purple, 17.6 ± 5.7), but (**C**) there was no consistent evidence for a consistent difference in PPI rates between these two groups (plasmid-related proteins in green, chromosomal proteins in purple—note log10-scaled y-axis and minor differences in binning per plot).

reduced when plasmid-related proteins are removed if there are no alternate paths between remaining chromosomal proteins. However, if there are many PPI paths between chromosomal proteins, then this high indirect connectivity may indicate robustness within the PPI network such that the elimination of plasmid-related proteins has no effect (plasmid-related proteins thus have few indirect connections). Moreover, if plasmid-related proteins tend toward the PPI network periphery, the total number of indirect connections may not change relative to a larger drop in the numbers of PPIs, again indicating plasmid-related proteins had few indirect connections. In contrast, if plasmid-related proteins are central and mixed among chromosomal proteins in the PPI network, the number of indirect connections and numbers of PPIs will be closely correlated (here, plasmid-related proteins have proportional indirect connection rates).

A majority of samples (96.3%, 471 out of 489 samples) showed no substantive differences in aPMNLE values between all proteins versus chromosomal ones alone, indicating that plasmid-encoded proteins had no large effects on the PPI network structure even though the plasmid-related proteins contributed large numbers of PPIs, consistent with robust networks. This was consistent with the idea that for these 471 samples the plasmid-encoded proteins overall had comparable indirect connection rates with the chromosomal proteins, but perhaps higher relative PPI rates because no evidence of excessive indirect connection numbers was observed.

3.7% (18 out of 489) samples had an aPMNLE for all proteins (including plasmid-encoded ones) exceeding the chromosomal aPMNLE by two standard deviations, indicating that plasmid-encoded proteins increased the indirect connection rates and that these PPI networks were less robust. No samples had a chromosomal aPMNLE that was two standard deviations above the full network aPMNLE, so no plasmid-encoded protein set strongly increased the rate of indirect connections per sample. When compared to the other 471 samples with plasmid genes, these 18 had altered aPMNLE values for chromosomal proteins (t-test p = 4.9e−5) but not for all proteins (p = 0.82, Figure S20). Furthermore, if the plasmid-encoded proteins were included for each sample, the number of PPI network connected components grew by 71% for the 471 samples with robust networks, whereas for these 18 samples with less robust networks it grew by 327%, indicating that these plasmid-encoded proteins had high PPI rates, fewer indirect connections and were often disconnected from the main network. We also observed these 18 had fewer genes per sample compared to the larger set of 471 (2314 ± 1429 vs 4013 ± 1336, mean ± SD), perhaps because smaller genomes have fewer PPIs overall and so were less robust to plasmid-related protein removal.

To quantify the extent to which plasmid-encoded proteins increased PPI rates but not indirect connections, we compared the correlation between plasmid-related proteins and PPIs per sample in the 489 samples with plasmid-related PPIs, which were positively correlated (r = 0.51, p = 1e−143, Fig. 5A). However, increasing the number of plasmid-encoded proteins did not have the same magnitude of effect on the number of indirect connections per sample (r = 0.06, p = 1, Fig. 5B). Similarly, the latter was negatively associated with the number of plasmid-related PPIs (r = − 0.08, p = 0.30, Fig. 5C). The partial correlations for each pair of variables were similar (0.86, 0.04, − 0.10, respectively). For example, the 69 plasmid-encoded proteins of *Serratia marcescens* subsp. marcescens Db11 had 92 plasmid-plasmid PPIs with one another and an additional 2214 plasmid-chromosome PPIs (with an average of 33.4 PPIs per plasmid protein), compared to 58,857 chromosomal PPIs among the 4614 chromosomal proteins (12.8 PPIs per chromosomal protein, Table S4, Figure S21). Overall, this implied that plasmid-related proteins tended to be at the PPI network periphery connected to chromosomal proteins, and not mixed usually with the main chromosomal proteins in the PPI networks.

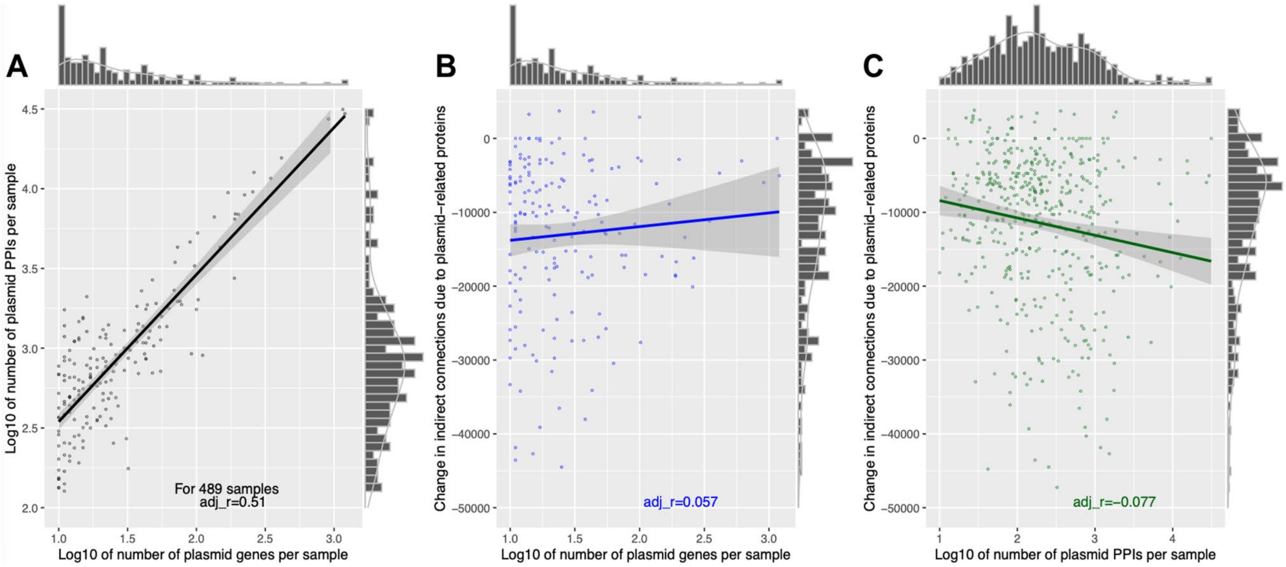

Figure 5. The (**A**) log10-scaled numbers of plasmid-related proteins per sample (x-axis) was positively correlated with the numbers of plasmid PPIs per sample (y-axis) (black, r = 0.51) in the 489 samples with plasmid-related PPIs, but less so with (**B**) the change in the numbers of indirect connections per sample due to the inclusion of the plasmid-encoded proteins (blue, r = 0.06). (**C**) The log10-scaled numbers of plasmid PPIs per sample (x-axis) was negatively associated with the numbers of indirect connections per sample due to the inclusion of the plasmid-encoded proteins (green, r = − 0.08). The lines of best fit for each plot shows the linear correlation of the number of PPIs per sample with the number of indirect connections per protein. The histograms indicate the marginal densities per axis for each dataset.

Plasmid-related proteins are generally peripheral to PPI networks. Given that plasmid-related proteins were highly connected but not central to PPI networks mainly composed of chromosomal proteins, we scaled the numbers of indirect connections by protein. We observed a trimodal pattern of indirect connections per protein. In the 489 samples with plasmid-related PPIs, the chromosomal proteins had more indirect connections per protein than plasmid-encoded proteins (average per sample 3.7 ± 2.1 vs 0.9 ± 0.6, t-test p = 5.2e−15). The indirect connections per protein in these 489 samples were positively correlated with the number of PPIs per sample for chromosomal (r = 0.61, Fig. 6A) but not all proteins (r = 0.02) due to the effect of plasmid-encoded proteins, confirming that the latter tended not to be mixed among the PPI networks. The chromosomal proteins of those 489 samples had a lower level of indirect connections per protein compared to the 3874 samples without plasmid-related PPIs (average per sample 3.7 ± 2.1 vs 5.5 ± 3.1, t-test p = 5.2e−15). The latter 3874 samples had higher levels of indirect connections that were positively correlated with the PPIs per sample, as previously observed in the chromosomal proteins for 489 samples (Fig. 6B).

Plasmid-related proteins had a median of 4.4-fold fewer indirect connections per sample than chromosomal proteins in the set of 489 samples (Fig. 7A). We also observed that additional plasmid-chromosome PPIs constituted 89.7% of all PPIs in these 489, and 93% of plasmid-encoded proteins had no plasmid-related PPIs. To explain this, consider a PPI network where part of it had four chromosomal proteins (A, B, C, D) sharing PPIs A-B, B-C, C-D and D-A, so it has one indirect connection (with a path A-B-C-D, Figure S22). If we added a plasmid protein P to this network such that it had PPIs with all four proteins, the indirect connection across A-B-C-D would be lost but the network would gain four PPIs. An alternative model where A, B, C and D had no PPIs with one another would be less likely because then the addition of P would create PPIs and more indirect connections, which was not observed in our findings. In addition, if A, B, C and D interact with P then it is more likely that they are both functionally related and interact with one another.

Finally, we assessed the numbers of PPI network connected components for all proteins versus chromosomal ones: the latter had fewer connected components compared to all proteins in the 489 samples with plasmid-related PPIs (50.0 ± 34.6 vs 227 ± 95.9, median ± SD, Fig. 7B). The numbers of proteins per connected component was more strongly associated with the chromosomal proteins than for all proteins in relation to the numbers of PPIs (r = 0.35 vs 0.01), chromosome-related PPIs (r = 0.35 vs 0.00), and proteins per sample (r = 0.18 vs 0.00) but not plasmid-related PPIs (r = 0.04 vs 0.00, Figure S23) in these 489 samples. Additionally, these 489 samples had far more connected components compared to 3872 samples without plasmid-related PPIs (means 222.4 vs 40.5, t-test p = 4.8e−15). This highlighted that when there were more PPIs and proteins present, chromosomal proteins were spread across fewer connected components, whereas plasmid proteins could be found in more separate connected components.

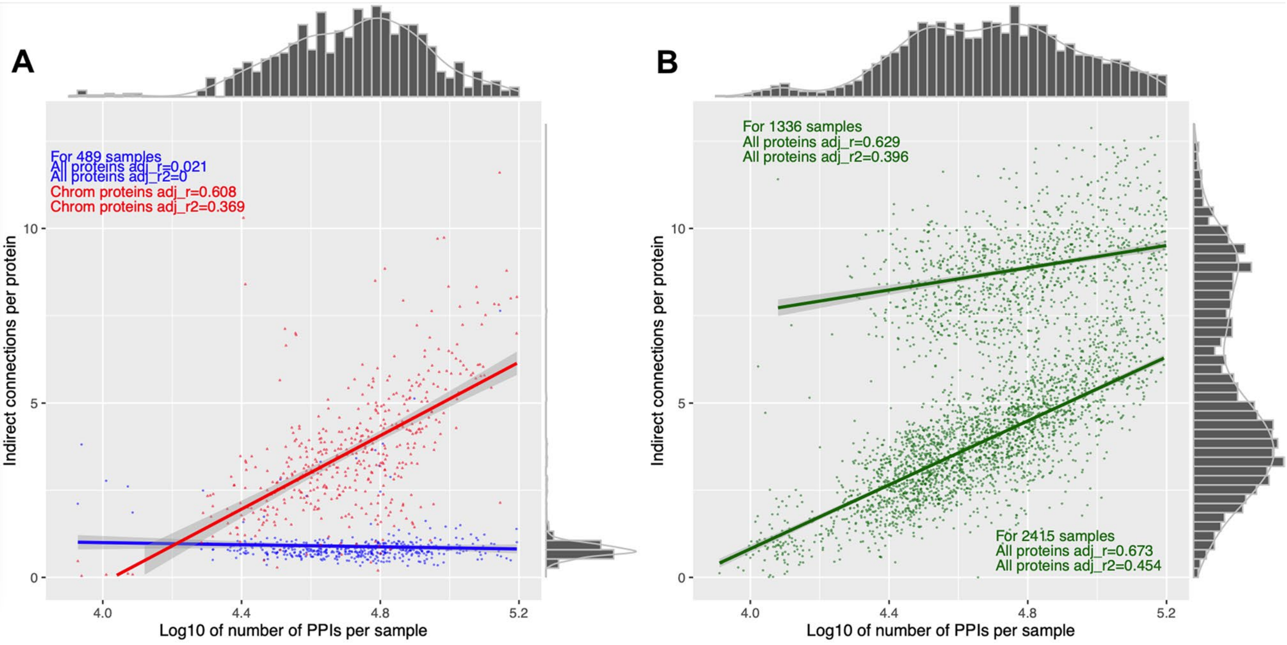

Figure 6. The number of PPIs per sample (x-axis) was positively correlated with the number of indirect connections per protein for (**A**) chromosomal (red, r = 0.61) but not all proteins (including plasmid-encoded ones, blue, r = 0.02) in the 489 samples with plasmid-encoded proteins. (**B**) The other samples with no plasmid-encoded proteins had positive correlations (green, r = 0.63 and r = 0.67) though with a bimodal pattern (groups of n = 2415 and n = 1336). The lines of best fit for each plot shows the linear correlation of the number of PPIs per sample with the number of indirect connections per protein. The histograms indicate the marginal densities per axis for each dataset.

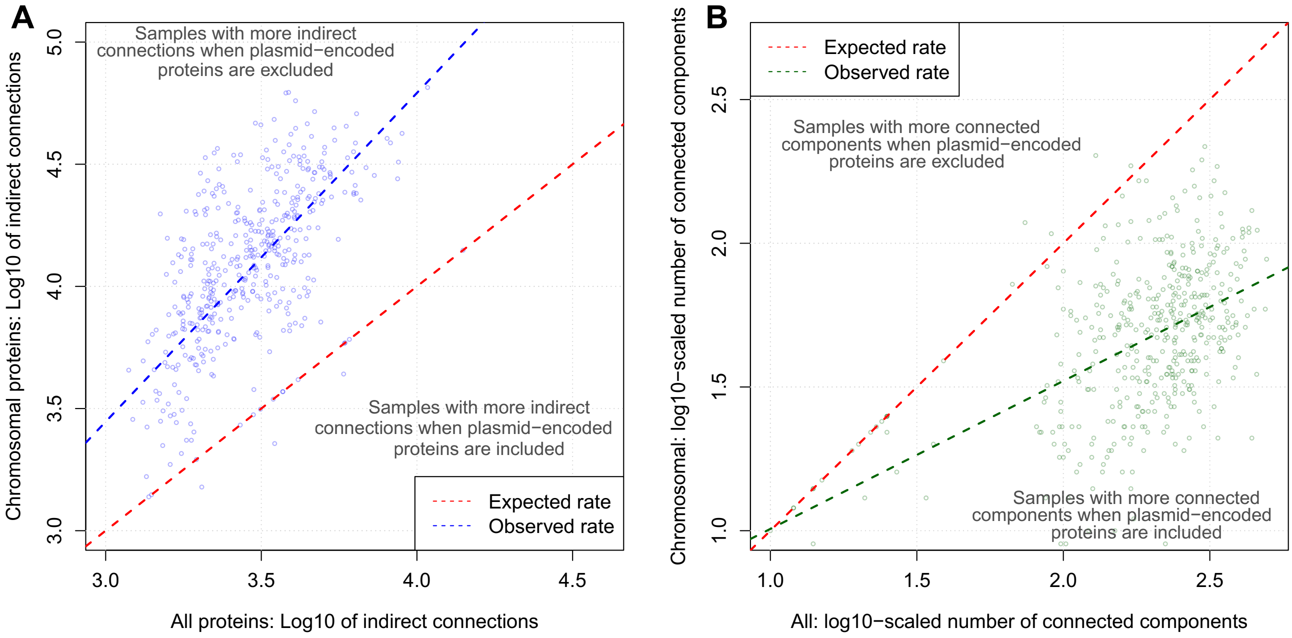

Figure 7. The log10-scaled numbers of (**A**) indirect connections (blue) and (**B**) connected components (green) for all (chromosomal and plasmid) proteins (x-axis) related to chromosomal proteins only (y-axis) for the 489 samples with plasmid-related PPIs. The red lines show the expected rate if (**A**) the numbers of indirect connections for all proteins and chromosomal proteins were identical and (**B**) the numbers of connected components for all proteins and chromosomal proteins were identical. (**A**) The blue line shows the correlation between the numbers of indirect connections for all proteins and chromosomal proteins (> 98% of samples had more indirect connections when plasmid-related proteins were omitted). (**B**) The green line shows the correlation between the numbers of connected components for all proteins and chromosomal proteins (> 97% of samples had more connected components when plasmid-related proteins were retained).

Discussion

Bacterial genome structure has been shaped by HGT more so than gene duplication in many species, such as *E. coli*[89,90]. Mobile plasmid-encoded genes can endow host cells with new phenotypes, including those linked to changed adhesion, AMR and virulence[91], and their fitness effects are associated with specific genetic, regulatory and protein interactions[8]. To further our understanding of how new plasmid-encoded proteins can affect the structure and signalling bacterial PPI networks, we compared the PPI profiles of plasmid-linked and chromosomal proteins across 4363 representative samples of well-characterised bacteria using data from StringDB to find 491 samples with evidence of plasmid-encoded genes within the same species.

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-encoded proteins with chromosomal ones. Most plasmid-encoded proteins were rare across bacteria, and the total fraction of the genome they constituted varied extensively. 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 within the same species, separate from chromosomal PPIs where neither protein was plasmid-linked. Only 11.2% of samples (489 out of 4363) had plasmid-related proteins and PPIs. Even in these samples, plasmid-related genes made up only 0.65% of all genes detected and thus were rare. As a consequence, plasmid-related PPIs made up only 0.14% of all PPIs, and PPIs that were exclusively plasmid-related were very rare (0.016%).

Plasmid-encoded proteins covaried with one another depending on the genetic background of the host cell such that clustering the samples based on their plasmid-encoded gene similarity identified six groups of covarying plasmid-encoded proteins and bacterial samples, including a specific one for *E. coli* alone. This aligned with existing expectations of plasmid-host co-evolution[30,31]. The *Enterobacterales* was the clearly the most variable order, but overall known taxonomical differences were not apparent. Similarly, the PPI rates per protein per sample had even less taxonomical resolution power, showing that the numbers of PPIs per protein was typically a more dynamic feature.

The complexity hypothesis[38] asserts that proteins undergoing HGT have fewer PPIs and are mostly operational genes (those involved in cellular processes) rather than informational genes (those involved in transcription or translation)[38,92,93]. It has been asserted that proteins with many PPIs are less likely to be encoding by genes that undergo HGT extensively[89,94–96], but we found that plasmid-linked proteins had two-fold more PPIs on average. Nonetheless, the predictive power of this across chromosomal versus plasmid-related proteins was low, suggesting that there is no intrinsic difference in PPI rates between these groups of proteins. Nuances to the complexity hypothesis have been found before: firstly, the being in a protein complex versus HGT chances[94]; secondly, the effect of cellular component state, function complexity and function conservation on adaptive evolution potential[41]; thirdly, HGT-linked genes have higher numbers of regulatory factors[89]; and fourthly, PPI connectivity has a large effect for proteins with a similar function[95,96]. 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 given our results here. An important caveat is that the general trends we observed may not hold for all genera, families or orders.

We found that plasmid-related proteins interacted with large numbers of other chromosomal and plasmid proteins. However, plasmid-related proteins contributed much less towards indirect (aka secondary) connections, unlike chromosomal proteins. We found evidence of PPI network structure between chromosomal and plasmid-encoded proteins, finding that plasmid proteins had a higher rate of plasmid rather than chromosomal protein partners than expected. In addition, by using the aPMNLE as a measure of indirect connections, most samples with plasmid-related PPIs (> 96%) were robust to the loss of plasmid-encoded proteins. The samples with networks more affected by plasmid-related protein loss had smaller genomes, and thus less redundancy in PPI network paths. Altogether, this was consistent with existing assertions that plasmid-encoded proteins tend to be at the periphery of a PPI network[88,89], perhaps because plasmids seldom encode essential proteins[5], though gene essentiality depends on the host[97].

Our findings were limited by a number of issues. First, 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 proteins were likely misclassified as chromosomal or plasmid-encoded. Second, many plasmids and bacterial samples are not annotated consistently, so gene name matches were undoubtedly missed. Third, 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. Fourth, the numbers of proteins and PPIs per sample varied widely due to the diverse ranges of genome sizes examined and the other limitations outlined here: testing if genome size affects plasmid-related protein context in PPI networks is an area for further exploration. Fifth, we obtained gene name data per sample from StringDB[59,60], and yet 5918 plasmid-encoded genes were not detected in these samples, which highlighted imprecision either in the plasmid annotation, the sample annotation or (most likely) both. Six, we did not perform sequence alignment nor read mapping to confirm at the sequence level matches, which is a next step following this work.

In summary, a broad scan of PPIs between plasmid-related proteins and chromosomal ones across most bacteria showed that very few genes in total were plasmid-related (0.65%) and that only 11% of samples had plasmid-related PPIs. There was a moderate but complex taxonomic signal arising from the plasmid-related gene profiles that was only clear for *E. coli*. Plasmid-related PPI rates varied more than chromosomal ones, and showed structure in the form of relatively higher rates of plasmid-plasmid PPIs compared to plasmid-chromosome ones.

From this, we inferred that plasmid-related proteins mainly occupy PPI network peripheries, and thus may be more expendable than chromosomal proteins. The data from this study (table below) is an asset for the community to explore further hypotheses that inform on the functional effects of plasmid gene gain/loss on host cells. Genes mediating AMR, virulence, intergration and DNA replication are examples of important functional groups whose context in plasmid-chromosome PPI networks that could be explored further to identify the signalling pathways affected by gene gain/loss. Studies of samples with known plasmid compositions would illuminate the effects of plasmid-encoded and chromosomal proteins on PPI networks. 5918 plasmid-encoded genes were not found in the samples studied here, indicating that sequence-level alignment, annotation and comparison of these sample-plasmid combinations may reveal the functional effects of these genes' products.

Data availability

Item	Brief description	FigShare link
Table S1	Plasmid-encoded 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
Figure S2	Heatmap of the bacterial plasmid-encoded genes	https://doi.org/10.6084/m9.figshare.19525453
Figure S3	Dendrogram of the bacteria based on plasmid-encoded gene similarity	https://doi.org/10.6084/m9.figshare.19525435
Figure S5	Heatmap of the bacteria and plasmid-encoded 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 S8	NMF heatmap of the metasample clustering of plasmid genes	https://doi.org/10.6084/m9.figshare.19525492
Figure S9	NMF heatmap of the mixture coefficient for the plasmid-encoded genes	https://doi.org/10.6084/m9.figshare.19525591
Figure S12	NMF heatmap of the metasample clustering of plasmid PPI rates	https://doi.org/10.6084/m9.figshare.19611276
Figure S13	NMF heatmap of the mixture coefficient for the plasmid-related PPI rates	https://doi.org/10.6084/m9.figshare.19611282
Data for all PPI analysis	4363 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	4363 samples' CSV files for each sample showing PPIs, associated genes and StringDB identifiers	shorturl.at/mtFNZ
Data for Figure S16	The numbers of chromosomal & plasmid genes for 343 as PDF image files	https://doi.org/10.6084/m9.figshare.20230299

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 https://github.com/arahm/HomologyLive and data files are there at https://figshare.com/projects/Bacteria_mobilome_2022/136720.

Received: 11 July 2022; Accepted: 19 September 2022
Published online: 10 November 2022

References

1. Canosi, U., Lüder, G. & Trautner, T. A. SPP1-mediated plasmid transduction. *J. Virol.* **44**(2), 431–436. https://doi.org/10.1128/JVI.44.2.431-436.1982 (1982).
2. Erdmann, S., Tschitschko, B., Zhong, L., Raftery, M. J. & Cavicchioli, R. A plasmid from an Antarctic haloarchaeon uses specialized membrane vesicles to disseminate and infect plasmid-free cells. *Nat. Microbiol.* **2**(10), 1446–1455. https://doi.org/10.1038/s41564-017-0009-2 (2017).
3. Zhang, X. *et al.* Improvement in the efficiency of natural transformation of *Haemophilus parasuis* by shuttle-plasmid methylation. *Plasmid* **98**, 8–14. https://doi.org/10.1016/j.plasmid.2018.07.001 (2018).
4. Wein, T. & Dagan, T. Plasmid evolution. *Curr. Biol.* **30**(19), R1158–R1163. https://doi.org/10.1016/j.cub.2020.07.003 (2020).

5. Wein, T. *et al.* Essential gene acquisition destabilizes plasmid inheritance. *PLoS Genet.* **17**(7), e1009656. https://doi.org/10.1371/journal.pgen.1009656 (2021).
6. Norman, A., Hansen, L. H. & Sørensen, S. J. Conjugative plasmids: Vessels of the communal gene pool. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**(1527), 2275–2289. https://doi.org/10.1098/rstb.2009.0037 (2009).
7. Downing, T. Tackling drug resistant infection outbreaks of global pandemic *Escherichia coli* ST131 using evolutionary and epidemiological genomics. *Microorganisms* **3**(2), 236–267 (2015).
8. Hall, J. P. J., Brockhurst, M. A. & Harrison, E. Sampling the mobile gene pool: Innovation via horizontal gene transfer in bacteria. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**(1735), 20160424. https://doi.org/10.1098/rstb.2016.0424 (2017).
9. Decano, A. G. *et al.* Plasmids shape the diverse accessory resistomes of *Escherichia coli* ST131. *Access Microbiol.* https://doi.org/10.1099/acmi.0.000179 (2020).
10. Ahmer, B. M. *et al.* The virulence plasmid of *Salmonella typhimurium* is self-transmissible. *J. Bacteriol.* **181**, 1364–1368 (1999).
11. Stasiak, G. *et al.* Functional relationships between plasmids and their significance for metabolism and symbiotic performance of *Rhizobium leguminosarum* bv. trifolii. *J. Appl. Genet.* **55**, 515–527 (2014).
12. San, M. A. Evolution of plasmid-mediated antibiotic resistance in the clinical context. *Trends Microbiol.* **26**, 978–985 (2018).
13. Niehus, R. *et al.* Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nat. Commun.* **6**, 8924 (2015).
14. Brockhurst, M. A. *et al.* The ecology and evolution of pangenomes. *Curr. Biol.* **29**(20), R1094–R1103. https://doi.org/10.1016/j.cub.2019.08.012 (2019).
15. Agyekum, A. *et al.* 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), 328–333. https://doi.org/10.1016/j.diagmicrobio.2015.12.010 (2016).
16. Huang, W. *et al.* Emergence and evolution of multidrug-resistant *Klebsiella pneumoniae* with both blaKPC and blaCTX-M integrated in the chromosome. *Antimicrob. Agents Chemother.* **61**(7), e00076-e117. https://doi.org/10.1128/AAC.00076-17 (2017).
17. Irrgang, A. *et al.* 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. https://doi.org/10.3389/fmicb.2017.02318 (2017).
18. Decano, A. G. & Downing, T. An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4071 isolates. *Sci. Rep.* https://doi.org/10.1038/s41598-019-54004-5 (2019).
19. Decano, A. G. *et al.* Complete assembly of *Escherichia coli* ST131 genomes using long reads demonstrates antibiotic resistance gene variation within diverse plasmid and chromosomal contexts. *mSphere* https://doi.org/10.1128/mSphere.00130-19 (2019).
20. Yoon, E. J. *et al.* Beneficial chromosomal integration of the genes for CTX-M extended-SPECTRUM β-lactamase in *Klebsiella pneumoniae* for stable propagation. *mSystems* **5**(5), e00459-e520. https://doi.org/10.1128/mSystems.00459-20 (2020).
21. Ludden, C. *et al.* Genomic surveillance of *Escherichia coli* ST131 identifies local expansion and serial replacement of subclones. *Microbial. Genom.* https://doi.org/10.1099/mgen.0.000352 (2020).
22. Bevan, E. R. *et al.* 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. https://doi.org/10.1016/j.jhin.2021.03.030 (2021).
23. Shawa, M. *et al.* 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. https://doi.org/10.1186/s13756-021-00941-8 (2021).
24. San Millan, A. & MacLean, R. C. Fitness costs of plasmids: A limit to plasmid transmission. *Microbiol. Spectr.* **5**, 5. https://doi.org/10.1128/microbiolspec.MTBP-0016-2017 (2017).
25. Baltrus, D. A. Exploring the costs of horizontal gene transfer. *Trends Ecol. Evol.* **28**(8), 489–495. https://doi.org/10.1016/j.tree.2013.04.002 (2013).
26. Harrison, E. & Brockhurst, M. A. Plasmid-mediated horizontal gene transfer is a coevolutionary process. *Trends Microbiol.* **20**(6), 262–267. https://doi.org/10.1016/j.tim.2012.04.003 (2012).
27. Stalder, T. *et al.* Emerging patterns of plasmid-host coevolution that stabilize antibiotic resistance. *Sci. Rep.* **7**(1), 4853. https://doi.org/10.1038/s41598-017-04662-0 (2017).
28. Ho, J. *et al.* Systematic review of human gut resistome studies revealed variable definitions and approaches. *Gut Microbes.* **12**(1), 1700755. https://doi.org/10.1080/19490976.2019.1700755 (2020).
29. Coelho, L. P. *et al.* Towards the biogeography of prokaryotic genes. *Nature* https://doi.org/10.1038/s41586-021-04233-4 (2021).
30. Loftie-Eaton, W. *et al.* Compensatory mutations improve general permissiveness to antibiotic resistance plasmids. *Nat. Ecol. Evol.* **1**(9), 1354–1363. https://doi.org/10.1038/s41559-017-0243-2 (2017).
31. Jordt, H. *et al.* Coevolution of host-plasmid pairs facilitates the emergence of novel multidrug resistance. *Nat. Ecol. Evol.* **4**(6), 863–869. https://doi.org/10.1038/s41559-020-1170-1 (2020).
32. Andam, C. P. & Gogarten, J. P. Biased gene transfer in microbial evolution. *Nat. Rev. Microbiol.* **9**, 543–555 (2011).
33. Forsberg, K. J. *et al.* Bacterial phylogeny structures soil resistomes across habitats. *Nature* **509**, 612–616 (2014).
34. Soucy, S. M. *et al.* Horizontal gene transfer: Building the web of life. *Nat. Rev. Genet.* **16**, 472–482 (2015).
35. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**(7376), 241–244 (2011).
36. Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**(7612), 435–439 (2016).
37. Porse, A. *et al.* Biochemical mechanisms determine the functional compatibility of heterologous genes. *Nat. Commun.* **9**(1), 522. https://doi.org/10.1038/s41467-018-02944-3 (2018).
38. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci. USA* **96**(7), 3801–3806 (1999).
39. Novick, A. & Doolittle, W. F. Horizontal persistence and the complexity hypothesis. *Biol. Philos.* **35**, 2 (2020).
40. Nakamura, Y. *et al.* Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet.* **36**, 760–776 (2004).
41. Aris-Brosou, S. Determinants of adaptive evolution at the molecular level: The extended complexity hypothesis. *Mol. Biol. Evol.* **22**(2), 200–209. https://doi.org/10.1093/molbev/msi00 (2005).
42. Puigbo, P., Wolf, Y. I. & Koonin, E. V. The tree and net components of prokaryote evolution. *Genome Biol. Evol.* **2**, 745–756 (2010).
43. Dewar, A. E. *et al.* Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range. *Nat. Ecol. Evol.* **5**(12), 1624–1636. https://doi.org/10.1038/s41559-021-01573-2 (2021).
44. Touchon, M. *et al.* Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *PLoS Genet.* **16**(6), e1008866. https://doi.org/10.1371/journal.pgen.1008866 (2020).
45. Babu, M. *et al.* Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J. Mol. Biol.* **358**, 614–633 (2006).
46. De Gelder, L., Ponciano, J. M., Joyce, P. & Top, E. M. Stability of a promiscuous plasmid in different hosts: No guarantee for a long-term relationship. *Microbiology (Reading)* **153**(Pt 2), 452–463. https://doi.org/10.1099/mic.0.2006/001784-0 (2007).
47. Dunn, S., Carrilero, L., Brockhurst, M. & McNally, A. 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-e121. https://doi.org/10.1128/mSystems.00083-21 (2021).

48. Alonso-Del Valle, A. *et al.* Variability of plasmid fitness effects contributes to plasmid persistence in bacterial communities. *Nat. Commun.* **12**(1), 2653. https://doi.org/10.1038/s41467-021-22849-y (2021).
49. Kottara, A., Hall, J. P. J., Harrison, E. & Brockhurst, M. A. Variable plasmid fitness effects and mobile genetic element dynamics across Pseudomonas species. *FEMS Microbiol. Ecol.* **94**(1), fix172. https://doi.org/10.1093/femsec/fix172 (2018).
50. Sheppard, R. J., Beddis, A. E. & Barraclough, T. G. The role of hosts, plasmids and environment in determining plasmid transfer rates: A meta-analysis. *Plasmid* **108**, 102489. https://doi.org/10.1016/j.plasmid.2020.102489 (2020).
51. Gama, J. A., Kloos, J., Johnsen, P. J. & Samuelsen, Ø. Host dependent maintenance of a blaNDM-1-encoding plasmid in clinical *Escherichia coli* isolates. *Sci. Rep.* **10**(1), 9332. https://doi.org/10.1038/s41598-020-66239-8 (2020).
52. Alderliesten, J. B. *et al.* Effect of donor-recipient relatedness on the plasmid conjugation frequency: A meta-analysis. *BMC Microbiol.* **20**(1), 135. https://doi.org/10.1186/s12866-020-01825-4 (2020).
53. Harrison, E., Guymer, D., Spiers, A. J., Paterson, S. & Brockhurst, M. A. Parallel compensatory evolution stabilizes plasmids across the parasitism-mutualism continuum. *Curr. Biol.* **25**(15), 2034–2039. https://doi.org/10.1016/j.cub.2015.06.024 (2015).
54. San Millan, A. *et al.* Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. *Nat. Commun.* **5**, 5208. https://doi.org/10.1038/ncomms6208 (2014).
55. Hall, J. P. J., Wright, R. C. T., Guymer, D., Harrison, E. & Brockhurst, M. A. Extremely fast amelioration of plasmid fitness costs by multiple functionally diverse pathways. *Microbiology (Reading)* **166**(1), 56–62. https://doi.org/10.1099/mic.0.000862 (2020).
56. Modi, R. I., Wilke, C. M., Rosenzweig, R. F. & Adams, J. Plasmid macro-evolution: Selection of deletions during adaptation in a nutrient-limited environment. *Genetica* **84**(3), 195–202. https://doi.org/10.1007/BF00127247 (1991).
57. Porse, A., Schønning, K., Munck, C. & Sommer, M. O. Survival and evolution of a large multidrug resistance plasmid in new clinical bacterial hosts. *Mol. Biol. Evol.* **33**(11), 2860–2873. https://doi.org/10.1093/molbev/msw163 (2016).
58. Lee, M. C. & Marx, C. J. Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet.* **8**(5), e1002651. https://doi.org/10.1371/journal.pgen.1002651 (2012).
59. Szklarczyk, D. *et al.* 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. https://doi.org/10.1093/nar/gkaa1074 (2021).
60. Szklarczyk, D. *et al.* STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**(D1), D607–D613 (2018).
61. R Core Team. 2021. R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.
62. Ihaka, R. & Gentleman, R. R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **3**(5), 299–314 (1996).
63. Winter, D. J. rentrez: An R package for the NCBI eUtils API. *R J.* **9**(2), 520–526 (2017).
64. Wickham, H. 2019. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr.
65. 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.
66. Wickham, H. 2021. forcats: Tools for Working with Categorical Variables (Factors). R package version 0.5.1. https://CRAN.R-project.org/package=forcats.
67. 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.
68. Wickham, H., Bryan, J. 2022. readxl: Read Excel Files. R package version 1.4.0. https://CRAN.R-project.org/package=readxl.
69. Müller, K., Wickham, H. 2021. tibble: Simple Data Frames. R package version 3.1.6. https://CRAN.R-project.org/package=tibble.
70. Wickham, H., Girlich, M. 2022. tidyr: Tidy Messy Data. R package version 1.2.0. https://CRAN.R-project.org/package=tidyr.
71. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**(43), 1686. https://doi.org/10.21105/joss.01686 (2019).
72. Chen, H. 2022. VennDiagram: Generate High-Resolution Venn and Euler Plots. R package version 1.7.3. https://CRAN.R-project.org/package=VennDiagram.
73. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
74. Slowikowski, K. 2021. ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.9.1. https://CRAN.R-project.org/package=ggrepel.
75. Zhou, H. *et al.* Functions predict horizontal gene transfer and the emergence of antibiotic resistance. *Sci. Adv.* **7**(43), eabj5056. https://doi.org/10.1126/sciadv.abj5056 (2021).
76. Brooks, L., Kaze, M. & Sistrom, M. A Curated, comprehensive database of plasmid sequences. *Microbiol. Resour. Announc.* **8**(1), e01325-e1418. https://doi.org/10.1128/MRA.01325-18 (2019).
77. Galata, V., Fehlmann, T., Backes, C. & Keller, A. PLSDB: A resource of complete bacterial plasmids. *Nucleic Acids Res.* **47**(D1), D195–D202. https://doi.org/10.1093/nar/gky1050 (2019).
78. Wright, S. The genetical structure of populations. *Ann. Eugen.* **15**, 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x (1951).
79. Chamberlain S, et al. 2020. taxize: Taxonomic information from around the web. R package version 0.9.98. https://github.com/ropensci/taxize.
80. Galili, T. dendextend: An R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btv428 (2015).
81. 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.
82. Rusakovica, J. *et al.* Probabilistic latent semantic analysis applied to whole bacterial genomes identifies common genomic features. *J. Integr. Bioinform.* **11**(2), 243. https://doi.org/10.2390/biecoll-jib-2014-243 (2014).
83. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
84. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinform.* **11**, 367 (2010).
85. Huber, V. J. *et al.* Orchestrating high-throughput genomic analysis with bioconductor. *Nat. Methods* **12**, 115 (2015).
86. Vietoris, L. Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. *Math. Ann.* **97**, 454–472 (1927).
87. Rahm, A. HomologyLive; 2019 https://github.com/arahm/HomologyLive.
88. Lercher, M. J. & Pál, C. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol. Biol. Evol.* **25**, 559 (2008).
89. Price, M. N., Dehal, P. S. & Arkin, A. P. Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli*. *Genome Biol.* **9**(1), R4. https://doi.org/10.1186/gb-2008-9-1-r4 (2008).
90. The LinBox Group Exact linear algebra over the Integers and finite rings, version 1.1.6; 2008
91. Lipworth, S. *et al.* The mobilome associated with Gram-negative bloodstream infections: A large-scale observational hybrid sequencing based study. *MedRxiv* https://doi.org/10.1101/2022.04.03.22273290 (2022).
92. Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. USA* **95**, 6239–6244 (1998).
93. Jain, R., Rivera, M. C., Moore, J. E. & Lake, J. A. Horizontal gene transfer in microbial genome evolution. *Theor. Popul. Biol.* **61**(4), 489–495. https://doi.org/10.1006/tpbi.2002.1596 (2002).

94. Wellner, A., Lurie, M. N. & Gophna, U. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. *Genome Biol.* **8**(8), R156. https://doi.org/10.1186/gb-2007-8-8-r156 (2007).
95. Cohen, O. *et al.* The complexity hypothesis revisited: Connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* **28**, 1481–1489 (2011).
96. Cohen, O., Gophna, U. & Pupko, T. The complexity hypothesis revisited: Connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* **28**(4), 1481–1489. https://doi.org/10.1093/molbev/msq333 (2011).
97. Rousset, F. *et al.* The impact of genetic diversity on gene essentiality within the *Escherichia coli* species. *Nat. Microbiol.* **6**(3), 301–312. https://doi.org/10.1038/s41564-020-00839-y (2021).

Acknowledgements

The authors thank Min Jie Lee for certain R code contributions. The authors thank the reviewers for their constructive suggestions.

Author contributions

T.D.—conceptualization, methodology, software, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, visualisation, project administration. A.R.—methodology, software, formal analysis, writing—review and editing. all authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-20809-0.

Correspondence and requests for materials should be addressed to T.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.