nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-021-01573-2

Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range

Anna E. Dewar[1,3]✉, Joshua L. Thomas[1,3], Thomas W. Scott[1], Geoff Wild[2], Ashleigh S. Griffin[1], Stuart A. West[1,4] and Melanie Ghoul[1,4]

Horizontal gene transfer via plasmids could favour cooperation in bacteria, because transfer of a cooperative gene turns non-cooperative cheats into cooperators. This hypothesis has received support from theoretical, genomic and experimental analyses. By contrast, we show here, with a comparative analysis across 51 diverse species, that genes for extracellular proteins, which are likely to act as cooperative 'public goods', were not more likely to be carried on either: (1) plasmids compared to chromosomes; or (2) plasmids that transfer at higher rates. Our results were supported by theoretical modelling which showed that, while horizontal gene transfer can help cooperative genes initially invade a population, it has less influence on the longer-term maintenance of cooperation. Instead, we found that genes for extracellular proteins were more likely to be on plasmids when they coded for pathogenic virulence traits, in pathogenic bacteria with a broad host-range.

The growth and success of many bacterial populations depends on the production of cooperative 'public goods'[1–4]. Public goods are molecules whose secretion provides a benefit to the local group of cells. Examples include iron-scavenging siderophores[5], exotoxins that disintegrate host cell membranes[6,7] and elastases that break down connective tissues[8–10]. A problem is that cooperation can be exploited by 'cheats': cells that avoid the cost of producing public goods but can still use and benefit from those produced by cooperative cells[3,11,12]. What prevents cheats from outcompeting cooperators and ultimately destabilizing cooperation?

In bacteria, some genetic elements are able to move between cells[13]. This horizontal gene transfer has been suggested as a mechanism to help stabilize the production of cooperative public goods[14–18] (Fig. 1a). If a gene coding for the production of a public good can be transferred horizontally, it would allow cheats to be 'infected' with the cooperative gene and turned into cooperators. Theoretical models have shown that this can facilitate the invasion of cooperative genes, in conditions where they would not be favoured on chromosomes[14–18]. Experiments on a synthetic *Escherichia coli* system have shown that location on a plasmid helped the gene for a cooperative public good to invade, particularly in structured populations[18]. In addition, bioinformatic analyses across a range of species found that genes that code for extracellular proteins, many of which act as public goods, are more likely to be found on plasmids than the chromosome[15,19,20].

There are, however, three potential problems for the hypothesis that horizontal gene transfer favours cooperation. First, previous bioinformatic analyses made important first steps but are not conclusive. One study examined only a single species, which may not be representative of all bacteria[15]. Two additional studies examined multiple species but assumed that genes and genomes from the same and different species can be treated as independent data points in a way that could have led to spurious results[19,20]. Statistical tests typically assume that data points are independent and even slight non-independence can lead to heavily biased results (type I errors)[21,22]. There is an extensive literature in the field of evolutionary biology showing that species share characteristics inherited through common descent, rather than through independent evolution and so cannot be considered independent data points[23–25]. Genomes are nested within species and genes are nested within genomes, multiplying this problem of non-independence, analogous to the problem of pseudoreplication in experimental studies[26–29]. Phylogenetically controlled bioinformatic analyses are required to address this problem of non-independence and test the robustness of previous conclusions.

Second, from a theoretical perspective, while horizontal gene transfer can favour the initial invasion of cooperation, it is not clear if it favours the maintenance of cooperation in the long run[16]. For example, after a plasmid carrying a cooperative gene has spread through a population, a loss-of-function mutation could easily lead to a cheat plasmid evolving, which could then potentially outcompete the plasmid carrying the cooperative gene[16,30]. Theory is required that examines the maintenance as well as the invasion of cooperation, while accounting for important biological details, such as how plasmid transmission depends on the population frequency of the plasmid and how frequently plasmids are lost, for example by segregation during cell division.

Third, there are alternative hypotheses for why genes coding for extracellular proteins might be preferentially carried on plasmids in some species (Fig. 1)[20,31]. Bacteria can rapidly adapt to new and/or changing environments by acquiring new genes via horizontal gene transfer and losing genes no longer required but costly to maintain (Fig. 1b)[32–34]. Genes that facilitate adaptation to environmental variability are often those that code for molecules secreted outside the cell[34–37]. Consequently, we might expect to find genes for extracellular proteins on plasmids to facilitate rapid gain and loss

[1]Department of Zoology, University of Oxford, Oxford, UK. [2]Department of Applied Mathematics, University of Western Ontario, London, Ontario, Canada. [3]These authors contributed equally: Anna E. Dewar, Joshua L. Thomas. [4]These authors jointly supervised this work: Stuart A. West, Melanie Ghoul. ✉e-mail: anna.dewar@zoo.ox.ac.uk

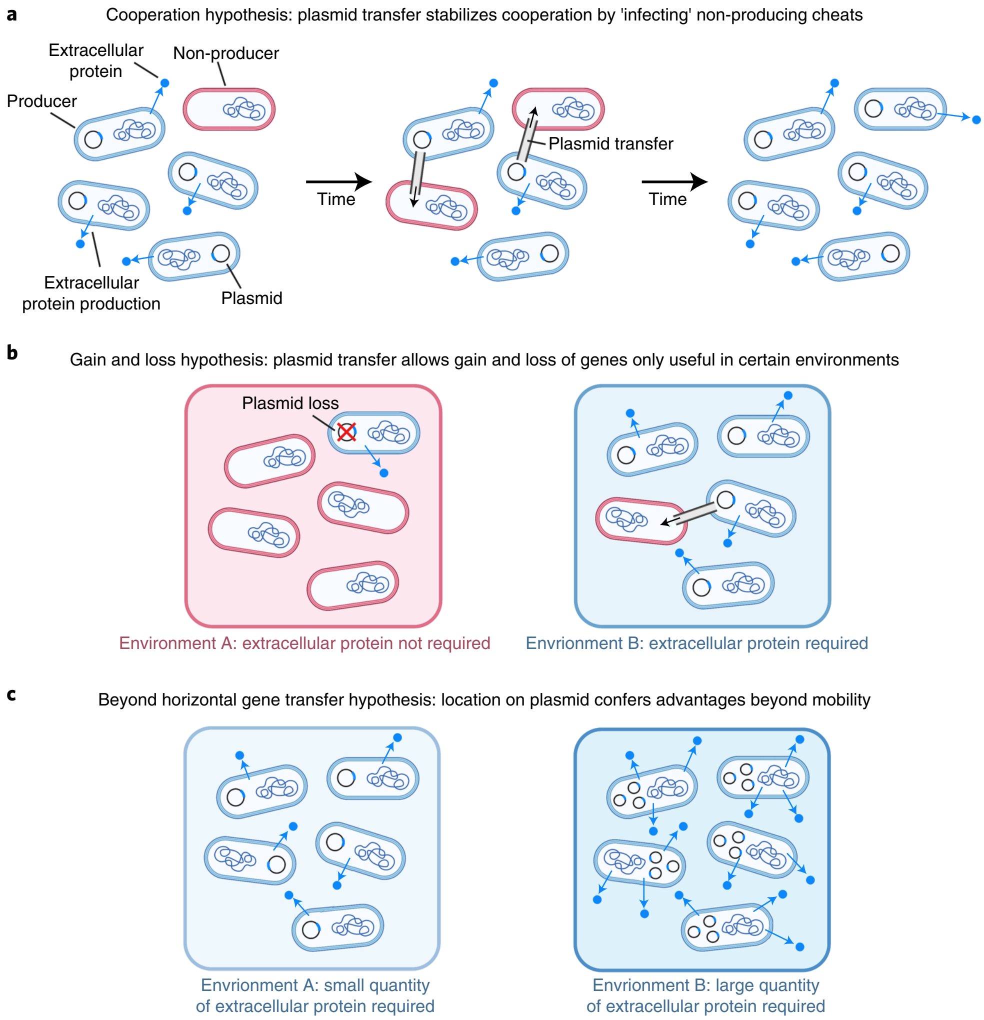

a
Cooperation hypothesis: plasmid transfer stabilizes cooperation by 'infecting' non-producing cheats
Extracellular protein
Non-producer
Producer
Extracellular protein production
Plasmid
Time
Plasmid transfer
Time
b
Gain and loss hypothesis: plasmid transfer allows gain and loss of genes only useful in certain environments
Plasmid loss
Environment A: extracellular protein not required
Envrionment B: extracellular protein required
c
Beyond horizontal gene transfer hypothesis: location on plasmid confers advantages beyond mobility
Envrionment A: small quantity of extracellular protein required
Envrionment B: large quantity of extracellular protein required

Fig. 1 | Three hypotheses about why selection might favour genes coding for extracellular proteins to be located on plasmids. a, Cooperation hypothesis. Blue cells produce extracellular proteins that act as cooperative public goods, while red cells are 'cheats' that exploit this cooperation. Over time, cheats grow faster than cooperators since they forgo the cost of public good production. However, because the gene for the extracellular protein is located on a plasmid, cooperators can transfer the gene to the cheats, turning them into cooperators, increasing genetic relatedness at the cooperative locus and stabilizing cooperation[14–18]. **b**, Gain and loss hypothesis. The production of the extracellular protein is required in some environments but not in others. Transitions between these environments can result from temporal or spatial change. Cells are selected to either lose (environment A) or gain (environment B) the plasmid coding for the production of the extracellular protein. **c**, Beyond horizontal gene transfer hypothesis. The location of a gene on a plasmid could provide a number of benefits, other than the possibility for horizontal gene transfer[38]. For example, when the quantity of extracellular protein required varies across environments (A versus B), plasmid copy number could be varied to adjust production[38]. Created with BioRender.com.

of genes depending on environmental conditions and not because they are cooperative per se. Alternatively, genes may be favoured to be on plasmids for reasons other than horizontal gene transfer (Fig. 1c)[38]. For example, a higher plasmid copy number offers a mechanism for more expression of a gene, potentially even conditionally, in response to certain environmental conditions[38]. The benefit of being able to regulate gene expression in this way could be higher in genes that code for molecules that are secreted outside the cell, when different quantities of molecule are required in different environments. These different hypotheses are not mutually exclusive.

We addressed all three of these potential problems for the hypothesis that horizontal gene transfer favours cooperation. We first tested two predictions that would be expected to hold if horizontal gene transfer favours cooperation. Specifically, cooperative genes would be more likely to be found on: (1) plasmids relative to chromosomes; and (2) more mobile plasmids relative to less mobile plasmids[14–20]. We used phylogeny-based statistical methods that control for the problem of non-independence, analysing 1,632 genomes from 51 bacterial species, to examine the location of genes that code for extracellular proteins. We then used theoretical models, to examine whether horizontal gene transfer facilitates the evolution as well as the initial spread of cooperation.

Finally, we also tested alternative hypotheses for why genes coding for extracellular proteins might be preferentially carried on plasmids. We used three measures of environmental variability to ask whether species that had more variable environments were those most likely to carry genes for extracellular proteins on their plasmids. Additionally, we examined one of these measures in more detail, to help determine whether genes for extracellular proteins were located on plasmids so that they could be gained and lost easily (Fig. 1b) or instead because of some additional benefit conferred by plasmid carriage (Fig. 1c).

Results

Genomic analyses. We use the approach developed by Nogueira et al.[15,19,20] of using PSORTb (ref. [39]) to predict the subcellular location of every protein encoded by 1,632 complete genomes from 51 diverse bacterial species (Extended Data Fig. 1 and Supplementary Table 3). We are also building on the work of researchers who pointed out that extracellular (secreted) proteins are likely to provide a benefit to the local population of cells and hence act as cooperative public goods[2,15,19,20,40]. The advantage of this method is that it allows a large number of genes to be examined, across multiple species.

Overall, we found that the average bacterial genome had 2,696 protein-coding genes on the chromosome(s) and 223 on the plasmid(s). Of these, an average of 57 genes (~2%) coded for the production of an extracellular protein, with 52 on the chromosome(s) and five on the plasmid(s). This means, on average, 1.9% of chromosome genes and 2.4% of plasmid genes coded for extracellular proteins. To control for the number of genomes per species, we first calculated the mean number of genes for each species and then the mean of these species means. Therefore, the values above give an indication of the location of genes coding for extracellular proteins in an average genome. Genes with unknown protein localizations were not included (chromosome, 26.2%; plasmid, 38.3%). Across species, the proportion of genes coding for extracellular proteins for plasmid(s) was generally more variable than for the chromosome(s) (Supplementary Fig. 2). These patterns are very similar to those found previously[15,19,20].

Extracellular proteins are not overrepresented on plasmids. We found that extracellular proteins were not more likely to be carried on plasmids compared to chromosomes (Fig. 2). The difference in the proportion of genes that coded for extracellular proteins between plasmid and chromosome was not significantly different from zero across all species (Markov Chain Monte Carlo generalized linear mixed effects model (MCMCglmm) (ref. [41]); posterior mean = 0.004, 95% credible interval (CI) = −0.063 to 0.057, pMCMC (generally interpreted in a similar way to a P value) = 0.87; $n = 1{,}632$ genomes; R^2 of species sample size = 0.47, R^2 of phylogeny = 0.17; Supplementary Table 2, row 1a). This result was robust to alternative forms of analysis. We also found no significant difference when we: (1) compared chromosomes to plasmids of only certain mobilities (Supplementary Fig. 3 and Supplementary Table 2, rows 20–22); (2) analysed our data by two alternative methods, by looking at the ratio of proportions instead of the difference or by considering only whether the plasmid proportion was greater than the chromosome proportion, removing any effect of the magnitude of this difference (Extended Data Fig. 2 and Supplementary Table 2, rows 2 and 3). Our analyses use a bacterial phylogeny, which assumes that plasmid evolution follows bacterial phylogeny but we also found no significant pattern if we ignored phylogeny and analysed species as independent data points (Fig. 2 and Supplementary Table 2, row 1b; pMCMC = 0.644).

The lack of an overall significant result was clear when looking at the raw data for the different species that we examined (Fig. 2 and Extended Data Fig. 2). There was considerable variation across species in the location of genes coding for extracellular proteins. Overall, extracellular proteins were more likely to be on plasmids in 51% of species (26/51) and more likely to be on the chromosome(s) in 49% (25/51) of species (Extended Data Fig. 2). For example, in *Bacillus anthracis*, genes coding for extracellular proteins were three times more likely to be on plasmids; whereas, in *Acinetobacter baumannii*, genes coding for extracellular proteins were three times more likely to be on the chromosome(s) (Extended Data Fig. 2). Clearly, across species, genes coding for extracellular proteins are not consistently more likely to be on plasmids.

As a control, we also analysed the genomic location of the genes coding for all other classes of protein (Extended Data Fig. 1). Specifically, we analysed genes that coded for the production of cytoplasmic, cytoplasmic membrane, periplasmic, outer membrane and cell wall proteins. We found that none of these protein localizations was significantly overrepresented on plasmids or chromosomes across the 51 species (Extended Data Fig. 3 and Supplementary Table 2, rows 5–10). Plasmids are highly variable in the genes they carry.

Importance of controlling for non-independence of genomes. Our results contrast with previous studies, which found that plasmid genes code for proportionally more extracellular proteins than do chromosomes[15,19,20]. The first of these studies found this pattern across 20 *E. coli* genomes[15]. We also found that genes coding for extracellular proteins in *E. coli* were more likely to be found on plasmids (Fig. 2 and Extended Data Fig. 2). However, Fig. 2 shows that this is not a consistent pattern across species: approximately half (25/51) of the species we analysed showed a pattern in the opposite direction, with genes coding for extracellular proteins more likely to be on their chromosome(s) than on their plasmid(s).

Two subsequent, multispecies studies found that plasmid genes were significantly more likely to code for extracellular proteins than were chromosome genes[19,20]. These studies used statistical tests such as Wilcoxon signed-rank test to ask whether there was a consistent pattern, using bacterial genomes as independent data points. When we analysed our data with the same statistical methods used in these studies, we also obtained a significant result (Wilcoxon signed-rank test; $V = 826{,}530$, $P < 0.001$, $R^2 = 0.385$; $n = 1{,}632$ plasmid–chromosome pairs). When analysing other questions, Garcia-Garcera and Rocha[20] used MCMCglmm to control for phylogeny.

Why does using bacterial genomes as independent data points lead to a significant result? By using a Wilcoxon signed-rank test, at the level of the genome, we are implicitly assuming that all the genomes analysed are: (1) independent from one another; and (2) a representative sample of bacteria in nature. Neither of these are true for multispecies genomic datasets. First, due to shared ancestry, species are not independent from one another and so neither are genomes in such analyses[24,42]. Even a slight lack of independence can lead to heavily biased results in statistical analyses and spurious conclusions[21]. Second, genomic databases tend to have a disproportionate abundance of certain species and genera. This will bias the results towards commonly sequenced species.

Consequently, when asking questions across species, it is inappropriate to treat all the genomes in genomic datasets as independent data points. When we performed an analysis analogous to the Wilcoxon signed-rank test, using the same untransformed data, which produced a significant result above but controlled for the number of genomes per species and the non-independence of species, we no longer found any significant difference between the proportion of plasmid and chromosome genes coding for extracellular proteins (MCMCglmm; posterior mean = 0.017, 95% CI = −0.021 to 0.057, pMCMC = 0.332; $n = 1{,}632$ plasmid–chromosome paired differences in extracellular proportion; R^2, species sample size = 0.46, phylogeny = 0.34; Supplementary Table 2, row 4). Furthermore, we found that the number of genomes per species and the non-independence of species explained 46 and 34% of the variation in data, respectively (paired plasmid and chromosome differences across our 1,632 genomes). Taken together, this illustrates that it is not our data that disagree with previous studies but instead our use of statistical analyses appropriate for multigenome, multispecies datasets[23–25].

These data also illustrate the importance of examining effect sizes and not just whether results are statistically significant. With large sample sizes it is possible to get results that are significant but not biologically important. The percentage of variance explained that is considered biologically significant can depend on the kind of data you are examining and the field of research but a baseline of 5–10% seems reasonable for many areas of evolutionary biology (Supplementary Information Section 1)[43–45]. When bacterial

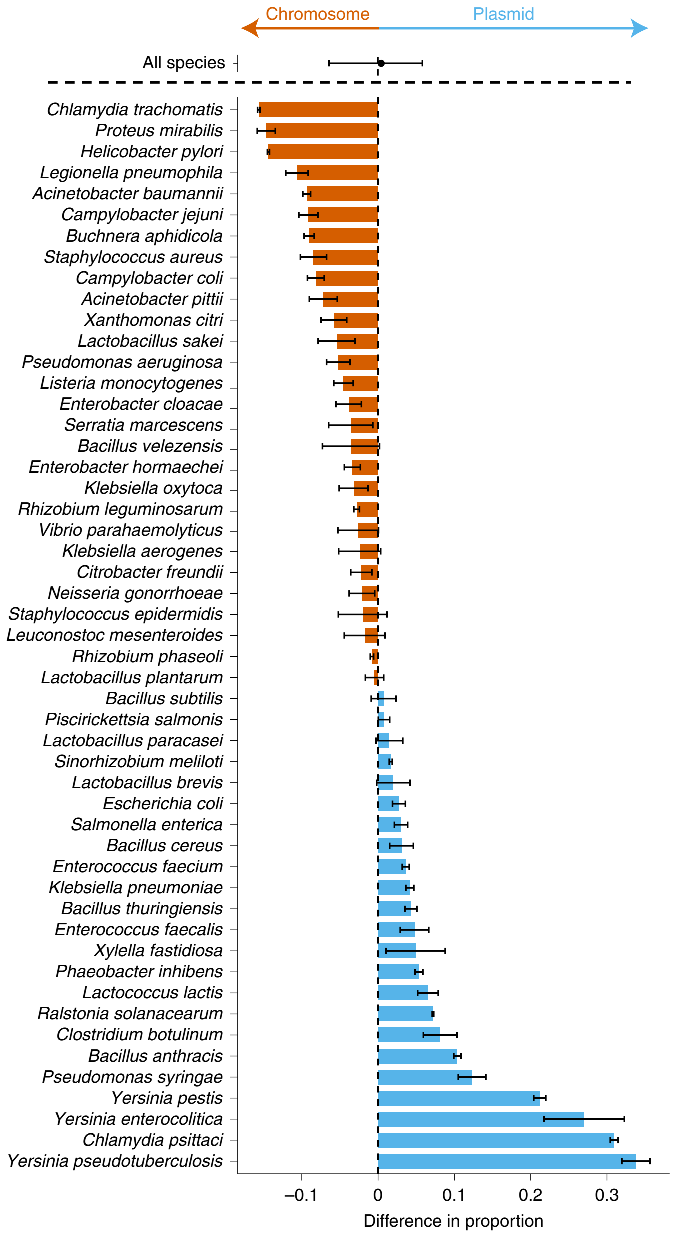


Fig. 2 | Extracellular proteins are not overrepresented on plasmids. For each species we calculated the mean difference between plasmid(s) and chromosomes in the proportion of genes coding for extracellular proteins. Species in blue have a difference greater than zero, meaning their plasmid genes code for a greater proportion of extracellular proteins than do chromosome genes. Species in red have a difference less than zero, meaning their chromosome genes code for a greater proportion of extracellular proteins than do plasmid genes. Error bars indicate the standard error. The dot and error bar at the top of the graph indicate the mean difference and 95% CI given by a MCMCglmm analysis across all species, controlling for phylogeny and sample size. We arcsine square root transformed proportion data before calculating the difference. Overall, there is no consistent trend that genes coding for extracellular proteins are more likely to be carried on plasmids (that is, no consistent trend towards species in blue).

genomes are assumed to be independent data points in across-species analyses, this leads to inflated sample sizes. Consequently, even when results are statistically significant at $P<0.05$, they can still only explain 1–2% of the variation in the data, which is clearly not biologically significant. The flip side of such considerations is that effects sizes and examination of raw data at the species level (for example, Fig. 2) are also useful checks against non-significant results due to a lack of statistical power (type II errors).

Plasmids with higher mobility do not carry more genes for extracellular proteins. We then tested another prediction of the cooperation hypothesis: cooperation is more likely to be favoured when coded for on more mobile plasmids[14–18]. We used data from the MOBsuite database to assign plasmids to one of three levels of mobility (Fig. 3a)[46,47]. We classify: conjugative plasmids, which carry all genes necessary to transfer, as the most mobile; mobilizable plasmids, which are dependent on conjugative plasmids' machinery to transfer, to have intermediate mobility; and non-mobilizable plasmids, which cannot be transferred via conjugation, to be the least mobile (Fig. 3a)[46,48].

Genes coding for extracellular proteins were not more likely to be on plasmids with higher transfer rates (Fig. 3b). Examining the slope of the regression between plasmid mobility and the proportion of genes coding for extracellular proteins, we found no consistent pattern across species (MCMCglmm; posterior mean = 0.006, 95% CI = −0.040–0.052, pMCMC = 0.73; n = 40; Supplementary Table 2, row 11). This lack of a significant relationship was robust to different forms of analysis, including an examination of the means of each mobility type of each species (Supplementary Fig. 4 and Supplementary Table 2, row 12). We also found no correlation between the proportion of a species' plasmids that can transfer and how overrepresented or under-represented extracellular proteins are on plasmids compared to chromosomes (Extended Data Fig. 4 and Supplementary Table 2, rows 16 and 17).

To examine our assumption that mobilizable plasmids are likely to be less mobile than conjugative plasmids, we examined how frequently these two kinds of plasmids co-occurred within a genome. If mobilizable plasmids are present in the same cell as conjugative plasmids, they could be transmitted at similar rates. However, we found that of genomes with a mobilizable plasmid(s), 60% did not also carry a conjugative plasmid (434/727). In addition, when mobilizable plasmids did co-occur with a conjugative plasmid, they did not have a higher proportion of genes coding for extracellular proteins (Supplementary Information Section 1 and Supplementary Fig. 6). A caveat here is that our estimates of transfer rates across different types of plasmid is relative and it would be very useful to obtain quantitative estimates of transfer rates.

Theoretical stability of cooperation. Our empirical results did not support the theoretical prediction that cooperative genes should be overrepresented on plasmids, relative to the chromosome[14–18,49]. Consequently, we then extended existing theory, to examine whether we could find conditions where cooperative genes were not predicted to be overrepresented on plasmids. We investigated the consequences of two factors: (1) allowing for a greater range of possible genetic architectures, especially plasmids that lacked the gene for cooperation (non-cooperative or 'cheat' plasmids); and (2) examining the evolutionary stability (maintenance) of cooperation, not just its initial invasion[16,49].

We examined two possible reasons for why cooperative genes could be overrepresented on plasmids, relative to the chromosome. First, horizontal gene transfer on a plasmid could allow cooperation to be favoured in conditions where it would otherwise not be favoured[14–18]. For example, because plasmid transfer can turn non-cooperators into cooperators and increase relatedness at the loci for cooperation[17]. Second, even if horizontal gene transfer did not increase the range of biological scenarios (parameter space) where cooperation was favoured, there could be selection for cooperation to be coded for on a plasmid, rather than on a chromosome.

We assumed an infinite population of haploid individuals (bacterial cells). Individuals may carry a cooperative gene that

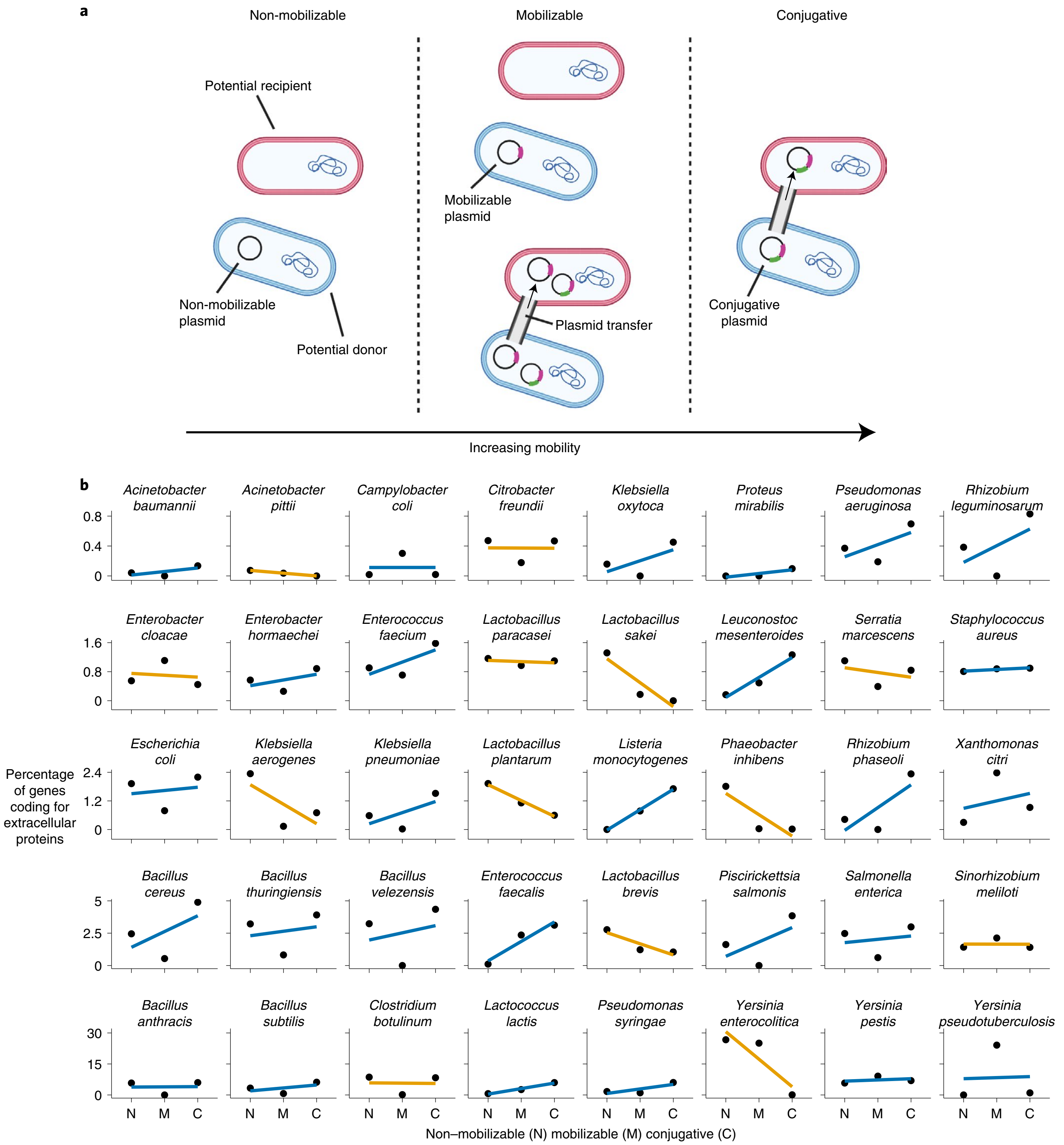


Fig. 3 | Plasmid mobility and extracellular proteins. a, We divided plasmids into three mobility types: non-mobilizable (lowest or no mobility); mobilizable (intermediate mobility); and conjugative (highest mobility). Blue cells are potential plasmid donors, while red cells are potential recipients. Each section shows when plasmid transfer is possible for one of the three plasmid mobility types. Non-mobilizable plasmids cannot be transferred. Mobilizable plasmids cannot be transferred alone but they carry enough genes to 'hijack' the machinery of a conjugative plasmid that is in the same cell. Conjugative plasmids carry all genes necessary to transfer independently. Created with BioRender.com. **b**, The 40 species that carried plasmids of all three mobilities are shown, with a graph for each of these species. Dots in each graph indicate the mean percentage of genes coding for extracellular proteins of all plasmids of each mobility level. The lines are the linear regression of these three points, coloured blue if the slope is positive and orange if the slope is negative. Note that each row of species has a different *y* axis scale, indicated on the left, which applies to all species in that row. We arcsine square root transformed proportion data before calculating the mean for each species and then back-transformed these values for display of the data. Overall, there is no consistent trend for genes that code for extracellular proteins to be on more mobile plasmids.

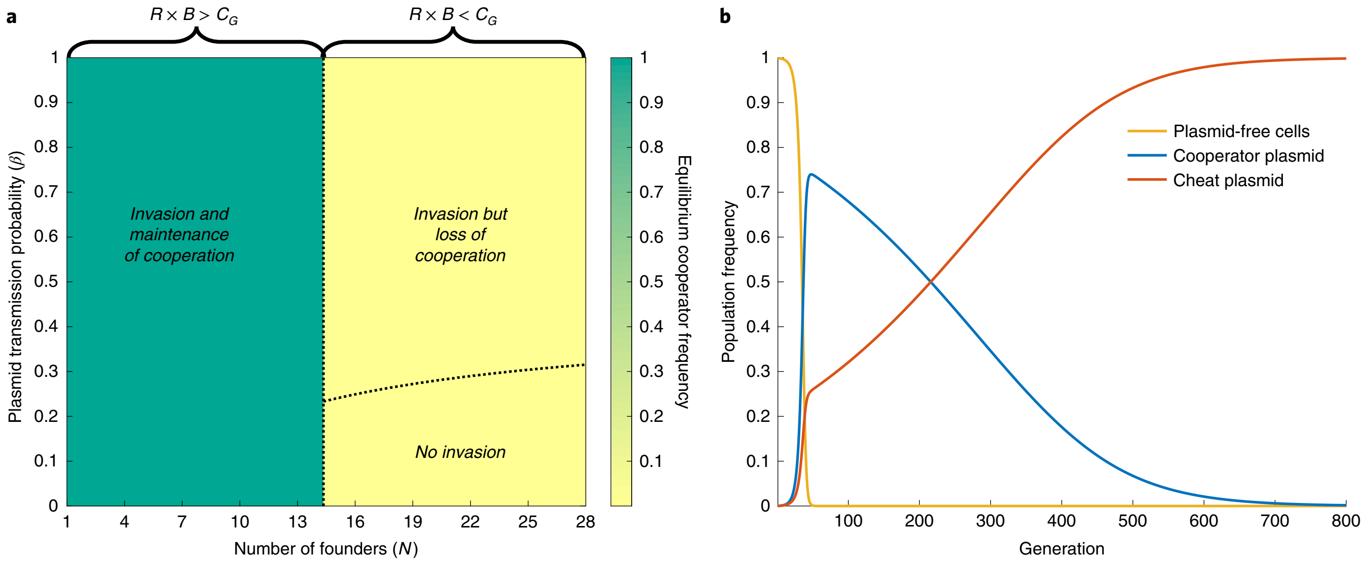


Fig. 4 | Plasmids facilitate the invasion but not the maintenance of cooperation. a,**b**, The results of our theoretical model for the case when there is no plasmid loss ($s = 0$): cooperation is only maintained at equilibrium (green shaded area) when it is favoured at the chromosomal level $R \times B > C_G$, which is unaffected by plasmid transfer (β) (**a**); plasmids can facilitate the invasion and initial spread of cooperation (blue line shoots above red line) but cooperative plasmids are eventually outcompeted by cheat plasmids (red line goes to 1) (**b**). We note that, in **b**, all individuals are chromosomal defectors—chromosomal cooperation was permitted but did not evolve in this run. To generate the plots in **a** and **b**, we assumed the following parameter values: for **a** and **b**, $B = 1.435$, $C_G = 0.1$, $C_C = 0.2$; **b**, $\beta = 0.5$, $N = 16$.

codes for public goods production, on a plasmid or the chromosome or both (redundancy). We also allowed for the possibility of: non-cooperative plasmids and chromosomes; plasmid-free cells; a cost of plasmid carriage (C_C).

Each generation, the population is divided into patches, each founded by N independent cells. Cells reproduce clonally until there is a large number of cells per patch. Cells are then randomly shuffled into pairs on their patch and, if a plasmid-free individual has a plasmid-bearing partner, with probability β, the plasmid-free individual acquires a copy of its partner's plasmid (horizontal gene transfer). Individuals with a gene for cooperation then produce a public good, at a cost C_G, which generates a benefit B that is shared between all members of the patch. Individuals then survive according to their fitness. Plasmid-bearing individuals lose their plasmid with probability s. Finally, individuals disperse to found new patches.

Cooperation invasion. Consistent with previous analyses, we found that, in the short term, horizontal gene transfer on a plasmid can initially help cooperation invade (Fig. 4)[14–18]. Horizontal gene transfer increased the frequency of cooperation, by turning non-cooperators into cooperators, which also increases relatedness at the cooperative locus on the plasmid[14–18,49]. Relatedness is increased because, in the short term, whilst plasmids are spreading from rarity, there are many plasmid-free cells available, meaning plasmids have many opportunities to be transferred, generating genetic similarity.

Cooperation stability. By contrast, we found that transfer on a plasmid did not appreciably increase the range of parameter space where cooperation was maintained at evolutionary equilibrium (Figs. 4a and 5 and Supplementary Information Section 4). First, in the absence of plasmid loss ($s = 0$), cooperation was only favoured when $R \times B - C_G > 0$, where R is the genetic relatedness at the chromosomal (individual) level ($R = 1/N$). Cooperation was therefore only favoured on the plasmid when it provided a kin selected benefit at the level of the chromosome (individual), as predicted by Hamilton's rule[50,51].

The reason for this result is that, in the absence of plasmid loss ($s = 0$), plasmids continue to increase in frequency after invasion, ultimately reaching fixation in the population. This means that, in the long-term, there are no plasmid-free individuals left to infect, which means that the overall level of horizontal gene transfer in the population goes to zero. Consequently, competition between plasmids with and without a cooperative gene (cooperators and cheats) becomes analogous to the scenario in which the gene for cooperation is on the chromosome[17].

Second, when plasmids can be lost ($s > 0$), this can favour cooperation on plasmids but only in certain areas of parameter space (Fig. 5). Plasmid loss means that plasmids do not reach fixation in the population and so some plasmid transfer still occurs in the evolutionary long-term, increasing relatedness at the cooperative plasmid locus. This increased relatedness may favour cooperation on the plasmid, when it would not otherwise be favoured on the chromosome, if plasmids are transferred rapidly (high β) and rates of plasmid loss are intermediate (Fig. 5). Specifically, plasmids need to be lost quickly enough that plasmid relatedness appreciably deviates from chromosomal relatedness but not too quickly that plasmids are not maintained (Fig. 5). Another factor that might prevent plasmids from reaching fixation is if there was a constant, high influx of plasmid-free cells (immigration).

Overall, our model suggests that horizontal gene transfer can help cooperation initially invade but will then often have less influence on whether cooperation is maintained in the long-term (Figs. 4 and 5). We are not saying that horizontal gene transfer can never favour cooperation, just that there is an appreciable area of parameter space where it does not. Consequently, our model provides an explanation for why cooperative genes are not consistently overrepresented on plasmids (Figs. 2 and 3). An analogous theoretical result for the case without plasmid loss ($s = 0$) was also found in a meta-population model by Mc Ginty et al.[16]. Our predictions are consistent with experiments carried out by Bakkeren et al.[30], who found that location on a conjugative plasmid could help a cooperative trait invade in *Salmonella enterica* serovar Typhimurium

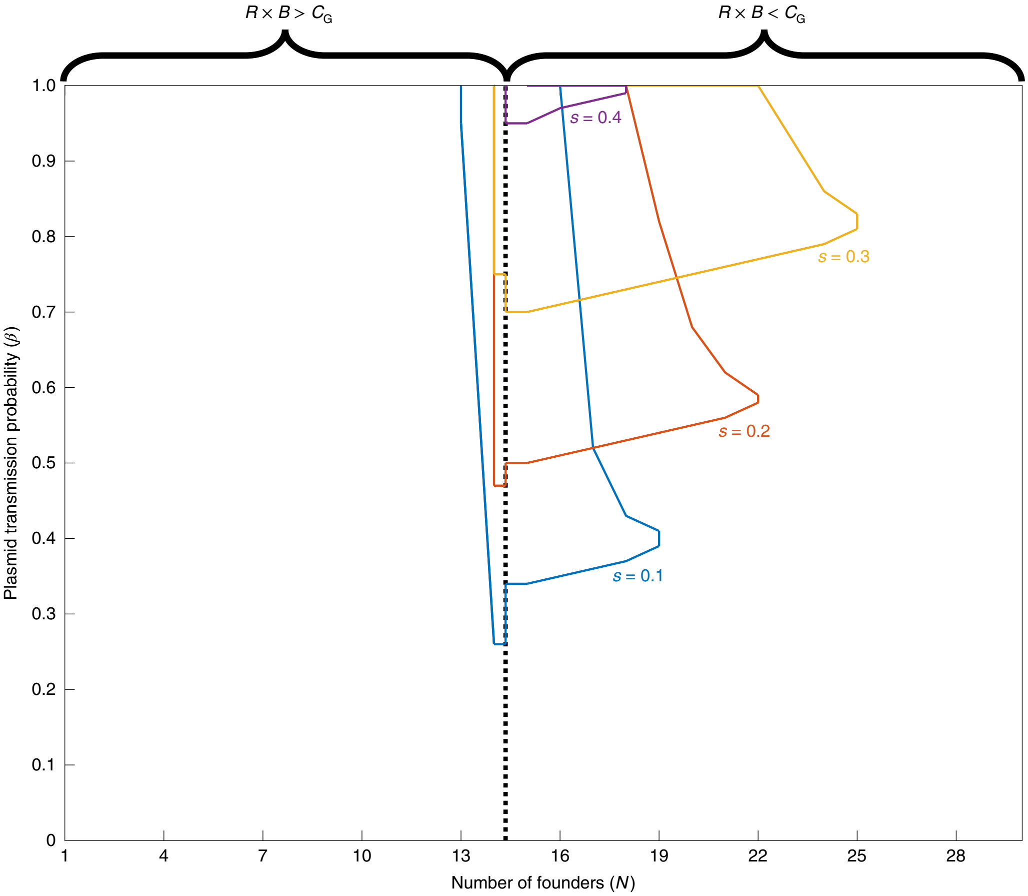


Fig. 5 | Plasmid loss can favour the maintenance of cooperation. We plot the results of our theoretical model for different levels of plasmid loss ($s = 0$–1). The areas encapsulated by the coloured lines show the regions of parameter space where cooperation is polymorphic at equilibrium (that is, population comprises some cooperators and some defectors). When plasmid loss is absent ($s = 0$), there is no polymorphism (encapsulated area collapses to nothing), meaning cooperation is only maintained at equilibrium (at fixation) when it is favoured at the chromosomal level $R \times B > C_G$ (to the left of the black dotted line) ($R = 1/N$). When plasmid loss is intermediate ($s = 0.1$, 0.2, 0.3, 0.4), cooperation can be polymorphic at equilibrium (encapsulated areas), with cooperation being disfavoured in the encapsulated areas to the left of the black dotted line and favoured in the encapsulated areas to the right of the black dotted line, relative to when plasmids are absent ($\beta = 0$). When plasmid loss is high ($s \geq 0.5$) or when transmission (β) is low, plasmids fail to persist at equilibrium, meaning they have no long-term effect on cooperation (encapsulated areas collapse to nothing). Overall, plasmid loss can facilitate cooperation but only if plasmid loss (s) is intermediate and transmission (β) is high. To generate this plot, we assumed the following parameter values: $B = 1.435$, $C_G = 0.1$, $C_C = 0.2$ (same as Fig. 4).

(*S*.Tm) but that this was only stable with strong population bottlenecks (high relatedness). Dimitriu et al.[18] found that cooperative plasmids were favoured in structured but not well-mixed populations and that cooperation was favoured more during 'epidemic spreads' into a population.

In addition, we found that, when cooperation is favoured, cooperative traits are not more likely to be favoured on, or transferred to, plasmids. The reason is that, when cooperation is favoured, non-cooperators (cheats) are purged from the population, which means there is no extra fitness benefit of coding for the cooperative trait on a plasmid rather than the chromosome. Consequently, our results suggest that horizontal gene transfer only favours cooperation in a restricted area of parameter space. Although, there could be interesting transient dynamics, with cooperation being favoured temporarily (Fig. 4) or when cooperation has other consequences, such as increasing plasmid transmission[52,53]. Another important factor is the rate of horizontal gene transfer. While plasmids clearly transmit fast enough to influence evolution, the transfer rates per cell per generation might not be high enough to significantly influence relatedness at the locus for cooperation (that is, a high enough β)[54].

Alternate hypotheses. Finally, we examined whether alternate hypotheses may better explain the considerable variation in the location of genes coding for extracellular proteins across species. Species that live in more variable environments may be more likely to carry extracellular genes on plasmids. This could be expected for different reasons, including plasmid transfer allowing genes for different environments to be gained and lost (Fig. 1b) or plasmids conferring some other advantage not associated with horizontal gene transfer, such as allowing copy number to be conditionally adjusted (Fig. 1c)[31,32,38,55]. A number of different ways can be used to classify environmental variability and so we used three different methods.

Broad host-range pathogens are most likely to carry genes for extracellular proteins on plasmids. We first used the diversity of pathogen hosts as a proxy for environmental variability. Although this does not capture all environmental variability experienced by species in our dataset, pathogenicity is a key aspect of bacterial lifestyle that has been suggested to be important for plasmid gene content, such as antibiotic resistance and virulence factors[6,40,56,57]. We divided species into three categories: pathogens with broad host-range, pathogens with narrow host-range and non-pathogens. Broad host-range pathogens are expected to encounter more variable environments than narrow host-range pathogens.

We found that pathogens with a broad host-range were more likely to carry genes coding for extracellular proteins on their plasmids, compared with both narrow host-range pathogens and non-pathogens (Fig. 6a). Specifically, we compared the difference in the proportion of genes coding for extracellular proteins between plasmid(s) and chromosome(s) across these three categories of species (MCMCglmm; narrow compared to broad host-range pathogens: posterior mean = −0.222, 95% CI = −0.322 to −0.123, pMCMC = <0.001; non-pathogens compared to broad host-range pathogens: posterior mean = −0.161, 95% CI = −0.252 to −0.067, pMCMC = <0.001; *n* = 701 genomes; R^2 of pathogenicity/host-range = 0.35, R^2 of species sample size = 0.28, R^2 of phylogeny = 0.11; Supplementary Table 2, row 23). There was no significant difference between narrow host-range pathogens and non-pathogens in the proportion of genes coding for extracellular proteins on their plasmids compared to chromosome(s) (MCMCglmm; non-pathogens compared to narrow host-range pathogens: posterior mean = 0.031, 95% CI = −0.065 to 0.127, pMCMC = 0.482; *n* = 389; Supplementary Table 2, row 25). These patterns hold irrespective of whether we included species that we could not reliably classify into either category, such as opportunistic pathogens, in our analyses (Extended Data Fig. 5).

Plasmids of broad host-range pathogens carry many pathogenicity genes. We suspected that the additional extracellular proteins coded for by plasmids of broad host-range species, compared to narrow host-range species, may be particularly involved in facilitating pathogenicity[40,56,57]. To investigate this, we used the programme MP3 (ref. [58]) to assign each extracellular protein as either 'pathogenic' or 'non-pathogenic'.

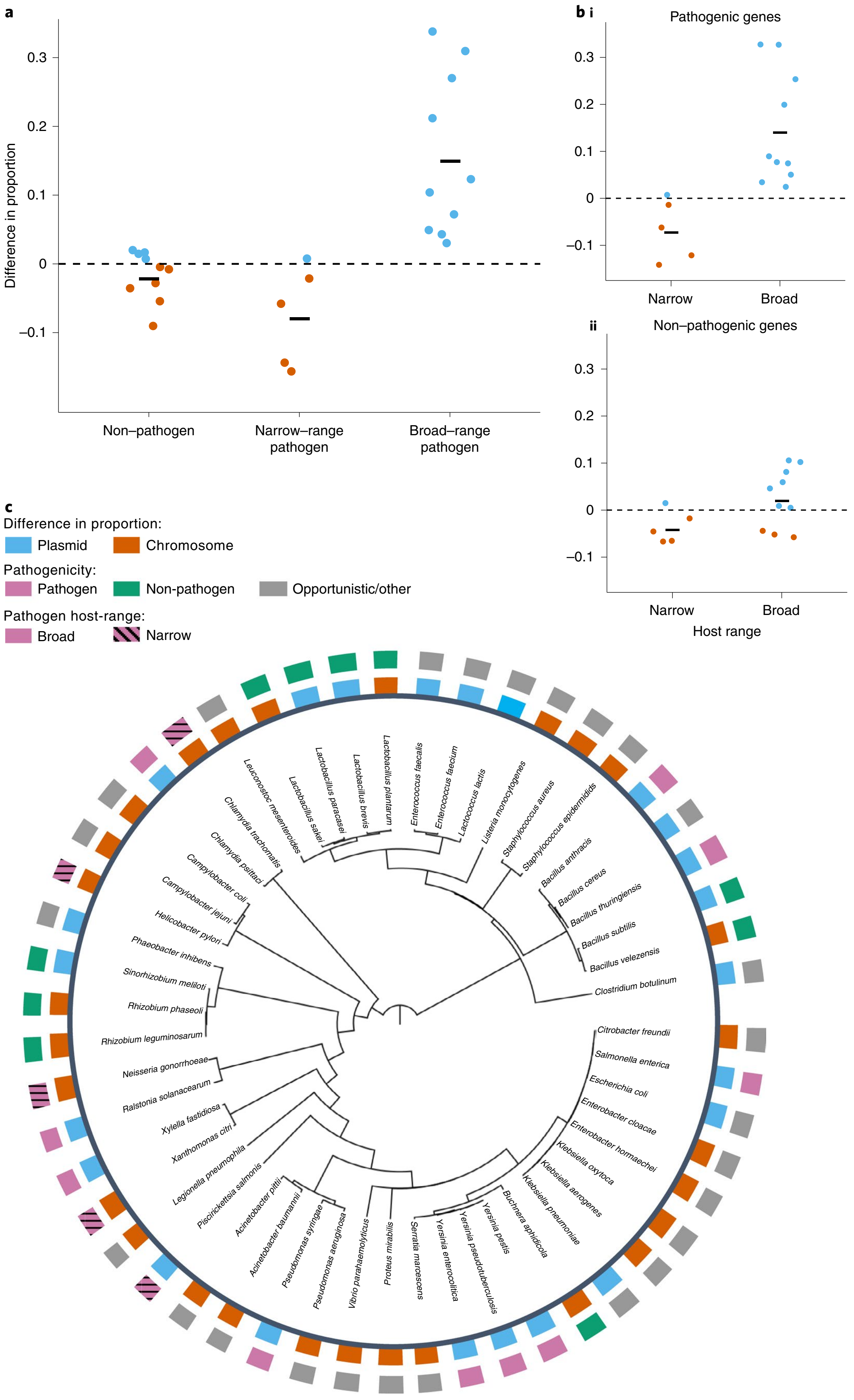
We found that plasmids of broad host-range pathogens were particularly enriched with extracellular proteins involved in facilitating pathogenicity, compared to plasmids of narrow host-range species (Fig. 6b(i)). Specifically, we found that pathogens with a broad host-range were significantly more likely to code for pathogenic extracellular proteins on their plasmids compared to narrow host-range species (Fig. 6b(i)) (MCMCglmm; narrow compared to broad host-range pathogens: posterior mean = −0.209, 95% CI = −0.350 to −0.086, pMCMC = 0.012; *n* = 474 genomes; Supplementary Table 2, row 26). By contrast, the relative location of non-pathogenic extracellular proteins did not vary between broad and narrow host-range pathogens (Fig. 6b(ii)) (MCMCglmm; narrow compared to broad host-range pathogens: posterior mean = −0.036, 95% CI = −0.115 to 0.040, pMCMC = 0.296; *n* = 474 genomes; Supplementary Table 2, row 27). Consequently, the excess of genes coding for extracellular proteins on the plasmids of broad host-range species (Fig. 6a) appears to arise due to an excess of pathogenicity genes coding for extracellular proteins (Fig. 6b).

Most genomic databases are biased towards species that interact with and/or infect humans, so we examined whether human pathogens had driven the above results. In our dataset, five out of ten broad host-range species and three out of five narrow host-range species can infect humans. We found no significant difference in how likely both pathogenic and non-pathogenic extracellular proteins were to be on plasmids of human pathogens compared to non-human pathogens. We also found that while host-range had a significant effect on how likely plasmids were to code for pathogenic extracellular proteins, whether a species could infect humans had no significant effect (Supplementary Table 2, rows 28 to 30).

Pathogenic extracellular proteins could be preferentially coded for on plasmids to facilitate their gain and loss (Fig. 1b, gain and loss hypothesis) or because of some other benefit provided by being carried on a plasmid (Fig. 1c, beyond horizontal gene transfer hypothesis). We tested these possibilities by examining whether pathogenic extracellular proteins were more likely to be on plasmids that transfer at higher rates. This would be predicted by the gain and loss hypothesis but not the beyond horizontal gene transfer hypothesis. We found that plasmids with higher mobility did not code for more pathogenic extracellular proteins. Specifically, across broad host-range pathogen species, the slope of the regression between plasmid mobility and the proportion of genes coding for pathogenic extracellular proteins was not consistently positive (Supplementary Fig. 7) (MCMCglmm; posterior mean = −0.020, 95% CI = −0.224 to 0.185, pMCMC = 0.774; *n* = 7; Supplementary Table 2, row 31). This lack of a significant relationship was robust to additional forms of analysis, such as considering all pathogenic species, including narrow host-range pathogens and those not carrying plasmids of all three mobility types (Supplementary Fig. 8 and Supplementary Table 2, rows 32 and 33).

Taken together, our results are most consistent with the hypothesis that genes coding for extracellular proteins are overrepresented on plasmids when plasmid carriage provides a benefit other than mobility (Fig. 1c). A number of other factors may influence which genes are carried on plasmids, beyond horizontal gene transfer. First, there is evidence that increasing the copy number of plasmids can lead to increasing rates of evolution in the genes they carry[59] and it also may act as a mechanism to increase the expression of genes carried on plasmids[60,61]. For example, increased expression of genes coding for extracellular public goods such as virulence factors could help invasion of a host and utilization of host resources. This could be particularly beneficial for broad host-range pathogens that frequently invade a variety of different hosts. Copy number of plasmids has also recently been shown to lead to genetic dominance effects[55], with likely implications for the phenotypes of genes selected for plasmid carriage[55]. Second, plasmids compete with their bacterial hosts for resources such as replication machinery and nucleotides[62,63]. To resolve this competition, plasmids should be under selection to reduce their cost to the host, with a likely impact

Fig. 6 | Pathogenicity, host-range and the location of genes coding for extracellular proteins. We have divided species into either pathogens or non-pathogens, with pathogens further categorized into those with a narrow or broad host-range. **a**, The *y* axis shows the difference in the proportion of genes on plasmids and chromosomes coding for extracellular proteins—this is the same as the *x* axis in Fig. 2—for non-pathogen, narrow host-range pathogen and broad host-range pathogen species. **b**, The *y* axes show the difference in the proportion of a subset of genes coding for extracellular proteins on plasmids and chromosomes which are predicted by MP3 as either (i) pathogenic or (ii) non-pathogenic, for narrow and broad host-range pathogen species. Each dot is the mean for all genomes in a species. Species in blue are those with the relevant subset of extracellular proteins overrepresented on plasmids, while species in red are those with the subset of extracellular proteins overrepresented on chromosomes. **c**, Phylogeny based on recently published maximum likelihood tree using 16S ribosomal protein data[80]. The inner ring indicates whether extracellular proteins were more likely to be coded for on the plasmid(s) or chromosome(s), as in Fig. 2. The outer ring indicates how we classified each species' pathogenicity and the presence or absence of diagonal lines for pathogens indicates narrow or broad host-range, respectively. Species with a pink or green label in the outer ring are those included in **a** and **b**, since for these we could be reasonably confident of whether or not pathogenicity was an important and consistent aspect of their lifestyle. Overall, pathogens with a broad host-range are more likely to have genes coding for extracellular proteins, and particularly those involved in pathogenicity, on their plasmids.



on their gene content. For example, extracellular proteins are, on average, cheaper to produce than are intracellular proteins[15,20]. Plasmid–host competition could consequently select for plasmids to carry more genes coding for cheaper proteins and so more extracellular proteins. Our conclusion here should be seen as tentative, as some form of the gain and loss hypothesis (Fig. 1b) could still be argued to be consistent with the data, if it is just the potential for horizontal gene transfer that matters and not the rate.

Number of environments and core versus accessory genes. To further examine a potential association with environmental variability, as could be predicted by both hypotheses b ('gain and loss') and c ('beyond horizontal gene transfer'), we also looked at two additional measures of environmental variability: (1) the number of five broad environments a species was sequenced from[20,64,65]; (2) the proportion of a species' genomes that is composed of 'core' genes, which are those found in all genomes of the species—species that experience more variable environments appear to have relatively smaller core genomes[32]. We found no significant correlation between either of these measures and the likelihood that genes coding for extracellular proteins were carried on plasmids (Extended Data Fig. 6, Supplementary Information Section 1 and Supplementary Table 2, rows 35 and 37). Garcia-Garcera and Rocha[20] previously analysed a different but related question, examining the type of environment and also used a MCMCglmm to control for the phylogenetic structure of the data (Supplementary Information Section 1). Our finding of no correlation between these two measures of environmental variability and whether plasmids code for extracellular proteins is in contrast to our above results with respect to pathogen host-range (Fig. 6). This suggests that hypothesis c, which our data are most consistent with, may be important for pathogens in particular but not necessarily across all bacterial species and lifestyles.

Complementary analyses. Our analyses could be expanded in a number of directions. We focused on plasmids because they have been the focus of previous theoretical and empirical work[14,16–18]. Other mobile genetic elements include bacteriophages and integrative conjugative elements[66,67]. Comparing core and accessory genes could be a potential way to lump all causes of horizontal gene transfer[15,19]. We considered the relative transfer rates among mobility types; quantitative estimates of plasmid transfer rates would be very useful for further examination of plasmid mobility[48,54,68–70]. We followed previous genomic studies by using extracellular proteins as indicators of cooperative traits[2,15,19,20]. The advantages of this approach are that: (1) we could compare our results with those from previous studies; (2) secretion systems are highly conserved, allowing us to examine a large number of species, where detailed genetic annotations are lacking; and (3) cooperation mediated by extracellular proteins is usually controlled by only one gene, making them potentially more suitable for plasmid carriage compared to cassettes of multiple genes[71,72]. However, while extracellular proteins are likely to be cooperative traits, not all cooperative genes code for extracellular proteins (for example, secondary metabolites such as siderophores) and not all extracellular proteins are involved in cooperation (for example, those involved in motility such as flagellin). It would be very useful to examine more detailed annotations of social genes and expand to other mobile genetic elements.

Discussion

We found no support for the hypothesis that horizontal gene transfer generally favours cooperation. Our genomic analyses showed that extracellular proteins are: (1) not overrepresented on plasmids compared to chromosomes (Fig. 2); and (2) not more likely to be carried by plasmids that transfer at higher rates (Fig. 3). These patterns could be explained by our theoretical modelling, which showed that while horizontal gene transfer may help cooperation to initially invade a population, it has less influence on the maintenance of cooperation in the long-term (Figs. 4 and 5). Once plasmids become common, cheat plasmids that do not code for cooperation are able to outcompete cooperative plasmids, analogous to selection at the level of the chromosome[16,30]. Our results suggest that horizontal gene transfer on plasmids has not consistently favoured cooperation across bacterial species—but it is still possible that horizontal gene transfer could have an influence in certain scenarios or species. By contrast, we found that genes coding for extracellular proteins involved in pathogenicity and virulence are preferentially located on plasmids in pathogens with a broad host-range (Fig. 6). These pathogenic virulence genes were not preferentially located on plasmids that transfer at a higher rate, suggesting that the benefit of being located on a plasmid is something other than horizontal gene transfer, such as the ability to vary copy number.

Methods

Genome collection. We retrieved 1,632 complete genomes comprising 51 bacterial species from GenBank RefSeq (https://www.ncbi.nlm.nih.gov) between February and November 2019. We used species on panX (http://pangenome.tuebingen.mpg.de)[73] as a list of potential species for our dataset, since these comprise the most sequenced bacterial species. To allow comparison of chromosome and plasmid genes within the same genome, we only retrieved genomes that contained at least one plasmid sequence. We included species with ten or more RefSeq genomes with one or more plasmids available in our analysis. We retrieved up to 100 genomes for each species; this was either all complete genomes available for the species or a random sample where >100 were available. Where two or more genomes had the same strain name, we randomly retrieved one genome to reduce the risk of pseudoreplication.

Prediction of subcellular location of proteins. We used PSORTb v.3 (ref. [39]) to predict the subcellular location of every protein encoded by each genome in our dataset. We used a Docker image of PSORTb developed by the Brinkman Lab, available at: https://github.com/brinkmanlab/psortb_commandline_docker. We chose PSORTb because it is widely regarded as one of the best-performing programmes of its kind[74]. It has also been used in previous analyses to identify 'cooperative' genes and/or extracellular proteins in bacteria[15,20]. The programme has a number of modules that are trained to recognize particular features of proteins. Results from these modules are combined to give a final prediction for each protein. We consulted the literature to confirm the Gram stain of each of our species. For Gram-positive species, PSORTb assigns proteins to one of four locations within the cell: cytoplasmic, cytoplasmic membrane, extracellular or cell wall (Extended Data Fig. 1). The locations for Gram-negative species are the same, except that cell wall is replaced with outer membrane and periplasmic, meaning that there are five possible locations for proteins of Gram-negative species (Extended Data Fig. 1). We used these predicted locations throughout all subsequent analyses in this work. PSORTb could not reliably assign a subcellular location to 27% of proteins we analysed, giving a final prediction of 'unknown' (Supplementary Table 1). Unless explicitly stated, we did not include these unknown proteins in our analyses.

Predicting plasmid mobility. We also predicted the mobility of every plasmid in our dataset using the MOB-typer tool of the programme MOBsuite[46]. This searches for features of plasmid sequences including the origin of transfer (oriT), relaxase and mating-pair formation to give each plasmid one of three mobility predictions: (1) conjugative, where plasmids encode all machinery required to transfer via conjugation; (2) mobilizable, where plasmids do not encode all machinery but encode oriT and/or relaxase, allowing them to 'hijack' another plasmid's conjugation machinery and mobilize; and (3) non-mobilizable, where plasmids do not encode the genes necessary to be mobilized by themselves or other plasmids and so cannot transfer via conjugation. A total of 628 of the 4,150 plasmids in our dataset were flagged as 'unverified' against the MOBsuite dataset, meaning their mobility prediction was unreliable and they were not included. This left 3,522 plasmids for subsequent analysis.

Effect of mobility on plasmid extracellular protein content. We next examined how plasmid mobility correlates with each plasmid's extracellular protein proportion. As part of its mobility prediction, MOBsuite[46] identifies sequences within each plasmid involved with conjugation. To control for the possibility that conjugative plasmids, by definition of being conjugative, must carry genes controlling this process, we subtracted the total number of these sequences from the total number of proteins when calculating the extracellular proportion of each plasmid. This is a highly conservative control, since it assumes none of the proteins predicted as extracellular are involved in conjugation. We did all analyses on these data with and without removing these mating-pair accessions to ensure any results were not affected by factors unrelated to plasmids' extracellular protein content.

Additionally, we used the plasmid mobility predictions to ask whether differences in the mobility of species' plasmids correlated with whether genes encoding extracellular proteins are overrepresented on plasmids compared to chromosomes. We calculated the proportion of plasmids in each genome capable of transferring via conjugation (conjugative and mobilizable plasmids) and averaged across all genomes to give a general measure of the mobility of each species' plasmids.

Measures of bacterial lifestyle and environmental variability. We classified a species as pathogenic if it was described in the literature as an obligate or facultative pathogen. Given that some bacterial species only rarely act as pathogens, such as opportunistic pathogens, we only included species where we could be sure pathogenicity was a key aspect of their lifestyle and a regular selection pressure acting on their genome content. For this reason, we decided not to include species described as opportunistic pathogens in the literature and those that frequently live as commensals in their hosts. We classified non-pathogens as species that are strictly environmental (never live in hosts) or strictly mutualists and/or commensals (never cause pathogenicity in their hosts). There were 26 species we could not definitively assign to either of these categories. These were not included in our main analyses, although we carried out additional analyses to ensure that removing these species did not bias our results (Extended Data Fig. 5).

To estimate the host-range of pathogens, we used information from the literature to determine the maximum taxonomic level of hosts each species is able to invade. We defined narrow host-range species as those that can invade either only one host species or host species within the same genus or family. By contrast, we defined broad host-range pathogens as those capable of invading host species within the same order, class or phylum. For example, *Xanthomonas citri* acts as a plant pathogen within the genus *Citrus*[75], while *Pseudomonas syringae* acts as plant pathogen across multiple orders of flowering plants[76]. For more details and references to the literature used for this classification, see Supplementary Table 3.

We completed additional analyses for another two measures and proxies of environmental variability, the details and results of which can be found in Supplementary Information Section 1. In brief, we used previously published data which classified the habitat diversity of species using 16S rRNA environmental datasets across five broad habitats: water, wastewater, sediment, soil and host[64,65]. We also supplemented this with information from the literature for species not included in the published data. We used this to ask whether species that lived in multiple habitats had genes encoding extracellular proteins more overrepresented on their plasmids.

We also looked at bacterial pangenomes as a proxy for environmental variability, since it has been noted that species with a high percentage of accessory genes, defined as genes found in only a subset of genomes within a species, are generally those with more variable environments. All pangenome data were collected from panX (ref. [73]; http://pangenome.tuebingen.mpg.de), since this calculates the pangenome using the same method across all of our species.

Pathogenicity categorization of extracellular proteins. We used MP3 (ref. [58]) to examine the pathogenicity of extracellular protein-coding genes in broad host-range and narrow host-range pathogens. MP3 compares protein sequences to a curated dataset of proteins known to be involved in various aspects of pathogenicity: adhesion, invasion, secretion and resistance[58]. MP3 uses two modules to produce a 'hybrid' prediction for each protein: either 'pathogenic' or 'non-pathogenic'. We used MP3 with default parameters to gain this prediction for every extracellular protein in all genomes of broad and narrow host-range species. MP3 was unable to give a prediction for ~9% of extracellular proteins and so these were not included in this analysis.

For each genome in broad and narrow host-range pathogens, we summed the MP3 predictions to give the total number of 'pathogenic' and 'non-pathogenic' extracellular proteins on the chromosome and on the plasmid(s). We then calculated the proportions of plasmid and chromosome genes that code for 'pathogenic' and 'non-pathogenic' extracellular proteins.

Statistical analyses. *MCMCglmm.* Many commonly used statistical methods in biology require data points to be independent from one another. However, due to shared ancestry, species cannot be considered as independent data points[24]. Recently developed statistical methods now allow for phylogenetic relationships to be controlled for within mixed effects models. For all statistical analyses we used the MCMCglmm package in R with phylogeny as a random effect[41,77]. This means the phylogeny is implemented in the model as a covariance matrix of the relationships between species, which is controlled for when considering whether patterns exist across species[41,77]. We also included sample size as a random effect when analysing at the genome level to control for differences in the number of genomes per species. Specific details of each model can be found in Supplementary Table 2. We extracted from each model the posterior mean, 95% credible intervals (functionally similar to 95% confidence intervals) and the pMCMC value (generally interpreted in a similar way to a P value). We also calculated R^2 values for models of particular interest using methods described in refs. [78,79]. A detailed description of MCMCglmm can be found elsewhere[41,77].

The response variable in all of our analyses is either a proportion or a measure calculated from proportions. Proportion data are bound between 0 and 1 and have a non-normal distribution. To control for this, all proportion data in our analyses have been arcsine square root transformed to improve normality.

Phylogeny. To control for species relationships, we generated a phylogeny including all 51 species in our dataset (Supplementary Fig. 1). We used a recently published maximum likelihood tree using 16S ribosomal protein data as the basis for our phylogeny[80]. This tree of life typically had only one representative species per genus. We used the R package 'ape' to extract all branches matching species in our dataset[81]. In cases where the genus representative was different to the species in our dataset, we swapped the tip name with our species, since all members of the same genus are equally related to members of a sister genus. In cases where we had multiple species within a single genus in our dataset, we used the R package 'phylotools' to add these species as additional branches into their genus[82]. We used published phylogenies from the literature to add any within-genus clustering of species' branches. We used this phylogeny in nexus format for all our MCMCglmm analyses (Supplementary Fig. 1 and Supplementary Table 2). Methods are also available to control for uncertainty in phylogenetic reconstruction[83,84], although we have not done this here.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The dataset of genomes analysed during this study, including PSORTb results and plasmid mobility predictions of MOBsuite, will be made available in the public repository Dryad at: https://doi.org/10.5061/dryad.gxd2547n4

Code availability

Code used to solve equations in the theoretical modelling section of the paper can be found at: https://github.com/ThomasWilliamScott/Plasmid_cooperation.git

Received: 19 February 2021; Accepted: 27 September 2021;
Published online: 08 November 2021

References

1. Foster, K. R. in *Social Be*haviour (eds Szekely, T. et al.) 331–356 (Cambridge Univ. Press, 2010). https://doi.org/10.1017/CBO9780511781360.027
2. McNally, L., Viana, M. & Brown, S. P. Cooperative secretions facilitate host range expansion in bacteria. *Nat. Commun.* **5**, 4594 (2014).
3. West, S. A., Griffin, A. S., Gardner, A. & Diggle, S. P. Social evolution theory for microorganisms. *Nat. Rev. Microbiol.* **4**, 597–607 (2006).
4. Simonet, C. & McNally, L. Kin selection explains the evolution of cooperation in the gut microbiota. *Proc. Natl Acad. Sci. USA* **118**, e2016046118 (2021).
5. Griffin, A. S., West, S. A. & Buckling, A. Cooperation and competition in pathogenic bacteria. *Nature* **430**, 1024–1027 (2004).
6. Hale, T. L. Genetic basis of virulence in *Shigella* species. *Microbiol. Rev.* **55**, 206–224 (1991).
7. Dinges, M. M., Orwin, P. M. & Schlievert, P. M. Exotoxins of *Staphylococcus aureus*. *Clin. Microbiol. Rev.* **13**, 16–34 (2000).
8. Diggle, S. P., Griffin, A. S., Campbell, G. S. & West, S. A. Cooperation and conflict in quorum-sensing bacterial populations. *Nature* **450**, 411–414 (2007).
9. Jones, S. et al. The lux autoinducer regulates the production of exoenzyme virulence determinants in *Erwinia carotovora* and *Pseudomonas aeruginosa*. *EMBO J.* **12**, 2477–2482 (1993).
10. Sandoz, K. M., Mitzimberg, S. M. & Schuster, M. Social cheating in *Pseudomonas aeruginosa* quorum sensing. *Proc. Natl Acad. Sci. USA* **104**, 15876–15881 (2007).
11. Ghoul, M., Griffin, A. S. & West, S. A. Toward an evolutionary definition of cheating. *Evolution* **68**, 318–331 (2014).
12. Butaitė, E., Baumgartner, M., Wyder, S. & Kümmerli, R. Siderophore cheating and cheating resistance shape competition for iron in soil and freshwater *Pseudomonas* communities. *Nat. Commun.* **8**, 414 (2017).
13. Thomas, C., Nielsen, K., Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721 (2005).
14. Smith, J. The social evolution of bacterial pathogenesis. *Proc. R. Soc. Lond. B* **268**, 61–69 (2001).
15. Nogueira, T. et al. Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Curr. Biol.* **19**, 1683–1691 (2009).
16. Mc Ginty, S. E., Rankin, D. J. & Brown, S. P. Horizontal gene transfer and the evolution of bacterial cooperation: mobile elements and bacterial cooperation. *Evolution* **65**, 21–32 (2011).
17. Mc Ginty, S. É., Lehmann, L., Brown, S. P. & Rankin, D. J. The interplay between relatedness and horizontal gene transfer drives the evolution of plasmid-carried public goods. *Proc. R. Soc. B* **280**, 20130400 (2013).
18. Dimitriu, T. et al. Genetic information transfer promotes cooperation in bacteria. *Proc. Natl Acad. Sci. USA* **111**, 11103–11108 (2014).
19. Nogueira, T., Touchon, M. & Rocha, E. P. C. Rapid evolution of the sequences and gene repertoires of secreted proteins in bacteria. *PLoS ONE* **7**, e49403 (2012).
20. Garcia-Garcera, M. & Rocha, E. P. C. Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. *Nat. Commun.* **11**, 758 (2020).

21. Kruskal, W. Miracles and statistics: the casual assumption of independence. *J. Am. Stat. Assoc.* **83**, 929–940 (1988).
22. Ives, A. R. & Zhu, J. Statistics for correlated data: phylogenies, space, and time. *Ecol. Appl.* **16**, 20–32 (2006).
23. Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15 (1985).
24. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, 1991).
25. Grafen, A. The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B.* **326**, 119–157 (1989).
26. Hurlbert, S. H. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54**, 187–211 (1984).
27. Ruxton, G. & Colegrave, N. *Experimental Design for the Life Sciences* (Oxford Univ. Press, 2011).
28. Stone, G. N., Nee, S. & Felsenstein, J. Controlling for non-independence in comparative analysis of patterns across populations within species. *Philos. Trans. R. Soc. B* **366**, 1410–1424 (2011).
29. Ives, A. R., Midford, P. E. & Garland, T. Jr. Within-species variation and measurement error in phylogenetic comparative methods. *Syst. Biol.* **56**, 252–270 (2007).
30. Bakkeren, E. et al. Cooperative virulence can emerge via horizontal gene transfer but is stabilized by transmission. Preprint at *bioRxiv* https://doi.org/10.1101/2021.02.11.430745 (2021).
31. Ghoul, M., Andersen, S. B. & West, S. A. Sociomics: using omic approaches to understand social evolution. *Trends Genet.* **33**, 408–419 (2017).
32. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040 (2017).
33. Niehus, R., Mitri, S., Fletcher, A. G. & Foster, K. R. Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nat. Commun.* **6**, 8924 (2015).
34. Cordero, O. X. et al. Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. *Science* **337**, 1228–1231 (2012).
35. Rakoff-Nahoum, S., Coyne, M. J. & Comstock, L. E. An ecological network of polysaccharide utilization among human intestinal symbionts. *Curr. Biol.* **24**, 40–49 (2014).
36. Nocelli, N., Bogino, P. C., Banchio, E. & Giordano, W. Roles of extracellular polysaccharides and biofilm formation in heavy metal resistance of rhizobia. *Materials* **9**, 418 (2016).
37. Ciofu, O., Beveridge, T. J., Kadurugamuwa, J., Walther-Rasmussen, J. & Høiby, N. Chromosomal β-lactamase is packaged into membrane vesicles and secreted from *Pseudomonas aeruginosa*. *J. Antimicrob. Chemother.* **45**, 9–13 (2000).
38. Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R. C. & San Millán, Á. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat. Rev. Microbiol.* **19**, 347–359 (2021).
39. Yu, N. Y. et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**, 1608–1615 (2010).
40. Rankin, D. J., Rocha, E. P. C. & Brown, S. P. What traits are carried on mobile genetic elements, and why? *Heredity* **106**, 1–10 (2011).
41. Hadfield, J. D. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33**, 1–22 (2010).
42. Clutton-Brock, T. H. & Harvey, P. H. Primate ecology and social organization. *J. Zool.* **183**, 1–39 (1977).
43. Jennions, M. D. & Møller, A. P. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behav. Ecol.* **14**, 438–445 (2003).
44. Crawley, M. J. *Statistics: An Introduction Using R* (John Wiley & Sons, 2014).
45. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Routledge, 1988).
46. Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genom.* **4**, e000206 (2018).
47. Robertson, J., Bessonov, K., Schonfeld, J. & Nash, J. H. E. Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance. *Microb. Genom.* **6**, mgen000435 (2020).
48. Smillie, C., Garcillan-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & de la Cruz, F. Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434–452 (2010).
49. Mc Ginty, S. É. & Rankin, D. J. The evolution of conflict resolution between plasmids and their bacterial hosts. *Evolution* **66**, 1662–1670 (2012).
50. Hamilton, W. D. Genetical evolution of social behaviour I & II. *J. Theor. Biol.* **7**, 1–52 (1964).
51. Hamilton, W. D. The evolution of altruistic behavior. *Am. Nat.* **97**, 354–356 (1963).
52. Ghigo, J. M. Natural conjugative plasmids induce bacterial biofilm development. *Nature* **412**, 442–445 (2001).
53. Di Venanzio, G. et al. Multidrug-resistant plasmids repress chromosomally encoded T6SS to enable their dissemination. *Proc. Natl Acad. Sci. USA* **116**, 1378–1383 (2019).
54. 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 (2020).
55. Rodríguez-Beltrán, J. et al. Genetic dominance governs the evolution and spread of mobile genetic elements in bacteria. *Proc. Natl Acad. Sci. USA* **117**, 15755–15762 (2020).
56. Cornelis, G. R. et al. The virulence plasmid of yersinia, an antihost genome. *Microbiol. Mol. Biol. Rev.* **62**, 1315–1352 (1998).
57. Köstlbacher, S., Collingro, A., Halter, T., Domman, D. & Horn, M. Coevolving plasmids drive gene flow and genome plasticity in host-associated intracellular bacteria. *Curr. Biol.* **31**, 346–357 (2021).
58. Gupta, A., Kapil, R., Dhakan, D. B. & Sharma, V. K. MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PLoS ONE* **9**, e93907 (2014).
59. San Millan, A., Escudero, J. A., Gifford, D. R., Mazel, D. & MacLean, R. C. Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat. Ecol. Evol.* **1**, 0010 (2016).
60. Carrier, T., Jones, K. L. & Keasling, J. D. mRNA stability and plasmid copy number effects on gene expression from an inducible promoter system. *Biotechnol. Bioeng.* **59**, 666–672 (1998).
61. Rodríguez-Beltrán, J. et al. Multicopy plasmids allow bacteria to escape from fitness trade-offs during evolutionary innovation. *Nat. Ecol. Evol.* **2**, 873–881 (2018).
62. Dietel, A.-K., Kaltenpoth, M. & Kost, C. Convergent evolution in intracellular elements: plasmids as model endosymbionts. *Trends Microbiol.* **26**, 755–768 (2018).
63. Rocha, E. P. C. & Danchin, A. Base composition bias might result from competition for metabolic resources. *Trends Genet.* **18**, 291–294 (2002).
64. Garcia-Garcera, M., Touchon, M., Brisse, S. & Rocha, E. P. C. Metagenomic assessment of the interplay between the environment and the genetic diversification of *Acinetobacter*. *Environ. Microbiol.* **19**, 5010–5024 (2017).
65. Kümmerli, R., Schiessl, K. T., Waldvogel, T., McNeill, K. & Ackermann, M. Habitat structure and the evolution of diffusible siderophores in bacteria. *Ecol. Lett.* **17**, 1536–1544 (2014).
66. Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L. & Brüssow, H. Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.* **6**, 417–424 (2003).
67. Burrus, V. & Waldor, M. K. Shaping bacterial genomes with integrative and conjugative elements. *Res. Microbiol.* **155**, 376–386 (2004).
68. O'Brien, F. G. et al. Origin-of-transfer sequences facilitate mobilisation of non-conjugative antimicrobial-resistance plasmids in *Staphylococcus aureus*. *Nucleic Acids Res.* **43**, 7971–7983 (2015).
69. Rodríguez-Rubio, L. et al. Extensive antimicrobial resistance mobilization via multicopy plasmid encapsidation mediated by temperate phages. *J. Antimicrob. Chemother.* **75**, 3173–3180 (2020).
70. Ramsay, J. P. & Firth, N. Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Curr. Opin. Microbiol.* **38**, 1–9 (2017).
71. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 3801–3806 (1999).
72. 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**, 1481–1489 (2011).
73. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* **46**, e5 (2018).
74. Gardy, J. L. & Brinkman, F. S. L. Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* **4**, 741–751 (2006).
75. Ference, C. M. et al. Recent advances in the understanding of *Xanthomonas citri* ssp. *citri* pathogenesis and citrus canker disease management. *Mol. Plant Pathol.* **19**, 1302–1318 (2018).
76. Morris, C. E., Lamichhane, J. R., Nikolić, I., Stanković, S. & Moury, B. The overlapping continuum of host range among strains in the *Pseudomonas syringae* complex. *Phytopathol. Res* **1**, 4 (2019).
77. Hadfield, J. D. *MCMCglmm Course Notes* (2019); https://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf
78. Nakagawa, S. & Schielzeth, H. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **4**, 133–142 (2013).
79. Nakagawa, S., Johnson, P. C. D. & Schielzeth, H. The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R. Soc. Interface* https://doi.org/10.1098/rsif.2017.0213 (2017).
80. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
81. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
82. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
83. Washburne, A. D. et al. Methods for phylogenetic analysis of microbiome data. *Nat. Microbiol.* **3**, 652–661 (2018).
84. Som, A. Causes, consequences and solutions of phylogenetic incongruence. *Brief. Bioinform.* **16**, 536–548 (2015).

Acknowledgements

We thank C. MacLean, K. Foster, S. Brown, L. Belcher, C. Hao and especially E. Rocha for their helpful comments, and R. Goldberg and V. Pike for their support with MCMCglmm. We thank J. Robertson for providing plasmid mobility data from the MOBsuite database. We thank the BBSRC (BB/M011224/1: A.E.D.), ERC (SESE: J.L.T., A.S.G. and M.G.; 834164: T.W.S and S.A.W.) and NSERC-CRSNG of Canada (G.W.) for funding. Conceptual figures were created with BioRender.com.

Author contributions

A.E.D., J.L.T., A.S.G., S.A.W. and M.G. conceived the genomic analyses and interpreted results. A.E.D. and J.L.T. collected and analysed the genomic data and A.E.D. produced the corresponding statistical analyses and figures. T.W.S, G.W. and S.A.W. conceived the theoretical modelling and interpreted results. T.W.S. completed the formal theoretical modelling. A.E.D., J.L.T., T.W.S., S.A.W. and M.G. wrote and/or edited the manuscript. A.E.D. wrote and put together Supplementary Sections 1, 2 and 3 and T.W.S. wrote and put together Supplementary Section 4. All authors commented on and approved the manuscript for submission.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-021-01573-2.

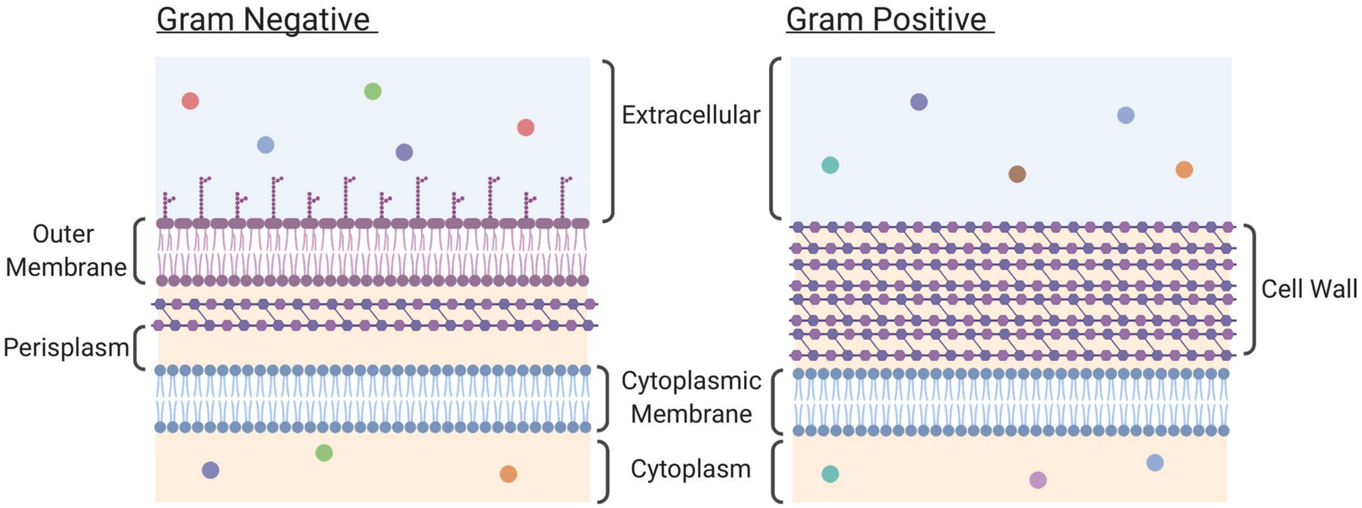
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-021-01573-2.

Correspondence and requests for materials should be addressed to Anna E. Dewar.

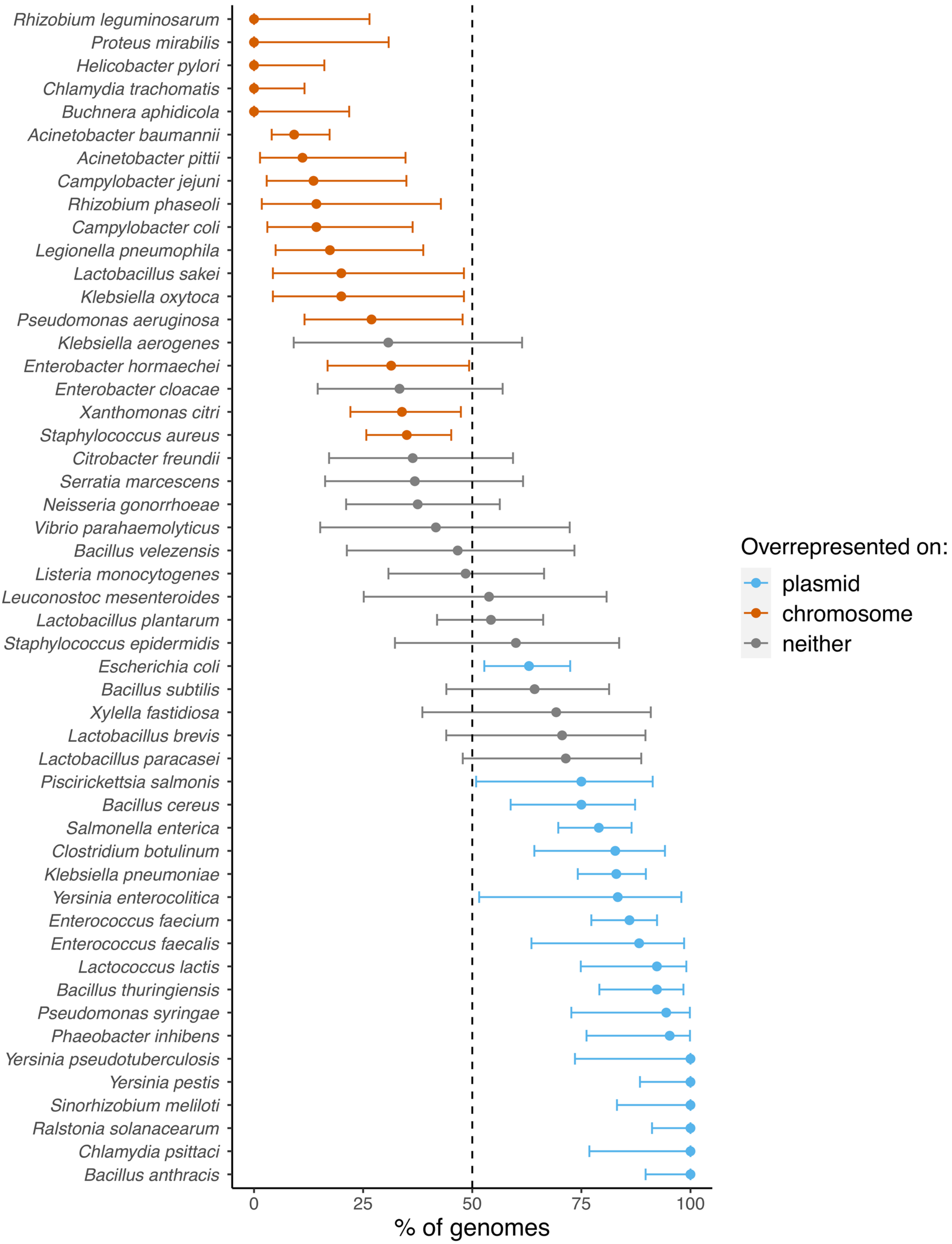
Peer review information *Nature Ecology & Evolution* thanks Isabel Gordo, Alex Washburne and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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.

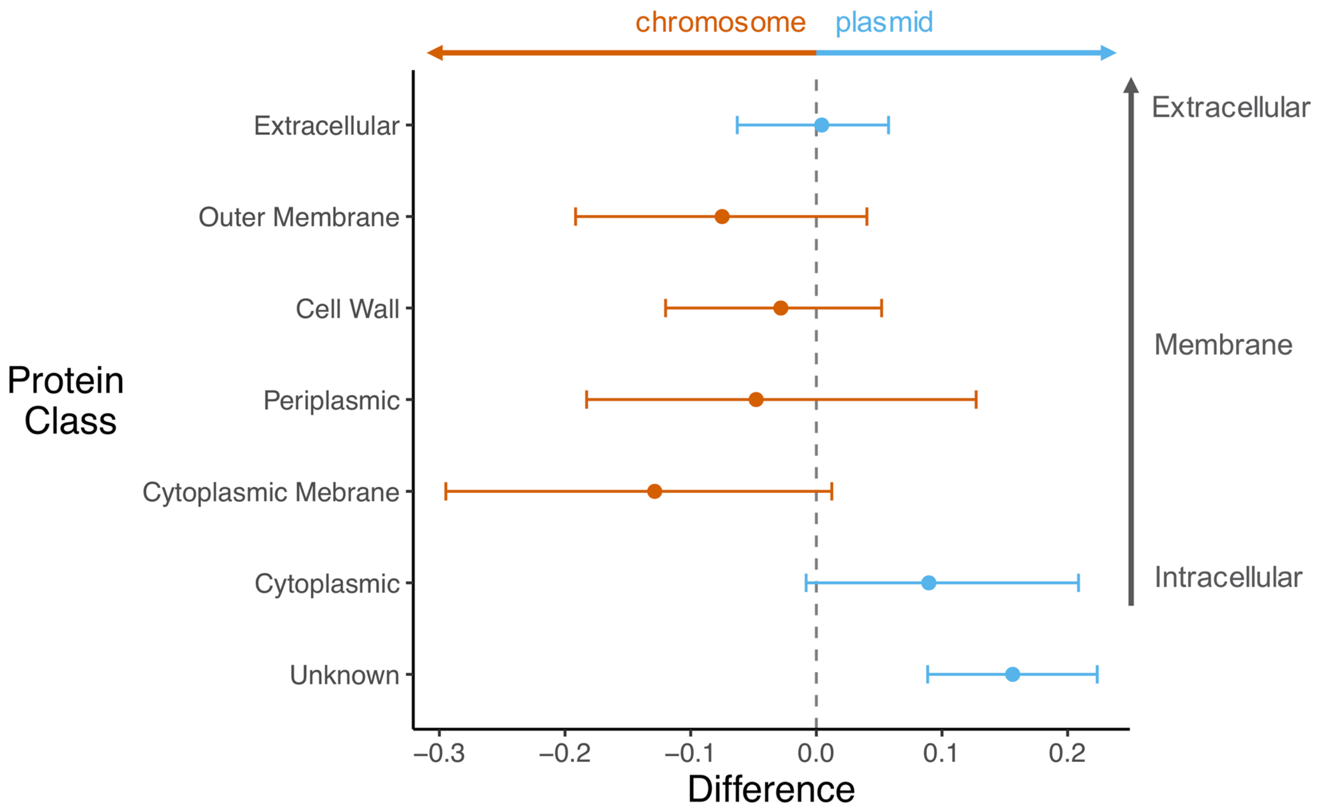


Extended Data Fig. 1 | Protein subcellular localizations. Visualization of all possible subcellular locations predicted by PSORTb. The left panel shows a cross-section of a typical Gram-negative bacterium and the right panel shows the equivalent for a Gram-positive bacterium. Both kinds of bacteria have an inner membrane, known as the cytoplasmic membrane. The main difference is that Gram-positive bacteria are surrounded by a thick layer of a molecule called peptidoglycan, while Gram-negative bacteria have a much thinner layer of peptidoglycan, and have an additional membrane. Created with BioRender.com.

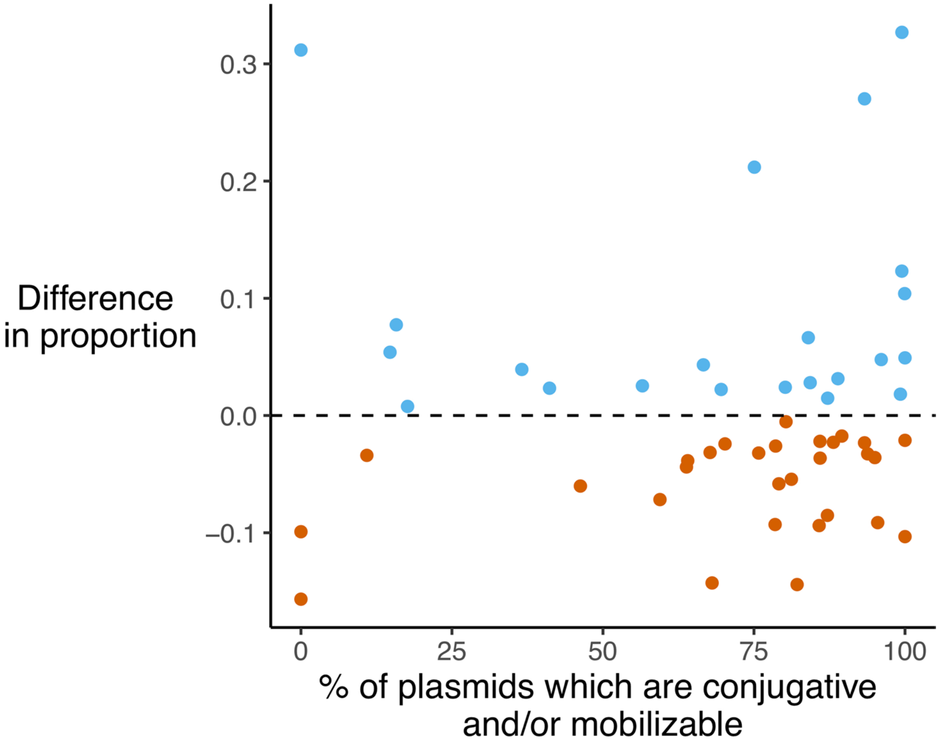


Extended Data Fig. 2 | See next page for caption.

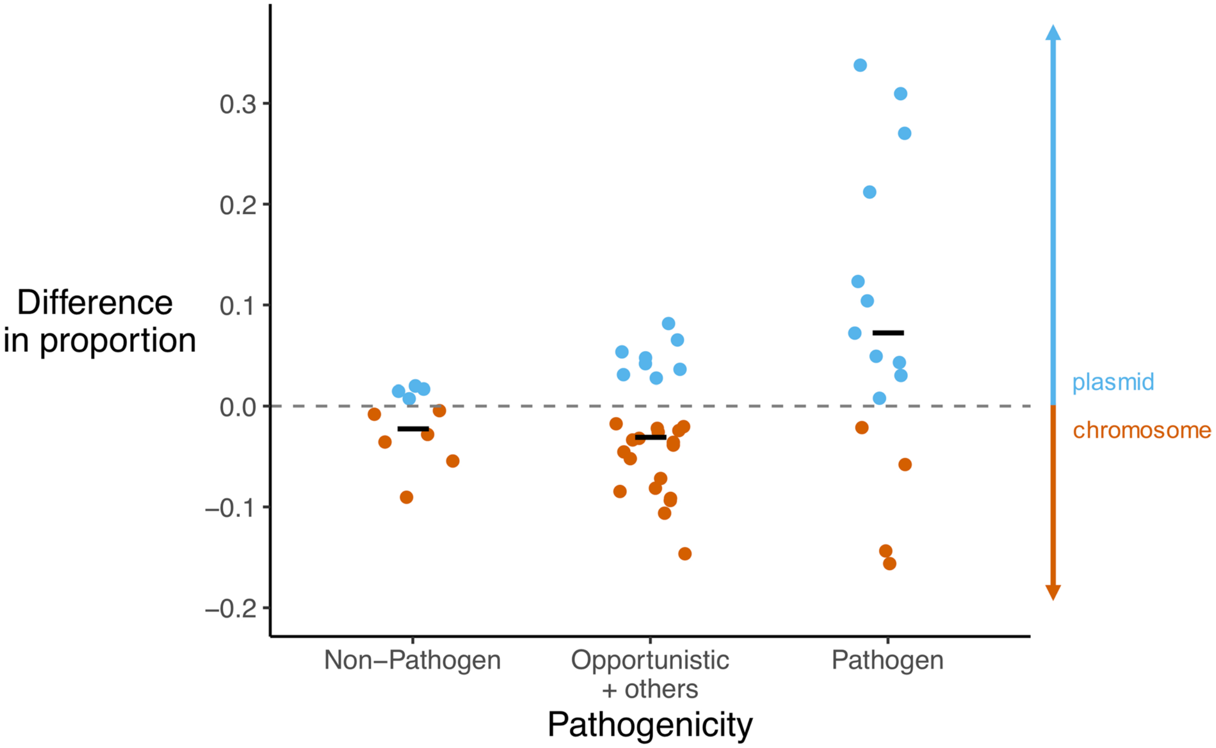
Extended Data Fig. 2 | Substantial variation within and between species in the genomic location of extracellular proteins. The x-axis is the % of genomes in each species where the proportion of plasmid proteins predicted as extracellular is greater than the proportion of chromosome proteins predicted as extracellular. Crucially, this considers only whether the plasmid proportion is greater than the chromosome proportion for each genome, rather than also considering the magnitude of the difference (Fig. 2). Error bars are the 95% Confidence Intervals from a binomial test on each species, comparing the number of genomes which have plasmid proportion > chromosome proportion to a null prediction of 50% of genomes. Species in blue have >50% of genomes where plasmid > chromosome extracellular proportion, meaning extracellular proteins are significantly over-represented on plasmids. Species in red have <50% of genomes where plasmid > chromosome extracellular proportion, meaning extracellular proteins are significantly over-represented on chromosomes. Species in grey have a 95% CI which overlaps 50%, so extracellular proteins are not significantly over-represented on either plasmids or chromosomes in these species.



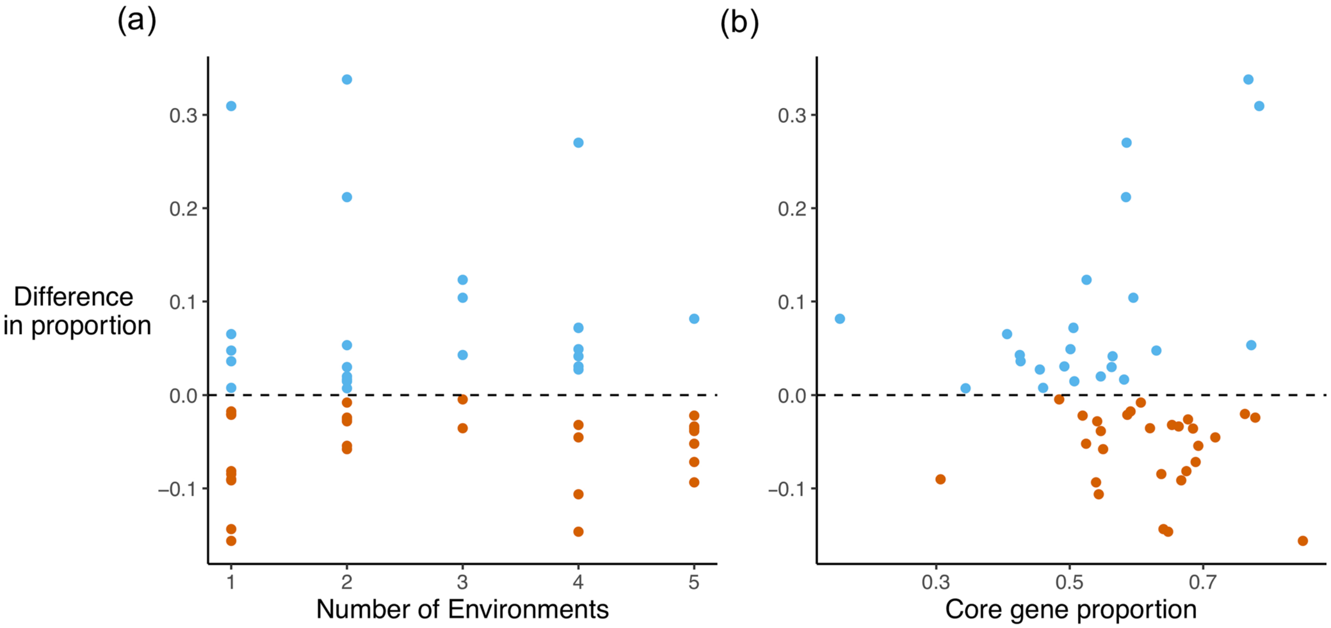
Extended Data Fig. 3 | Difference in plasmid and chromosome proportion for all protein classes predicted by PSORTb. The x-axis is the difference in plasmid and chromosome extracellular proportions, as in Fig. 2. The y-axis is all possible subcellular locations predicted by PSORTb. These protein 'classes' are ordered along the y-axis by location within the cell, from intracellular to increasingly extracellular. Each dot is the posterior mean and 95% Credible Intervals from a MCMCglmm[42] on the difference in plasmid and chromosome proportion across all species, accounting for phylogeny and sample size. The only proteins significantly over-represented in either direction are unknown proteins, which make up a higher proportion of plasmid proteins in all species we analysed.



Extended Data Fig. 4 | No effect of plasmid mobility on the difference in plasmid and chromosome proportion of genes coding for extracellular proteins. The x-axis is the % of a species' plasmids which are conjugative or mobilizable. The y-axis shows the difference in the plasmid and chromosome proportions of genes coding for extracellular proteins, as in Fig. 2. Each dot is the mean for all genomes in a species. Species in blue are those with genes coding for extracellular proteins over-represented on plasmids, while species in red have genes coding for extracellular proteins over-represented on chromosomes.



Extended Data Fig. 5 | No difference in where extracellular proteins are coded for in pathogens compared to non-pathogens. The y-axis shows the difference in the plasmid and chromosome proportion of genes coding for extracellular proteins. Each dot is the mean for all genomes in a species. Species in blue are those with genes coding for extracellular proteins over-represented on plasmids, while species in red have genes coding for extracellular proteins over-represented on chromosomes. Species were categorized as pathogens or non-pathogens; those we could not classify as either are shown in the 'Opportunistic + others' category. The black bars indicate the mean for all species in each category.



Extended Data Fig. 6 | Additional measures of environmental variability. We used two additional methods to estimate the environmental variability encountered by these species. (a) The x-axis shows published data on the number of five broad environments each species was recorded in, which we supplemented with information from the literature to include all species. (b) The x-axis shows the proportion of each species' genes which are 'core' genes, meaning they are found in all members of the species. The y-axis in both graphs shows the difference in the proportion of genes on plasmids and chromosomes coding for extracellular proteins. Each dot is the mean for all genomes in a species. Species in blue are those with extracellular proteins over-represented on plasmids, while species in red are those with extracellular proteins over-represented on chromosomes. For both these measures, we found no significant correlation with the genomic location of genes coding for extracellular proteins across species.

nature portfolio

Corresponding author(s): Anna Dewar

Last updated by author(s): Sep 10, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Programs used: prediction of protein subcellular location with PSORTb v3, via a Docker image developed by the Brinkman Lab (https://github.com/brinkmanlab/psortb_commandline_docker); mobility data from MOBsuite database (https://github.com/phac-nml/mob-suite); pathogenicity of proteins using MP3(http://metagenomics.iiserb.ac.in/mp3/index.php); pipeline using python v3.7.2 for data extraction and combining of datasets. Code used to solve equations in the theoretical modelling section of the paper can be found at: https://github.com/ThomasWilliamScott/Plasmid_cooperation.git
Data analysis	All data analysis in R v 3.5.2. Packages used: MCMCglmm v2.29; ape 5.3; phylotools v0.2.2, dplyr v1.0.2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The dataset of genomes analysed during this study, including PSORTb results and plasmid mobility predictions of MOBsuite, will be made available in the public repository Dryad when published at the following DOI: https://doi.org/10.5061/dryad.gxd2547n4

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We retrieved between 10 and 100 genomes of 51 bacterial species from GenBank RefSeq. Species were included if found in PanX (http://pangenome.tuebingen.mpg.de) and had at least 10 complete genomes with at least one plasmid sequence annotated. We retrieved up to 100 genomes for each species; this was either all complete genomes available for the species, or a random sample where more than 100 were available. This gave 1632 genomes, each with at least one chromosome and one plasmid sequence.
Data exclusions	No data was excluded. Some of our analyses, such as examining the effect of host-range of pathogens, is on only a subset of species, which were those we could definitively assign to a category. In such cases, this is stated in the study, and we completed additonal analyses to ensure this did not affect our conclusions.
Replication	We collected genomes from species with at least 10 genomes available to ensure any conclusions at a species-level were not based on very few genomes. Each genome is treated as a replicate for the species in our analyses, rather than an independent data point.
Randomization	For species which had more than 100 genomes that fit our criteria (complete & at least one plasmid sequence), we used R to generate 100 random numbers out of the total number of genomes available, and collected the corresponding genomes.
Blinding	All genomes were collected prior to data analysis on which genomes or species were driving any patterns. While blinding was not possible when categorising the pathogenicity and host-range of species, we categorised species using data from the literature with pre-determined criteria to reduce subjectivity.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a | Involved in the study

☒ □ Antibodies
☒ □ Eukaryotic cell lines
☒ □ Palaeontology and archaeology
☒ □ Animals and other organisms
☒ □ Human research participants
☒ □ Clinical data
☒ □ Dual use research of concern

Methods

n/a | Involved in the study

☒ □ ChIP-seq
☒ □ Flow cytometry
☒ □ MRI-based neuroimaging