Horizontal Gene Transfer and Cooperation in Bacteria

Anna E. Dewar

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2021

Declaration

I declare that this thesis was composed by myself and that the work contained herein is my own except where explicitly stated in the text. This work has not been submitted for any degree or professional qualification except as specified.

Anna Dewar

Michaelmas Term 2021

Acknowledgments

First, thank you to my supervisors Stuart West, Melanie Ghoul and Ashleigh Griffin – I couldn't have asked for a better trio of academic role models. I am so grateful for your support and guidance, especially during the past year and a half.

While not an official supervisor, I'd also like to thank Lindsay Turnbull for your mentorship during my seven years in Oxford, and for sharing my enthusiasm in sea slugs, woodpigeons and everything in between.

Next, thank you to the following people whose feedback, discussions and expertise has greatly improved the work contained in this thesis: Joshua Thomas, Thomas Scott, Chunhui Hao, Geoff Wild, Craig MacLean, Kevin Foster, Laurie Belcher, Rebecca Goldberg, Victoria Pike, Asher Leeks, Guy Cooper, Mati Patel, Ming Liu and Sam Levin.

Thank you to everyone in the West and Griffin Labs, I'm so lucky to be part of such a great group of scientists. In particular, thank you to Josh, Chunhui, Tom and Laurie, it's been brilliant to collaborate with you.

Somehow, finishing a PhD during a global pandemic has been much better than it sounds, and that is largely thanks to the support of my amazing friends and family. Léa and Sofia, I could not have asked for better house mates, thank you for the tea, chocolate and friendship. To everyone in my DTP cohort, and especially Vero, Signe, Lois, Tati, Stephen, Nick, Steffi, Josh, Andreas and Vicky, you have made the last four years such a joy. And to Becca and Josh, thank you for being such brilliant conference, seminar and pub companions.

Finally, to Omer and my wonderful Mum, Dad and brother: I am beyond lucky to have your continued support, encouragement and love. I would not be where I am without you.

Publications and Contributions

Chapter 2

The following paper arose from this thesis and is presented in Chapter 2:

- **Dewar A.E**., Thomas J.L., Scott T.W., Wild G., Griffin A.S., West S.A., Ghoul M. (2021) Plasmids do not consistently stabilize cooperation across bacteria, but may promote broad pathogen host-range. *Nature Ecology & Evolution.* In press.
    - 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 S1, S2 and S3, and T.W.S. wrote and put together S4. All authors commented on and approved the manuscript for submission.

While I have presented this work in the format of a single paper, the work I contributed to the paper is equivalent to between three and four thesis chapters. For clarity, if this was not an integrated thesis, Chapter 2 would be split into the following four chapters:

1. Where are genes coding for extracellular proteins located in bacterial genomes? (In collaboration with fellow DPhil student Joshua Thomas)
2. Do more mobile plasmids code for more extracellular proteins?
3. Do alternate hypotheses explain the genomic location of genes for extracellular proteins?
4. Are plasmid genes coding for extracellular proteins more likely to be involved in pathogenicity when comparing broad to narrow host-range pathogen species?

Additionally, the theoretical modelling in Chapter 2 would be have been included in the appendix, rather than the main text, of my thesis if this was not an integrated thesis.

Chapters 3 and 4

The work in Chapters 3 and 4 is my own, supervised by Stuart West and Melanie Ghoul.

- I conceived both Chapters, conducted all analyses, produced all graphs and figures, and wrote both Chapters, all with comments and support from West S.A and Ghoul M.

Chapter 5

Chapter 5 is a collaboration between myself and fellow DPhil student Chunhui Hao, supervised by Stuart West.

- A.E.D, C.H. and S.A.W. conceived of the study. C.H. collected the genomic data. A.E.D. produced the phylogeny which C.H. used for statistical analysis of the data. A.E.D, C.H. and S.A.W. interpreted the data. C.H. produced the figures with input from A.E.D. A.E.D. wrote the majority of the Chapter with input from C.H. and S.A.W.

Appendix

The appendix contains a paper in review and an early draft of a manuscript, both of which I contributed to during my DPhil:

A. Belcher L.J., **Dewar A.E.**, Ghoul M., West S.A. Kin selection for cooperation in natural bacterial populations. *PNAS*. In review.
B. Hao C., **Dewar A.E.**, Ghoul M., West S.A. Are plasmid-carried genes for cooperation less complex? In prep.

Abstract

Bacteria are capable of a wide range of cooperative behaviours. It has been suggested that horizontal gene transfer, which is common in bacteria, could help to stabilise cooperation and prevent the invasion of non-cooperative cheats. Research on this hypothesis has largely focused on plasmids: genetic sequences found across bacteria that can often transfer to other cells. Here, I test two key predictions of this hypothesis across 51 bacterial species. Contrary to these predictions, I find that genes for cooperation are not more likely to be carried on: (1) plasmids compared to chromosomes; (2) more mobile plasmids compared to less mobile plasmids. Next, I explore characteristics of plasmids themselves. First, I examine correlations between three potential 'life-history' traits of plasmids: size, mobility and range. Second, I find that plasmid sequences are consistently enriched with A and T nucleotide bases compared to chromosomes, and explore two hypotheses for why this is the case. Finally, horizontal gene transfer can impact the content of bacterial genomes. To explore these impacts, I test whether bacterial species' genomes become more variable with increasing environmental variability. Overall, in this thesis I consider the evolution of cooperation and horizontal gene transfer in bacteria, and how they may interact.

Contents

Introduction

Since the discovery of bacteria over 300 years ago, we now know that our health, the growth of our crops, and even the stability of the climate, are all dependent upon bacteria (Gest 2004; Burrows *et al.* 2009; Hayat *et al.* 2010; Zhang *et al.* 2015). In additon to their interactions with animals, plants, and the climate, we also now know that bacteria interact with each other. Rather than acting in isolation, as it was largely assumed until the beginning of this century, bacteria have remarkably active social lives (West *et al.* 2006, 2007a; Foster 2010). A social behaviour is one which affects the fitness of both the performer and the recipient of the behaviour (West *et al.* 2007b; Davies *et al.* 2013). When the behaviour increases the fitness of the recipient, and is selected for because of this benefit, this is described as a cooperative behaviour (Hamilton 1963, 1964; West *et al.* 2007b).

In this thesis, I will explore the role of horizontal gene transfer in bacteria, with a particular focus on the evolution of cooperation. Before describing the outline of my thesis, I introduce the concepts of cooperation and horizontal gene transfer in bacteria. I provide only a brief introduction here, because each of my Chapters contains its own introduction.

Cooperation in bacteria

Bacteria are capable of performing a wide array of cooperative behaviours. Many of these cooperative behaviours occur through the extracellular secretion of molecules that act as 'public goods' (West *et al.* 2006, 2007a; Foster 2010). The small size of these molecules can cause them to diffuse away from the producing cell, meaning their effects are shared with neighbouring cells (Kümmerli *et al.* 2009; Mund *et al.* 2017). These molecules are costly to produce, but provide benefits to both the producing cell and its neighbours (Diggle *et al.* 2007; Ghoul *et al.* 2014b). Examples of these benefits include invasion of hosts, breakdown of food sources, and scavenging for scarce but essential nutrients (Griffin *et al.* 2004; Rumbaugh *et al.* 2009; McNally *et al.* 2014; Orsi *et al.* 2018).

Cooperative behaviours lead to the problem that they could potentially be exploited by non-producing 'cheats'. These are individuals who pay no costs, yet are able to reap the benefits provided by those still performing the behaviour (shown as red in Figure 1) (Ghoul *et al.* 2014a). In bacteria, 'cheats' are usually cells which downregulate expression or no longer have a functional copy of the cooperative gene (Cordero *et al.* 2012; Andersen *et al.* 2015; Ghoul *et*

al. 2017). Why cooperation remains stable, despite the threat from non-cooperative cheats, was a major problem for evolutionary biology.

However, kin selection theory tells us that cooperation can remain stable if the benefits of a behaviour are directed preferentially towards individuals who also share the gene for the behaviour (Hamilton 1963, 1964; Ghoul *et al.* 2014a). Hamilton realised that relatedness, defined as the relative genetic similarity of two individuals relative to the population of potential recipients, was key for understanding the evolution of cooperation (Grafen 1985; Davies *et al.* 2013). Hamilton's rule describes the conditions under which an altruistic behaviour, which has no direct benefit to its actor, will be favoured by selection. Specifically, a gene for altruism will spread if $rB - c > 0$, where r is the relatedness between actor and recipient at the altruistic locus, B is the fitness benefit gained by the recipient from the behaviour, and c is the fitness cost to the actor due to the behaviour (Hamilton 1963, 1964). In bacteria, mechanisms that introduce population structure can increase relatedness, because cells will tend to be near related cells. Consequently, individuals who do not carry the gene, and so do not produce the public good, cannot benefit from and cheat the behaviour.

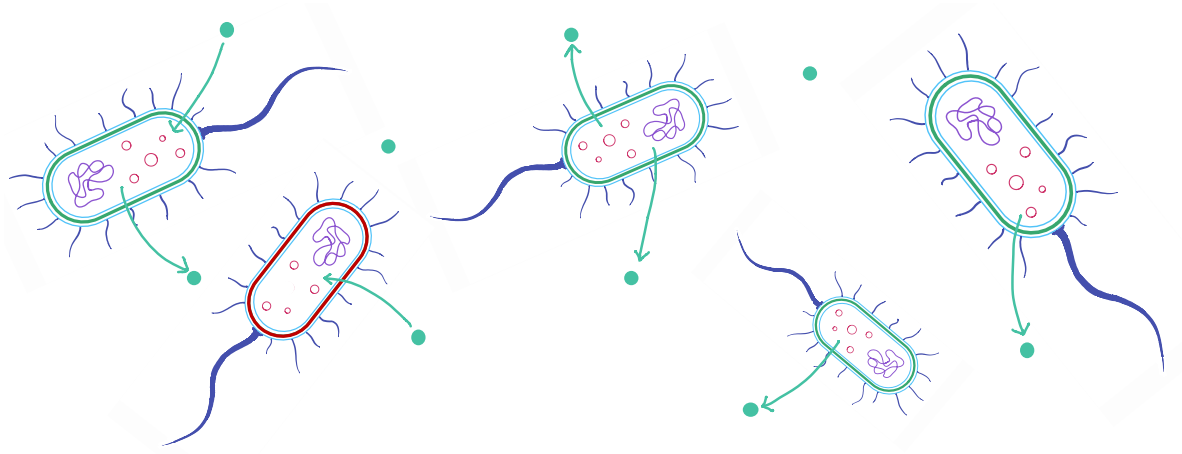


Figure 1. Public goods and cheats

An example of how cheats can exploit cooperation. Green bacteria are co-operators, which produce and take up public goods (green dots). The red bacterium is a cheat, which does not produce yet still takes up public goods.

There is also another potential problem: why would the other 'non-cooperative' genes in the individual be selected to pay a fitness cost to help other individuals? The answer is that because genes are passed on vertically to an organism's descendants, potential recipients that carry the

cooperative gene are also more likely to carry the other genes in the actor's genome (Grafen 1985, 1989). Therefore, the rest of the genome will gain the indirect fitness benefit of helping a related recipient along with the cooperative gene. This 'common ancestry' via vertical inheritance is key to much of social evolution theory, since it leads to more or less equal relatedness at all genomic loci and prevents the other genes from suppressing the cooperative behaviour (Figure 2a) (Grafen 1985).

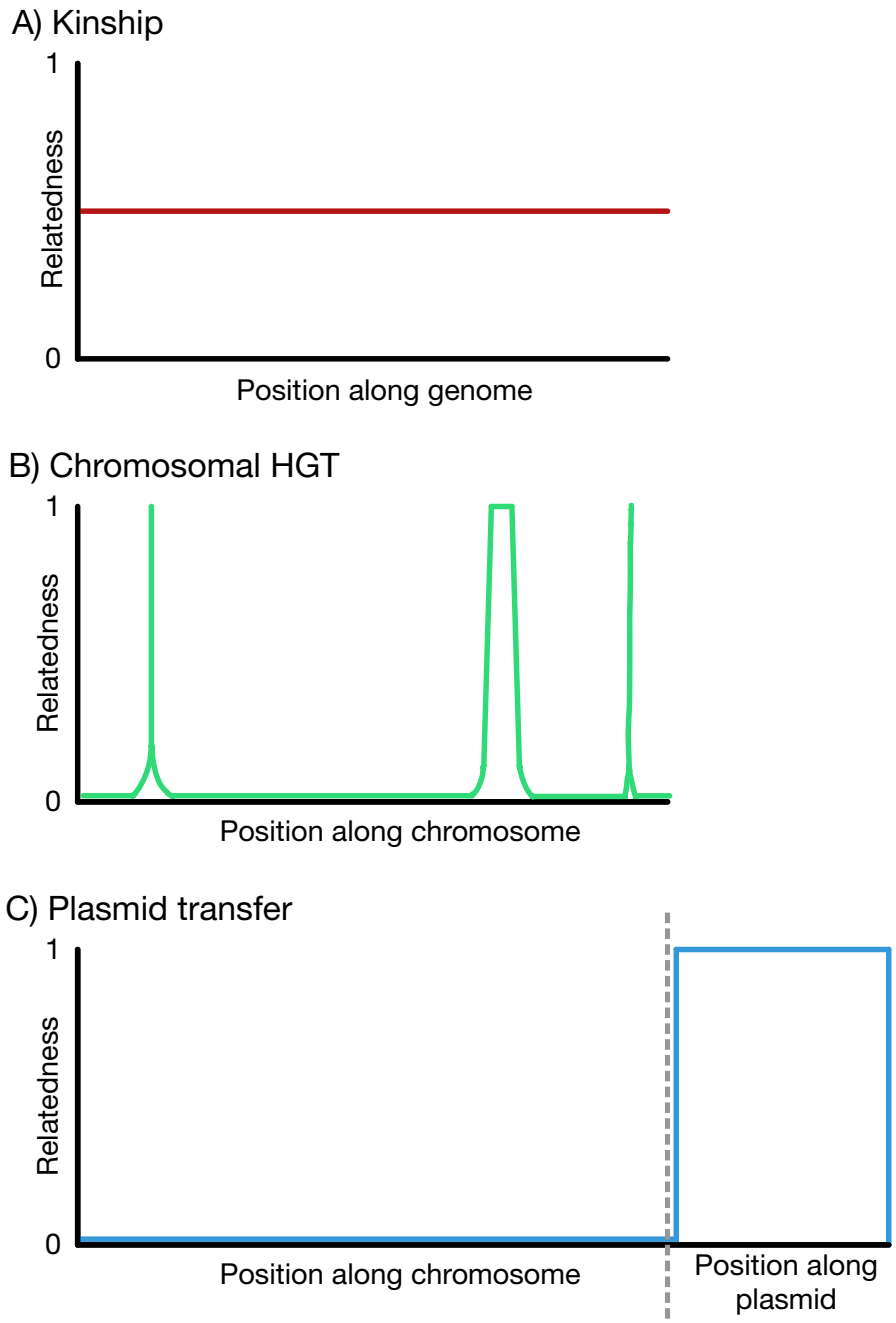


Figure 2. Relatedness through kinship, chromosomal HGT and plasmid transfer

(a) The relatedness between outbred diploid siblings is the same across all loci in the genome due to common ancestry. (b) Horizontal gene transfer (HGT) can occur between unrelated bacteria, meaning individuals could be $r = 1$ at the transferred loci and $r = 0$ everywhere else. (c) Plasmids can transfer between unrelated bacteria, a form of HGT called conjugation, meaning that individuals could be $r = 1$ at all plasmid loci and $r = 0$ at all chromosome loci. Adapted from Grafen, 1985.

Horizontal Gene Transfer

However, more or less equal relatedness across the genome may not be the case for all organisms (Figure 2b & 2c). Horizonal gene transfer (HGT), where genetic material is transferred between organisms other than vertically, has been found to be remarkably common (Land *et al.* 2015; Soucy *et al.* 2015). This is particularly true for bacteria, with much of the average bacterial genome made up of previously 'foreign' DNA (McInerney *et al.* 2017). Horizontal gene transfer can mean that the evolutionary history of a gene may be very different to the organism in which it is present (Figure 3).

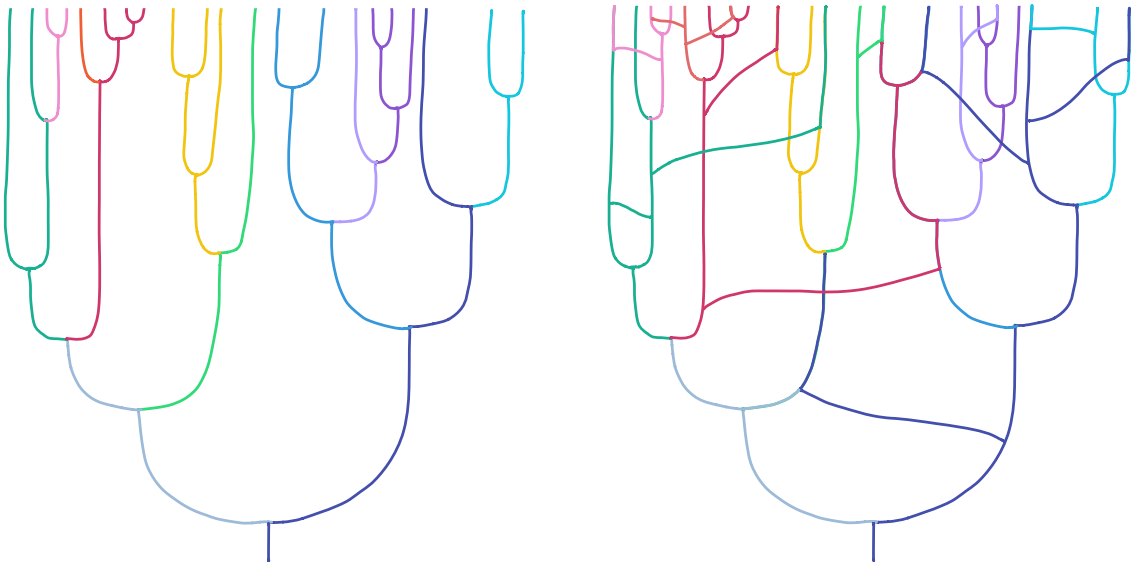


Figure 3. Horizontal gene transfer and evolution

Two illustrations of a phylogenetic gene tree, with different alleles of a gene in different colours. The left-hand tree is without horizontal gene transfer: any gene will have the same evolutionary history as the individual that carries it. In contrast, the right-hand tree is with horizontal gene transfer: a gene may have a different evolutionary history to both the individual that carries it and the other genes in the genome.

There are three major types of HGT in bacteria which vary in terms of how much control the donor and/or recipient have on the transfer, how random or non-random it is, and how frequently the genes transferred actually incorporate into the recipient genome (Thomas & Nielsen 2005; Hall *et al.* 2017) (Figure 4). First is transformation, where a bacterial cell enters a state called competence and takes up free DNA from the environment, some of which may end up incorporated into the genome (Lorenz & Wackernagel 1994; Chen & Dubnau 2004).

Second is transduction, which is mediated by infection by bacteriophages (viruses which infect bacteria) (Canchaya *et al.* 2003). Here, a phage inserts itself into the bacterial genome, and then can take sections with it when it becomes encapsulated to exit the cell (Jiang & Paul 1998). Third, conjugation is where semi-autonomous segments of DNA transfer to nearby cells via a tube called a pili, usually encoded by the segment itself (Llosa *et al.* 2002). The conjugation process has been particularly well studied in plasmids, which are small, usually circular pieces of DNA that can replicate independently from the rest of the genome (Stewart & Levin 1977; Willetts & Skurray 1980; Pinilla-Redondo *et al.* 2018).

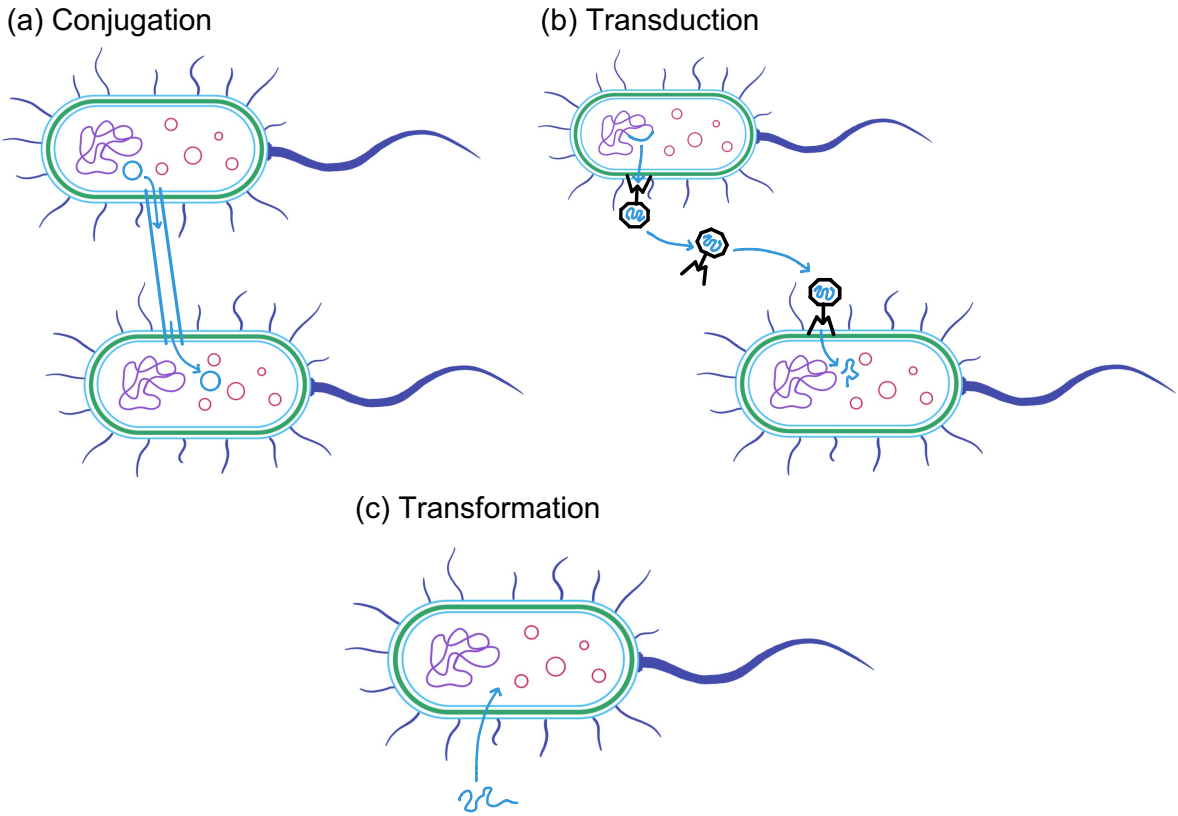


Figure 4. Types of Horizontal Gene Transfer (HGT) in bacteria

Blue indicates genetic material which is being transferred from one cell to the other. The arrows indicate the direction of this transfer. (a) A conjugative plasmid (shown as a blue circle) moves through a pili into the cytoplasm of the recipient cell. (b) A bacteriophage genome inserts itself into the bacterial chromosome (or plasmid), where it replicates and is repackaged into the phage capsule (shown in black). Along with its genome, sections of DNA either side of the phage can be incorporated into the capsule. This capsule

then infects other bacterial cells, transferring the DNA it carries. (c) A bacterial cell enters a state called competence where it can take up free DNA from the environment into the cytoplasm.

How could horizontal gene transfer affect bacterial cooperation?

There is a possibility that horizontal gene transfer could promote and stabilise cooperation in bacteria. If cooperative genes were able to be transferred horizontally, then any cheats that lose the gene could be 're-infected' with the cooperative gene (Smith 2001) (Figure 5). Transfer of the gene between individuals could increase relatedness at the cooperative locus and favour cooperation (Smith 2001; Nogueira *et al.* 2009; Mc Ginty *et al.* 2011, 2013; Dimitriu *et al.* 2014). Furthermore, transfer of the cooperative gene could favour cooperation even in scenarios where individuals were unable to direct the benefits of the behaviour towards other cooperators. Of the three methods of horizontal gene transfer in bacteria, conjugation via plasmids has received by far the most research attention as a potential driver of this cooperation hypothesis (Figure 5) (Smith 2001; Mc Ginty *et al.* 2013; Dimitriu *et al.* 2014).

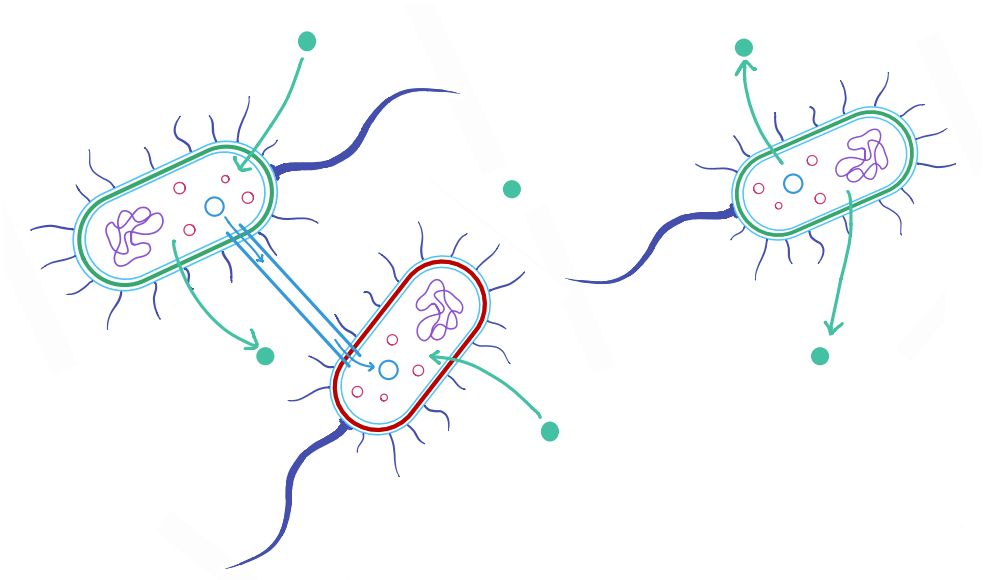


Figure 5. HGT increases relatedness by 're-infecting' those without the cooperative gene

Green bacteria are co-operators, which produce and take up public goods (green dots). They carry a plasmid which encodes the production of this public good (shown in blue). The red bacterium is a cheat, which does not produce yet still takes up public goods. It can cheat because it does not carry the cooperative plasmid. In the figure, a cooperator is able to transfer the

cooperative plasmid to the cheat via a conjugative pili, increasing relatedness and restoring the cheat to a co-operator.

Compared to other mobile genetic elements, plasmids could be a particularly good candidate for driving this hypothesis. First, plasmids have been sequenced in virtually all bacterial phyla (Rodríguez-Beltrán *et al.* 2021). As such, a bacterial genome is usually made up of one large chromosome, and several smaller plasmids. Since cooperation is thought to occur in almost all bacteria, the high prevalence of plasmids means they could help to stabilise cooperation in each of these species. Second, many of these plasmids can move to other bacterial cells via conjugation, and this can occur extremely rapidly within populations (Smillie *et al.* 2010; Sheppard *et al.* 2020). Third, in addition to replication and mobilization machinery, many plasmids carry genes that code for beneficial and ecologically relevant traits (Rankin *et al.* 2011). These include genes for antibiotic resistance, heavy-metal resistance, and toxins for virulence (Hale 1991; Cornelis *et al.* 1998; Gullberg *et al.* 2014; Lopatkin *et al.* 2017; Stevenson *et al.* 2017). It has not gone unnoticed that many of these genes may also act as cooperative public goods (Rankin *et al.* 2011; Nogueira *et al.* 2012; Garcia-Garcera & Rocha 2020).

This cooperation hypothesis has received some support. Theoretical and experimental studies have shown that transfer of cooperative genes on plasmids could select for cooperation in situations where it would not otherwise be favoured (Smith 2001; Mc Ginty *et al.* 2011, 2013; Dimitriu *et al.* 2014). Additionally, there is some evidence from studies on bacterial genomes that plasmids may carry proportionally more genes coding for extracellular proteins, which are likely to act as cooperative public goods, than the less mobile chromosome (Nogueira *et al.* 2009, 2012; Garcia-Garcera & Rocha 2020). This would be expected if plasmid transfer consistently helped to maintain cooperation in bacteria.

However, there are also potential issues with this hypothesis. One is that the rate of transfer of the cooperative gene must be greater than the fitness benefit gained by 'cheats', and fast enough to have a real influence on relatedness (Ghoul *et al.* 2017). Given that there is considerable variation in how quickly plasmids can transfer, it is possible that many plasmids do not transfer fast enough between cells for this to be the case (Sheppard *et al.* 2020). Second, plasmid incompatibility, where certain plasmids are unable to coexist stably within a cell, could lead to 'cheat plasmids' emerging (Pinto *et al.* 2012; Hülter *et al.* 2017). One factor that can cause

plasmids to be incompatible is if they share similar replication or partitioning systems (Hülter *et al.* 2017). If a cell emerged with a version of the plasmid that had lost the cooperative gene, the similarity of this 'cheat plasmid' to the original 'cooperative plasmid' could prevent the cell from being 're-infected' with the cooperative gene. Over time, the 'cheat plasmid' would be expected to increase in frequency in the population, preventing stable cooperation.

Additionally, with such frequent horizontal gene transfer in bacteria, the linkage between the inheritance of a cooperative gene and the rest of the genome may be much weaker in bacteria than in multicellular eukaryotes (Figure 2). This could open up the possibility of intragenomic conflict and prevent cooperative genes from spreading (Scott & West 2019; Hall *et al.* 2020). Therefore, horizontal gene transfer of cooperative genes could actually cause a problem for the evolution of cooperation in bacteria.

Furthermore, this is not the only hypothesis that might predict that plasmids should carry proportionally more genes for public goods (Nogueira *et al.* 2009; Ghoul *et al.* 2017). First, while public goods are likely to have cooperative effects, their location in the extracellular space means they are also likely to interact with the environment (Garcia-Garcera & Rocha 2020). Genes that help bacteria adapt to certain environments could also be expected to be favoured if located on mobile elements, since horizontal gene transfer would allow the gene to be easily gained when the trait is required and easily lost when no longer needed. Second, carriage of genes on plasmids may provide benefits beyond their ability to transfer (Rodríguez-Beltrán *et al.* 2021). There are many plasmids actually incapable of transferring via conjugation (Smillie *et al.* 2010). Despite this, these non-mobilizable plasmids are still found across many bacterial species, and often code for traits which are useful to their hosts (Smillie *et al.* 2010). This suggests that rather than simple vehicles of horizontal gene transfer, there may be other selection pressures that could cause plasmids to carry certain kinds of genes (Rodríguez-Beltrán *et al.* 2021).

Thesis Outline

Here, I explore the role of horizontal gene transfer in bacterial cooperation and evolution. For most of the thesis, I focus in particular on plasmids. I first examine a potential role of plasmids in stabilising cooperation via conjugation, before analysing other features of plasmids, including their size, mobility, range, and base content. Finally, I consider how horizontal gene

transfer might impact the evolution and structure of bacterial genomes more generally. Specifically:

In **Chapter 2**, I use comparative genomics across 51 bacterial species to test two key predictions of the hypothesis that conjugation via plasmids could stabilise cooperation. Contrary to these predictions, I show that genes coding for extracellular proteins are not more likely to be found on plasmids compared to chromosomes, and on more mobile plasmids compared to less mobile plasmids. I then discuss the reasons why previous studies found support for this hypothesis. Instead, I find evidence that the lifestyle of a species may determine whether their plasmids carry more genes coding for extracellular proteins. Specifically, I show that plasmids of pathogenic species with a broad host-range are particularly enriched with genes coding for extracellular proteins, compared to non-pathogens and narrow host-range pathogens. I find that this is because plasmids in these species code for many more extracellular proteins involved in pathogenicity. This suggests that these species carry pathogenicity genes on their plasmids because of benefits other than being able to re-infect cheats with these genes.

In **Chapter 3**, I further explore the potential role of plasmids in bacterial evolution by considering how potential 'life-history' traits of plasmids, specifically their size, mobility, and range, correlate with one another. I find that, consistent with previous studies, conjugative plasmids are generally largest, while mobilizable plasmids are smallest. I also find that plasmid mobility and range are positively correlated. Additionally, I find the correlation between plasmid size and range is different depending on the mobility of plasmids. Together, these analyses provide a comprehensive study of the variation in key characteristics of bacterial plasmids, and are a basis for future work.

In **Chapter 4**, I test the predictions of two hypotheses for why plasmid sequences are often observed to be enriched with A and T bases, compared to chromosomes. These two hypotheses suggest that AT-bias is: (1) an adaptation to reduce plasmid cost; (2) an artefact of increased mutation and genetic drift in plasmids. To test which of these hypotheses is more likely, I explore how plasmid AT-content varies with respect to plasmid mobility and plasmid range. Overall, I find more evidence for the hypothesis that AT-bias of plasmids is due to mutation and genetic drift. I then discuss how future studies could further explore evidence for these two hypotheses.

In **Chapter 5**, I use comparative genomics to provide an initial test of the general observation that bacterial species' genomes become more variable with increasing environmental variability. I examine the structure of 126 species' pangenomes, defined as the total number of genes sequenced in a species. I specifically consider how the percentage of core genes, found in all genomes of a species, and accessory genes, found in only a subset of genomes, varies across bacteria. I then compare these measures to two proxies of environmental variability, before discussing limitations and future directions to further explore this question.

In **Chapter 6**, I summarise and discuss the main results presented in Chapters 2-5. I consider what more we now know about a potential role of horizontal gene transfer, and particularly plasmids, in bacterial cooperation. Additionally, I discuss how we can define cooperative behaviours, and how we can identify cooperative genes in bacteria. I also consider the advantages and limitations of using comparative genomics to study bacterial cooperation.

Finally, the **appendix** contains two additional manuscripts which I contributed to during my DPhil. The first tests for signatures of kin selection in the social genes of the bacterial species *Pseudomonas aeruginosa* (Belcher et al. Submitted, PNAS). The second tests the hypothesis that genes carried on plasmids have a lower complexity than chromosome genes, and whether this holds for genes coding for extracellular proteins, which could act as cooperative public goods (Hao et al. Draft, Unsubmitted).

References

Andersen, S.B., Marvig, R.L., Molin, S., Johansen, H.K. & Griffin, A.S. (2015). Long-term social dynamics drive loss of function in pathogenic bacteria. *PNAS*, 112, 10756–10761.

Burrows, S.M., Elbert, W., Lawrence, M.G. & Pöschl, U. (2009). Bacteria in the global atmosphere – Part 1: Review and synthesis of literature data for different ecosystems. *Atmospheric Chemistry and Physics*, 9, 9263–9280.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.L. & Brüssow, H. (2003). Phage as agents of lateral gene transfer. *Curr Opin Microbiol*, 6, 417–424.

Chen, I. & Dubnau, D. (2004). DNA uptake during bacterial transformation. *Nat Rev Microbiol*, 2, 241–249.

Cordero, O.X., Ventouras, L.-A., DeLong, E.F. & Polz, M.F. (2012). Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. *PNAS*, 109, 20059–20064.

Cornelis, G.R., Boland, A., Boyd, A.P., Geuijen, C., Iriarte, M., Neyt, C., *et al.* (1998). The Virulence Plasmid of Yersinia, an Antihost Genome. *Microbiol Mol Biol Rev*, 62, 1315–1352.

Davies, N.B., Krebs, J.R. & West, S.A. (2013). *An Introduction to Behavioural Ecology*. Wiley-Blackwell.

Diggle, S.P., Griffin, A.S., Campbell, G.S. & West, S.A. (2007). Cooperation and conflict in quorum-sensing bacterial populations. *Nature*, 450, 411–414.

Dimitriu, T., Lotton, C., Benard-Capelle, J., Misevic, D., Brown, S.P., Lindner, A.B., *et al.* (2014). Genetic information transfer promotes cooperation in bacteria. *Proceedings of the National Academy of Sciences*, 111, 11103–11108.

Foster, K.R. (2010). Social behaviour in microorganisms. In: *Social Behaviour* (eds. Szekely, T., Moore, A.J. & Komdeur, J.). Cambridge University Press, Cambridge, pp. 331–356.

Garcia-Garcera, M. & Rocha, E.P.C. (2020). Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. *Nature Communications*, 11, 758.

Gest, H. (2004). The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society. *Notes Rec R Soc Lond*, 58, 187–201.

Ghoul, M., Andersen, S.B. & West, S.A. (2017). Sociomics: Using Omic Approaches to Understand Social Evolution. *Trends in Genetics*, 33, 408–419.

Ghoul, M., Griffin, A.S. & West, S.A. (2014a). Toward an evolutionary definition of cheating. *Evolution*, 68, 318–331.

Ghoul, M., West, S.A., Diggle, S.P. & Griffin, A.S. (2014b). An experimental test of whether cheating is context dependent. *J. Evol. Biol.*, 27, 551–556.

Grafen, A. (1985). A geometric view of relatedness. *Oxford surveys in evolutionary biology*, 2, 28–89.

Grafen, A. (1989). The phylogenetic regression. *Philos Trans R Soc Lond B Biol Sci*, 326, 119–157.

Griffin, A.S., West, S.A. & Buckling, A. (2004). Cooperation and competition in pathogenic bacteria. *Nature*, 430, 1024–1027.

Gullberg, E., Albrecht, L.M., Karlsson, C., Sandegren, L. & Andersson, D.I. (2014). Selection of a Multidrug Resistance Plasmid by Sublethal Levels of Antibiotics and Heavy Metals. *mBio*, 5, 2014 v.5 no.5.

Hale, T.L. (1991). Genetic basis of virulence in Shigella species. *Microbiol. Rev.*, 55, 206–224.

Hall, J.P.J., Brockhurst, M.A. & Harrison, E. (2017). Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372, 20160424.

Hall, R.J., Whelan, F.J., McInerney, J.O., Ou, Y. & Domingo-Sananes, M.R. (2020). Horizontal Gene Transfer as a Source of Conflict and Cooperation in Prokaryotes. *Front Microbiol*, 11, 1569.

Hamilton, W.D. (1963). The Evolution of Altruistic Behavior. *The American Naturalist*, 97, 354–356.

Hamilton, W.D. (1964). Genetical evolution of social behaviour I & II. *J. Theor. Biol.*, 7, 1–52.

Hayat, R., Ali, S., Amara, U., Khalid, R. & Ahmed, I. (2010). Soil beneficial bacteria and their role in plant growth promotion: a review. *Ann Microbiol*, 60, 579–598.

Hülter, N., Ilhan, J., Wein, T., Kadibalban, A.S., Hammerschmidt, K. & Dagan, T. (2017). An evolutionary perspective on plasmid lifestyle modes. *Current Opinion in Microbiology*, 38, 74–80.

Jiang, S.C. & Paul, J.H. (1998). Gene Transfer by Transduction in the Marine Environment. *Applied and Environmental Microbiology*, 64, 2780–2787.

Kümmerli, R., Griffin, A.S., West, S.A., Buckling, A. & Harrison, F. (2009). Viscous medium promotes cooperation in the pathogenic bacterium Pseudomonas aeruginosa. *Proceedings of the Royal Society B: Biological Sciences*, 276, 3531–3538.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M.R., Ahn, T.-H., *et al.* (2015). Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics*, 15, 141–161.

Llosa, M., Gomis-Rüth, F.X., Coll, M. & Cruz, F. de la. (2002). Bacterial conjugation: a two-step mechanism for DNA transport. *Molecular Microbiology*, 45, 1–8.

Lopatkin, A.J., Meredith, H.R., Srimani, J.K., Pfeiffer, C., Durrett, R. & You, L. (2017). Persistence and reversal of plasmid-mediated antibiotic resistance. *Nat Commun*, 8, 1689.

Lorenz, M.G. & Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. *Microbiological Reviews*, 58, 563–602.

Mc Ginty, S.É., Lehmann, L., Brown, S.P. & Rankin, D.J. (2013). The interplay between relatedness and horizontal gene transfer drives the evolution of plasmid-carried public goods. *Proc. R. Soc. B*, 280, 20130400.

Mc Ginty, S.E., Rankin, D.J. & Brown, S.P. (2011). Horizontal gene transfer and the evolution of bacterial cooperation: mobile elements and bacterial cooperation. *Evolution*, 65, 21–32.

McInerney, J.O., McNally, A. & O'Connell, M.J. (2017). Why prokaryotes have pangenomes. *Nat Microbiol*, 2, 17040.

McNally, L., Viana, M. & Brown, S.P. (2014). Cooperative secretions facilitate host range expansion in bacteria. *Nat Commun*, 5, 4594.

Mund, A., Diggle, S.P. & Harrison, F. (2017). The Fitness of Pseudomonas aeruginosa Quorum Sensing Signal Cheats Is Influenced by the Diffusivity of the Environment. *mBio*, 8, e00353-17.

Nogueira, T., Rankin, D.J., Touchon, M., Taddei, F., Brown, S.P. & Rocha, E.P.C. (2009). Horizontal Gene Transfer of the Secretome Drives the Evolution of Bacterial Cooperation and Virulence. *Current Biology*, 19, 1683–1691.

Nogueira, T., Touchon, M. & Rocha, E.P.C. (2012). Rapid Evolution of the Sequences and Gene Repertoires of Secreted Proteins in Bacteria. *PLoS One*, 7, e49403.

Orsi, W.D., Richards, T.A. & Francis, W.R. (2018). Predicted microbial secretomes and their target substrates in marine sediment. *Nature Microbiology*, 3, 32–37.

Pinilla-Redondo, R., Cyriaque, V., Jacquiod, S., Sørensen, S.J. & Riber, L. (2018). Monitoring plasmid-mediated horizontal gene transfer in microbiomes: recent advances and future perspectives. *Plasmid*, 99, 56–67.

Pinto, U.M., Pappas, K.M. & Winans, S.C. (2012). The ABCs of plasmid replication and segregation. *Nat Rev Microbiol*, 10, 755–765.

Rankin, D.J., Rocha, E.P.C. & Brown, S.P. (2011). What traits are carried on mobile genetic elements, and why? *Heredity*, 106, 1–10.

Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R.C. & San Millán, Á. (2021). Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology*, 1–13.

Rumbaugh, K.P., Diggle, S.P., Watters, C.M., Ross-Gillespie, A., Griffin, A.S. & West, S.A. (2009). Quorum Sensing and the Social Evolution of Bacterial Virulence. *Current Biology*, 19, 341–345.

Scott, T.W. & West, S.A. (2019). Adaptation is maintained by the parliament of genes. *Nat Commun*, 10, 5163.

Sheppard, R.J., Beddis, A.E. & Barraclough, T.G. (2020). The role of hosts, plasmids and environment in determining plasmid transfer rates: A meta-analysis. *Plasmid*, 108, 102489.

Smillie, C., Garcillan-Barcia, M.P., Francia, M.V., Rocha, E.P.C. & de la Cruz, F. (2010). Mobility of Plasmids. *Microbiology and Molecular Biology Reviews*, 74, 434–452.

Smith, J. (2001). The social evolution of bacterial pathogenesis. *Proc. R. Soc. Lond. B*, 268, 61–69.

Soucy, S.M., Huang, J. & Gogarten, J.P. (2015). Horizontal gene transfer: building the web of life. *Nat Rev Genet*, 16, 472–482.

Stevenson, C., Hall, J.P., Harrison, E., Wood, Aj. & Brockhurst, M.A. (2017). Gene mobility promotes the spread of resistance in bacterial populations. *ISME J*, 11, 1930–1932.

Stewart, F.M. & Levin, B.R. (1977). The Population Biology of Bacterial Plasmids: A Priori Conditions for the Existence of Conjugationally Transmitted Factors. *Genetics*, 87, 209–228.

Thomas, C.M. & Nielsen, K.M. (2005). Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nat Rev Microbiol*, 3, 711–721.

West, S.A., Diggle, S.P., Buckling, A., Gardner, A. & Griffin, A.S. (2007a). The Social Lives of Microbes. *Annu. Rev. Ecol. Evol. Syst.*, 38, 53–77.

West, S.A., Griffin, A.S. & Gardner, A. (2007b). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *J Evolution Biol*, 20, 415–432.

West, S.A., Griffin, A.S., Gardner, A. & Diggle, S.P. (2006). Social evolution theory for microorganisms. *Nat Rev Microbiol*, 4, 597–607.

Willetts, N. & Skurray, R. (1980). The Conjugation System of F-Like Plasmids. *Annual Review of Genetics*, 14, 41–76.

Zhang, Y.-J., Li, S., Gan, R.-Y., Zhou, T., Xu, D.-P. & Li, H.-B. (2015). Impacts of Gut Bacteria on Human Health and Diseases. *International Journal of Molecular Sciences*, 16, 7493–7519.

Chapter 2. Plasmids do not consistently stabilize cooperation across bacteria, but may promote broad pathogen host-range

The work contained in the following Chapter is currently in press for publication in the journal Nature Ecology & Evolution. Therefore, I have included the final manuscript of the accepted paper, as is permitted in an integrated thesis.

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

Anna E. Dewar[1,a,*], Joshua L. Thomas[1,a], Thomas W. Scott[1], Geoff Wild[2], Ashleigh S. Griffin[1], Stuart A. West[1,b], Melanie Ghoul[1,b]

[1]Department of Zoology, University of Oxford, Oxford, OX1 3SZ, United Kingdom

[2]Department of Applied Mathematics, University of Western Ontario, London, Ontario N6A 3K7, Canada

a Joint first author

b Joint last author

*Corresponding author. Email: anna.dewar@zoo.ox.ac.uk

Abstract

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. In 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: (i) plasmids compared to chromosomes; or (ii) 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.

Introduction

The growth and success of many bacterial populations depends upon 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 which 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 destabilising 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] (Figure 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 though 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 (Figure 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 (Figure 1b)[32–34]. Genes which facilitate adaptation to environmental variability are often those which 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 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 (Figure 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 which 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: (i) plasmids relative to chromosomes; (ii) 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 1632 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 which 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 (Figure 1b), or instead because of some additional benefit conferred by plasmid carriage (Figure 1c).

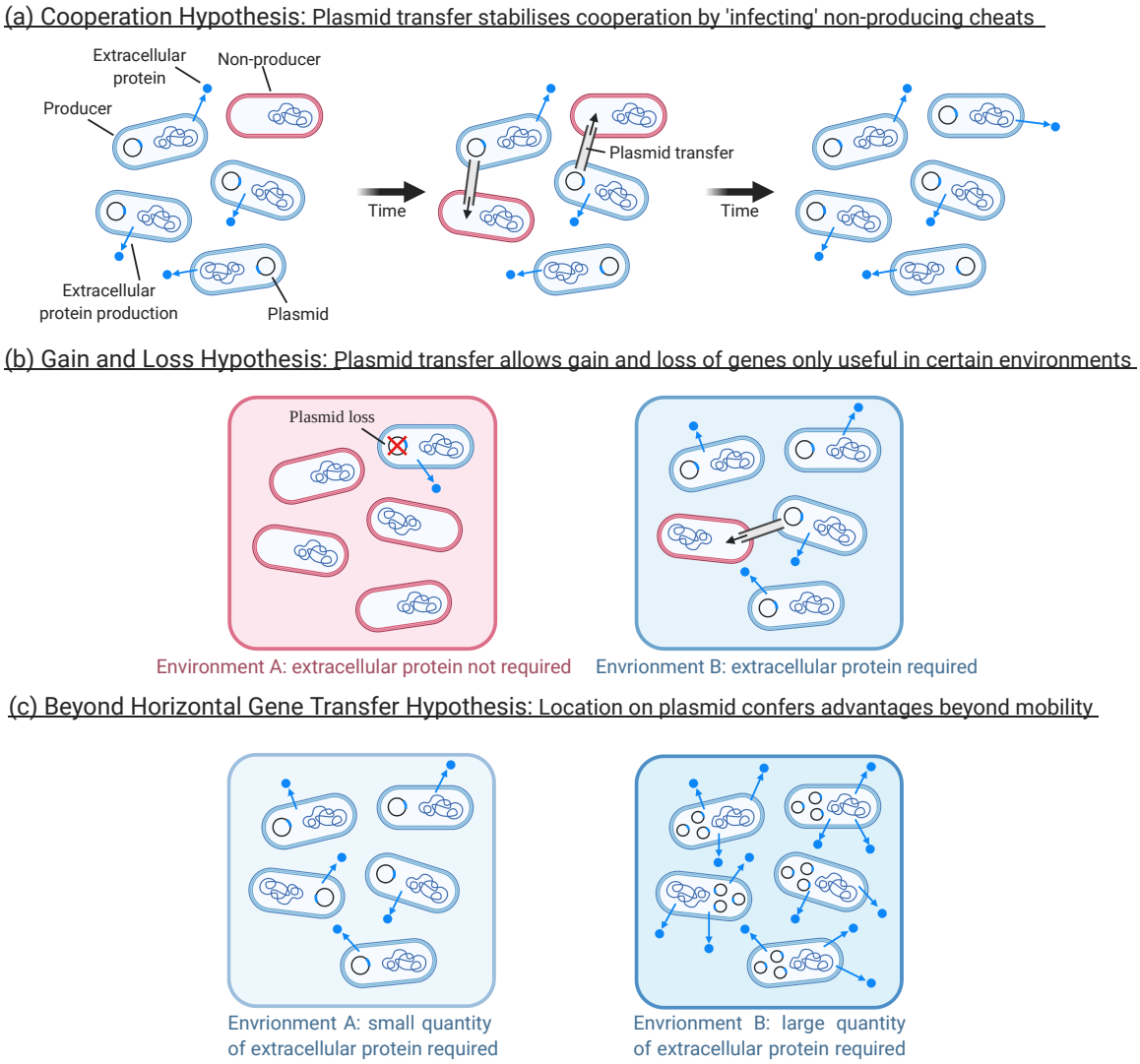


Figure 1. Three hypotheses for why selection might favour genes coding for extracellular proteins to be located on plasmids. (a) Cooperation Hypothesis. Blue cells produce extracellular proteins which act as cooperative public goods, while red cells are 'cheats' which exploit this cooperation. Over time 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 stabilising cooperation[14–18]. (b) Gain and Loss Hypothesis. The production of the extracellular protein is required in some environments, but not 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.

Results

Genomic Analyses.

We use the approach developed by Nogueira *et al.*[15,19,20], of using PSORTb[39] to predict the subcellular location of every protein encoded by 1632 complete genomes from 51 diverse bacterial species (Extended Data Figure 1; Table S3). We are also building upon 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 the average bacterial genome had 2696 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 5 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 localisations 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) (Figure S2). These patterns are very similar to those found previously[3,15,19,20].

Extracellular proteins are not overrepresented on plasmids.

We found that extracellular proteins were not more likely to be carried on plasmids compared to chromosomes (Figure 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 (MCMCglmm[41]; posterior mean = 0.004, 95% CI = -0.063 to 0.057, pMCMC= 0.87; n = 1632 genomes; R^2 of species sample size = 0.47, R^2 of phylogeny = 0.17; Table S2, row 1a). This result was robust to alternative forms of analysis. We also found no significant difference when we: (i) compared chromosomes to plasmids of only certain mobilities (Fig S3; Table S2, rows 20-22); (ii) 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 Figure 2; Table S2, rows 2 and 3). Our analyses use a bacterial phylogeny, which assumes plasmid evolution follows bacterial phylogeny, but we also found no significant pattern if we ignored phylogeny and analysed species as independent data points (Figure 2; Table S2, 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 (Figure 2; Extended Data Figure 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 Figure 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 Figure 2). Clearly, across species, genes coding for extracellular proteins are not consistently more likely to be on plasmids.

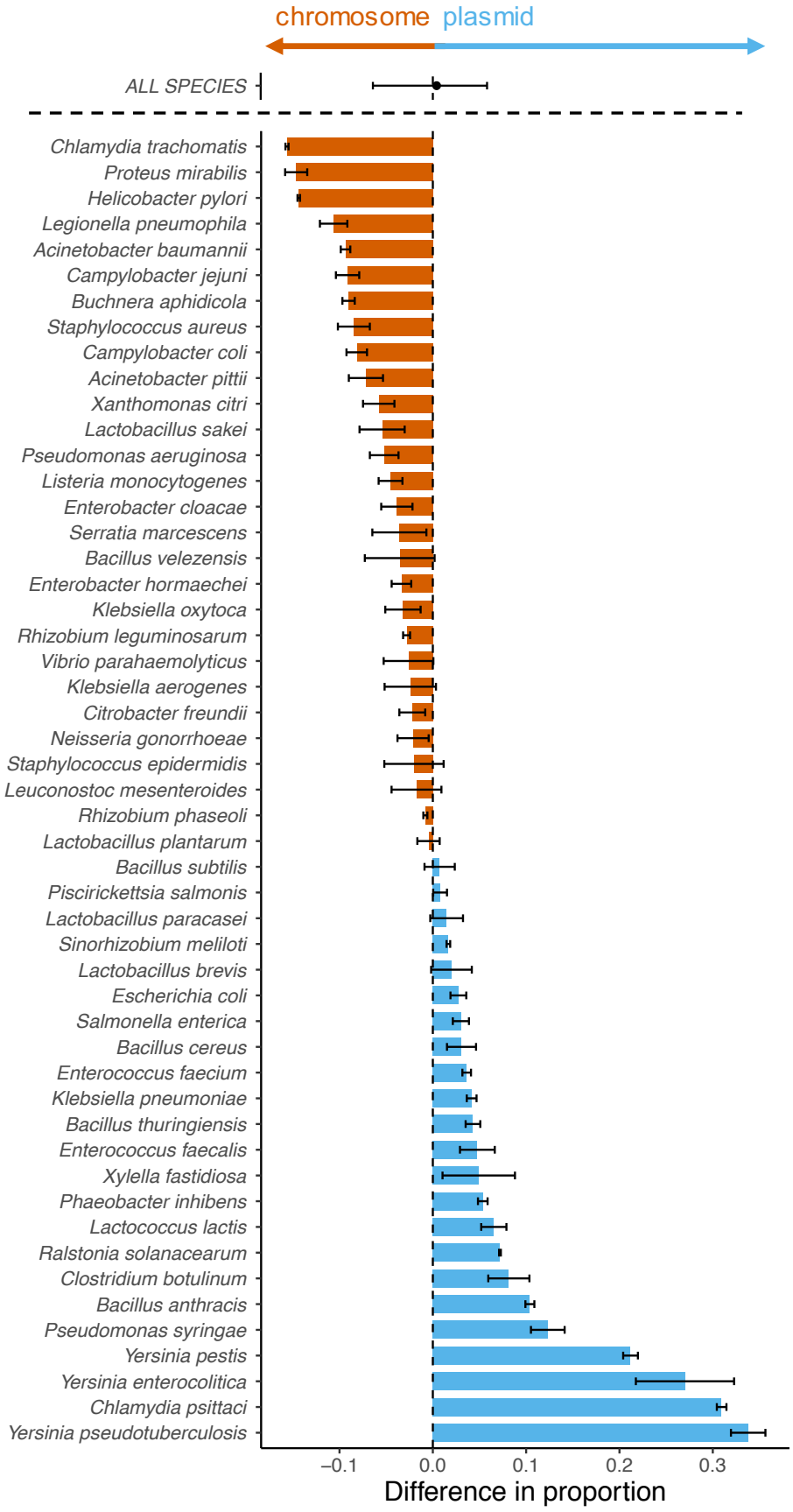


Fig 2. Extracellular proteins are not overrepresented on plasmids. For each species we calculated the mean difference between plasmid(s) and

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 chromosome genes. Species in red have a difference less than zero, meaning their chromosome genes code for a greater proportion of extracellular proteins than 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% Credible Interval 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 (i.e. no consistent trend towards species in blue).

As a control, we also analysed the genomic location of the genes coding for all other classes of protein (Extended Data Figure 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 localisations were significantly overrepresented on plasmids or chromosomes across the 51 species (Extended Data Figure 3; Table S2, 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 chromosomes[15,19,20]. The first of these studies found this pattern across 20 *Escherichia coli* genomes[15]. We also found that genes coding for extracellular proteins in *E. coli* were more likely to be found on plasmids (Figure 2; Extended Data Figure 2). However, Figure 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 their plasmid(s).

Two subsequent, multi-species studies found that plasmid genes were significantly more likely to code for extracellular proteins than 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= 826530, p-value <0.001, R^2 = 0.385; n = 1632 plasmid-chromosome pairs). When analysing other questions, Garcia-Garcera & 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: (i) independent from one another; (ii) a representative sample of bacteria in nature. Neither of these are true for multi-species 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 = 1632 plasmid-chromosome paired differences in extracellular proportion; R^2: species sample size = 0.46, phylogeny = 0.34; Table S2, 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 1632 genomes). Taken together, this illustrates that it is not our data which disagrees with previous studies, but instead our use of statistical analyses appropriate for multi-genome, multi-species 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 upon 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 (Supp. Info. 1)[43–45]. When bacterial 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 (e.g. Figure 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 upon conjugative plasmids' machinery to transfer, to have intermediate mobility; 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 (Figure 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 to 0.052, pMCMC = 0.73; n = 40; Table S2, 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 (Figure S4; Table S2, row 12). We also found no correlation between the proportion of a species' plasmids which can transfer and how overrepresented or underrepresented extracellular proteins are on plasmids compared to chromosomes (Extended Data Figure 4; Table S2, 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 (Supp. Info. 1; Figure S6). 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.

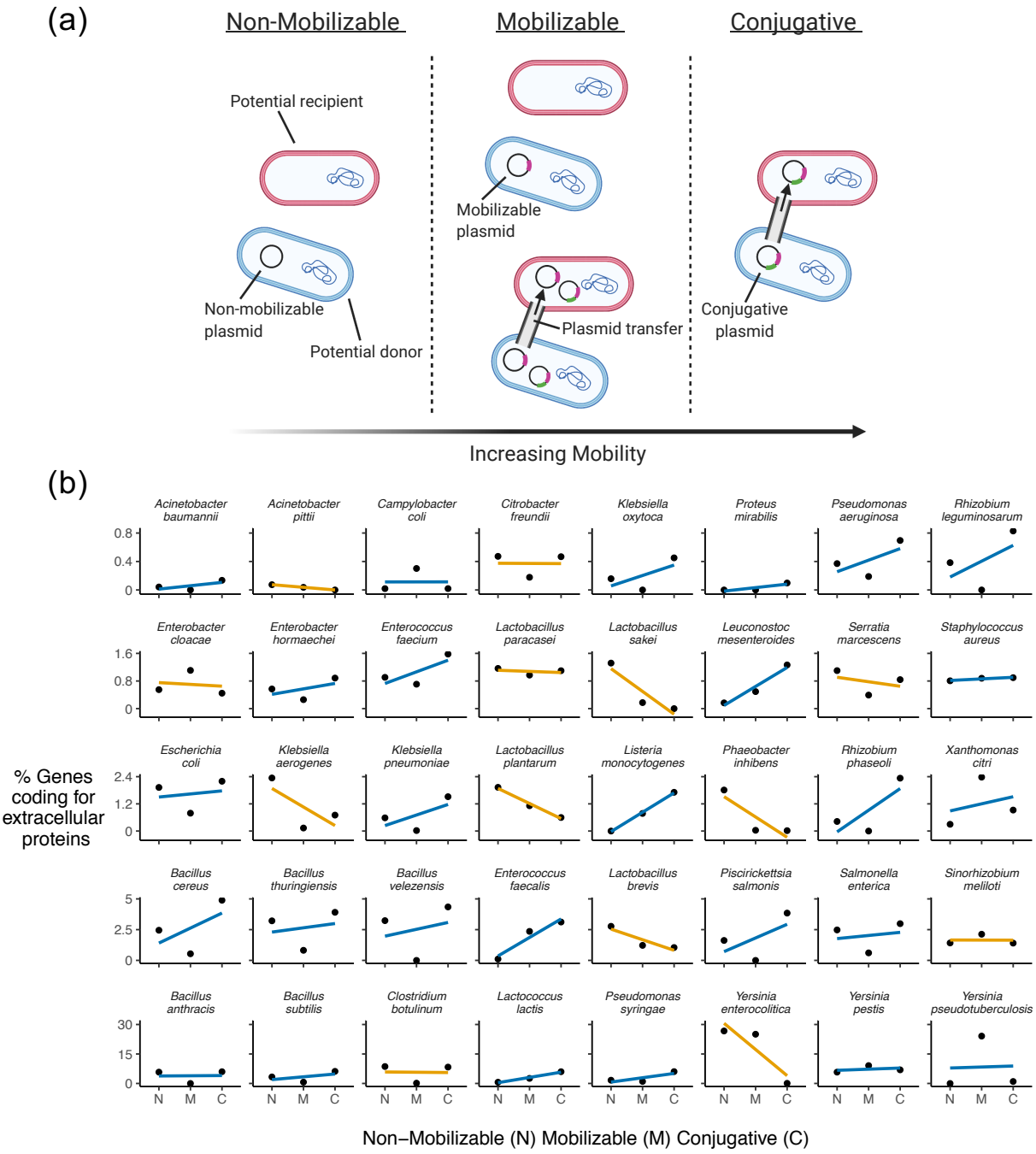


Figure 3. Plasmid mobility and extracellular proteins. (a) We divided plasmids into three mobility types: non-mobilizable (lowest or no mobility); mobilizable (intermediate mobility); conjugative (highest mobility). Blue cells are potential plasmid donors, while red cells are

potential recipients. Each panel 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 which carried plasmids of all three mobilities are shown, with a panel for each of these species. Dots in each panel indicate the mean % 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.

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 a chromosome.

We assumed an infinite population of haploid individuals (bacterial cells). Individuals may carry a cooperative gene, that codes for public goods production, either 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 are 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.

Consistent with previous analyses, we found that, in the short term, horizontal gene transfer on a plasmid can initially help cooperation invade (Figure 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.

In contrast, we found that transfer on a plasmid did not appreciably increase the range of parameter space where cooperation was maintained at evolutionary equilibrium (Fig 4a & 5) (Supp. Info. 4). First, in the absence of plasmid loss (s=0), cooperation was only favoured when RB-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 (Figure 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 (Figure 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 (Figure 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 (Figures 4 & 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 (Figures 2 & 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* Typhimurium (*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 (Figure 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 (i.e. a high enough β)[54].

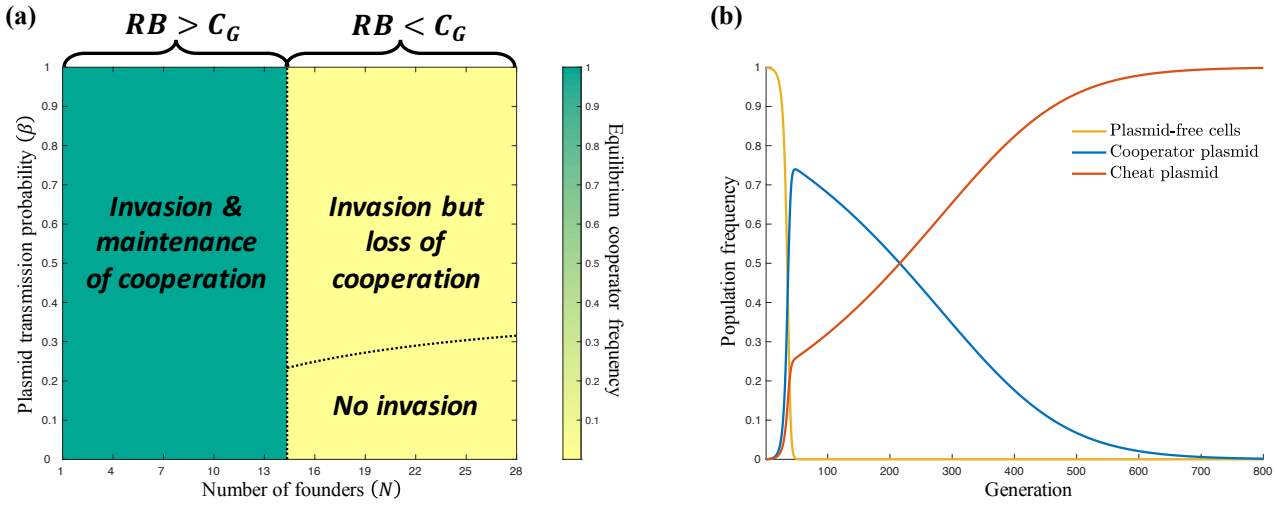


Figure 4. Plasmids facilitate the invasion but not the maintenance of cooperation. In parts (a) and (b), we plot the results of our theoretical model for the case when there is no plasmid loss ($s=0$). (a) Cooperation is only maintained at equilibrium (green shaded area) when it is favoured at the chromosomal level $RB > C_G$, which is unaffected by plasmid transfer (β). (b) 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). 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: (a & b) $B = 1.435, C_G = 0.1, C_C = 0.2$; (b) $\beta = 0.5, \mathrm{N} = 16$.

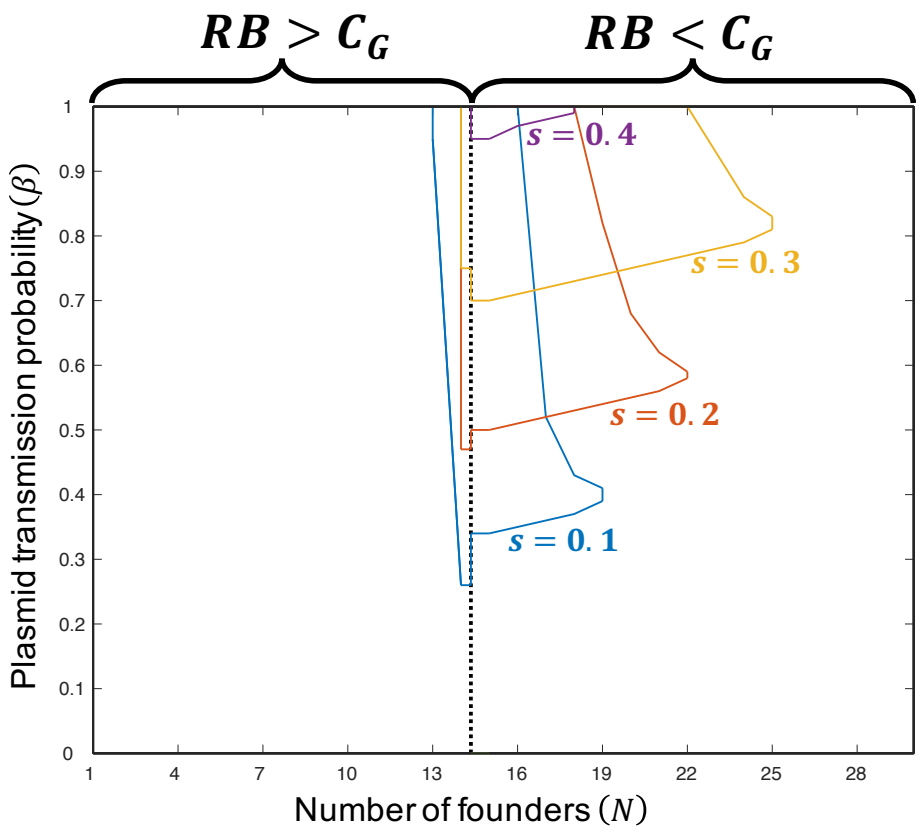


Figure 5. Plasmid loss can favour the maintenance of cooperation. We plot the results of our theoretical model for different levels of plasmid loss (s=0-1). The areas encapsulated by the coloured lines show the regions of parameter space where cooperation is polymorphic at equilibrium (i.e. population comprises some cooperators & 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 $RB > C_G$ (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 (β=0). When plasmid loss is high (s≥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).

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 which 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 (Figure 1b), or plasmids conferring some other advantage not associated with horizontal gene transfer, such as allowing copy number to be conditionally adjusted (Figure 1c)[31,32,38,55]. There are a number of different ways 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 data set, 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; Table S2, 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; Table S2, 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 Figure 5).

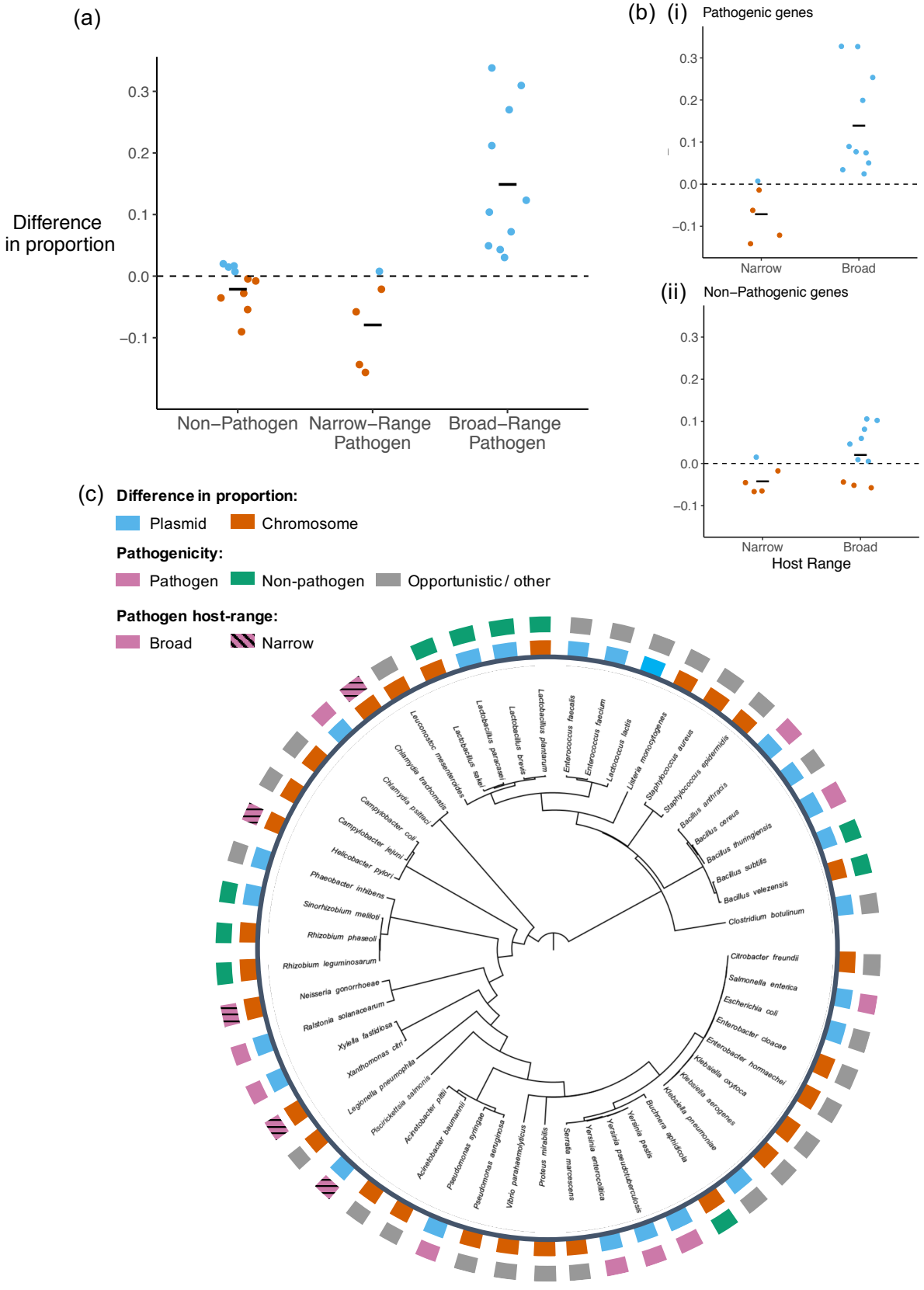


Figure 6. Pathogenicity, host-range and the location of genes coding for extracellular proteins. We have divided species into either pathogens or non-pathogens, with pathogens further categorised into those with a narrow or broad host-range. The y-axis in (a) shows the difference in the proportion of genes on plasmids and chromosomes coding for extracellular proteins – this is the same as the x-axis in Figure 2. The y-axes in (b)(i) and (b)(ii) 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. 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[64]. The inner ring indicates whether extracellular proteins were more likely to be coded for on the plasmid(s) or chromosome(s), as in Figure 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.

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 program MP3[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 (Figure 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 (Figure 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; Table S2, row 26). In contrast, the relative location of non-pathogenic extracellular proteins did not vary between broad and narrow host-range pathogens

(Figure 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; Table S2, row 27). Consequently, the excess of genes coding for extracellular proteins on the plasmids of broad host-range species (Figure 6a) appears to arise due to an excess of pathogenicity genes coding for extracellular proteins (Figure 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, 5 out of 10 broad host-range species and 3 out of 5 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 (Table S2, rows 28 to 30).

Pathogenic extracellular proteins could be preferentially coded for on plasmids to facilitate their gain and loss (Figure 1b: Gain and loss hypothesis), or because of some other benefit provided by being carried on a plasmid (Figure 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 (Figure S7) (MCMCglmm; posterior mean = -0.020, 95% CI = -0.224 to 0.185, pMCMC = 0.774; n=7; Table S2, 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 (Figure S8; Table S2, 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 (Figure 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 utilisation 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 on their gene content. For example, extracellular proteins are, on average, cheaper to produce than 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 (Figure 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 vs 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: (i) the number of five broad environments a species was sequenced in[20,65,66]; (ii) the proportion of a species' genomes that is composed of 'core' genes, which are those found in all genomes of the species – species which 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 Figure 6) (Supp. Info. 1; Table S2, rows 35 and 37). Garcia-Garcera & 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 (Supp. Info. 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 (Figure 6). This suggests that hypothesis c, which our data is most consistent with, may be important for pathogens in particular, but not necessarily across all bacterial species and lifestyles.

Complementary Analyses

There a number of directions in which our analyses could be expanded. 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[67,68]. 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[48,54,69–71]. 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: (i) we could compare our results with those from previous studies; (ii) secretion systems are highly conserved, allowing us to examine a large number of species, where detailed genetic annotations are lacking; (iii) 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[72,73]. However, while extracellular proteins are likely to be cooperative traits, not all cooperative genes code for extracellular proteins (e.g. secondary metabolites such as siderophores), and not all extracellular proteins are involved in cooperation (e.g. 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 not: (i) overrepresented on plasmids compared to chromosomes (Figure 2); (ii) more likely to be carried by plasmids that transfer at higher rates (Figure 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 (Figures 4 & 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. In 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 (Figure

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 1632 complete genomes comprising 51 bacterial species from GenBank RefSeq (https://www.ncbi.nlm.nih.gov) between February-November 2019. We used species on panX (http://pangenome.tuebingen.mpg.de)[74] 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 10 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 more than 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[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 programs of its kind[75]. It has also been used in previous analyses to identify 'cooperative' genes and/or extracellular proteins in bacteria[15,20]. The program has a number of modules which are trained to recognise 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 Figure 1). The locations for Gram-negative species are the same, except that cell wall is replaced with outer membrane and periplasmic, meaning there are five possible locations for proteins of Gram-negative species (Extended Data Figure 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' (Table S1). 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 program 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: (i) conjugative, where plasmids encode all machinery required to transfer via conjugation; (ii) 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; (iii) non-mobilizable, where plasmids do not encode the genes necessary to be mobilized by themselves or other plasmids, and so cannot transfer via conjugation. 628 of the 4150 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 3522 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 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 which frequently live as commensals in their hosts. We classified non-pathogens as species which 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 Figure 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 which can invade either only one host species, or host species within the same genus or family. In 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[76], while *Pseudomonas syringae* acts as plant pathogen across multiple orders of flowering plants[77]. For more details and references to the literature used for this classification, please see Table S3.

We completed additional analyses for other two measures and proxies of environmental variability, the details and results of which can be found in Supp. Info. 1. In brief, we used previously published data which classified the habitat diversity of species using 16S RNA environmental datasets across five broad habitats: water, wastewater, sediment, soil and host[65,66]. We also supplemented this with information from the literature for species not included in the published data. We used this to ask whether species which 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 % 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 was collected from panX[74] (http://pangenome.tuebingen.mpg.de), since this calculates the pangenome using the same method across all of our species.

Pathogenicity categorisation of extracellular proteins

We used MP3[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 approximately 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 which 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 (Markov Chain Monte Carlo generalised linear mixed effects model) package in R with phylogeny a random effect[41,78]. 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[41,78]. 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 Table S2. 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[79,80]. A detailed description of MCMCglmm can be found elsewhere[41,78].

The response variable in all of our analyses is either a proportion or a measure calculated from proportions. Proportion data is bound between 0 and 1 and has a non-normal distribution. To control for this, all proportion data in our analyses has 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 (Figure S1). We used a recently published maximum likelihood tree using 16S ribosomal protein data as the basis for our phylogeny[64]. 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 (Fig S1, Table S2). Methods are also available to control for uncertainty in phylogenetic reconstruction[83,84], although we have not done this here.

Data Availability Statement

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

Code Availability Statement

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

Acknowledgements

We thank: Craig MacLean, Kevin Foster, Laurence Belcher, Chunhui Hao, and especially Eduardo Rocha for their helpful comments; James Robertson for providing plasmid mobility data from the MOBsuite database; 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. We also thank Alex Washburne and three anonymous reviewers for comments which greatly improved the manuscript. 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 S1, S2 and S3, and T.W.S. wrote and put together S4. All authors commented on and approved the manuscript for submission.

Competing Interests

The authors declare no competing interests.

References

1. Foster, K. R. Social behaviour in microorganisms. in *Social Behaviour* (eds. Szekely, T., Moore, A. J. & Komdeur, J.) 331–356 (Cambridge University Press, 2010). doi:10.1017/CBO9780511781360.027.
2. McNally, L., Viana, M. & Brown, S. P. Cooperative secretions facilitate host range expansion in bacteria. *Nat. Commun.* **5**, (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.* **118**, (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, table of contents (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.* **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 CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Micro 3: 711-721. *Nat. Rev. Microbiol.* **3**, 711–21 (2005).
14. Smith, J. The social evolution of bacterial pathogenesis. *Proc. R. Soc. Lond. B Biol. Sci.* **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 Biol. Sci.* **280**, 20130400 (2013).
18. Dimitriu, T. *et al.* Genetic information transfer promotes cooperation in bacteria. *Proc. Natl. Acad. Sci.* **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. Publ. Ecol. Soc. Am.* **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 University Press, 1991).
25. Grafen, A. The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **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*. (OUP Oxford, 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 Biol. Sci.* **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. *bioRxiv* 2021.02.11.430745 (2021) doi:10.1101/2021.02.11.430745.
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**, (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).

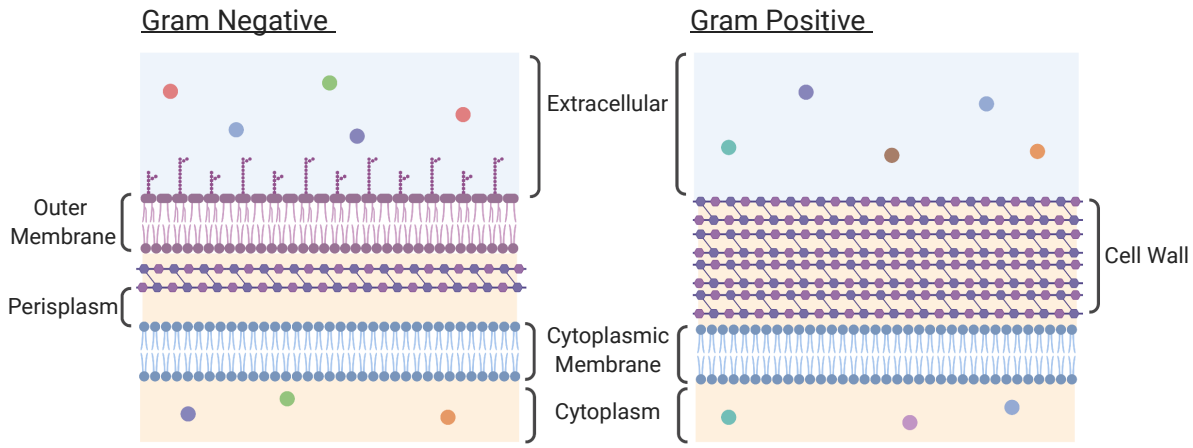
36. Nocelli, N., Bogino, P. C., Banchio, E. & Giordano, W. Roles of Extracellular Polysaccharides and Biofilm Formation in Heavy Metal Resistance of Rhizobia. *Materials* **9**, 418 (2016).
37. Ciofu, O., Beveridge, T. J., Kadurugamuwa, J., Walther-Rasmussen, J. & Høiby, N. Chromosomal β-lactamase is 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.* 1–13 (2021) doi:10.1038/s41579-020-00497-1.
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. Genomics* **4**, (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. Genomics* **6**, (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. work(s):, W. D. H. R. 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. U. S. A.* **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.* **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.e3 (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 (2017).
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. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
65. 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).

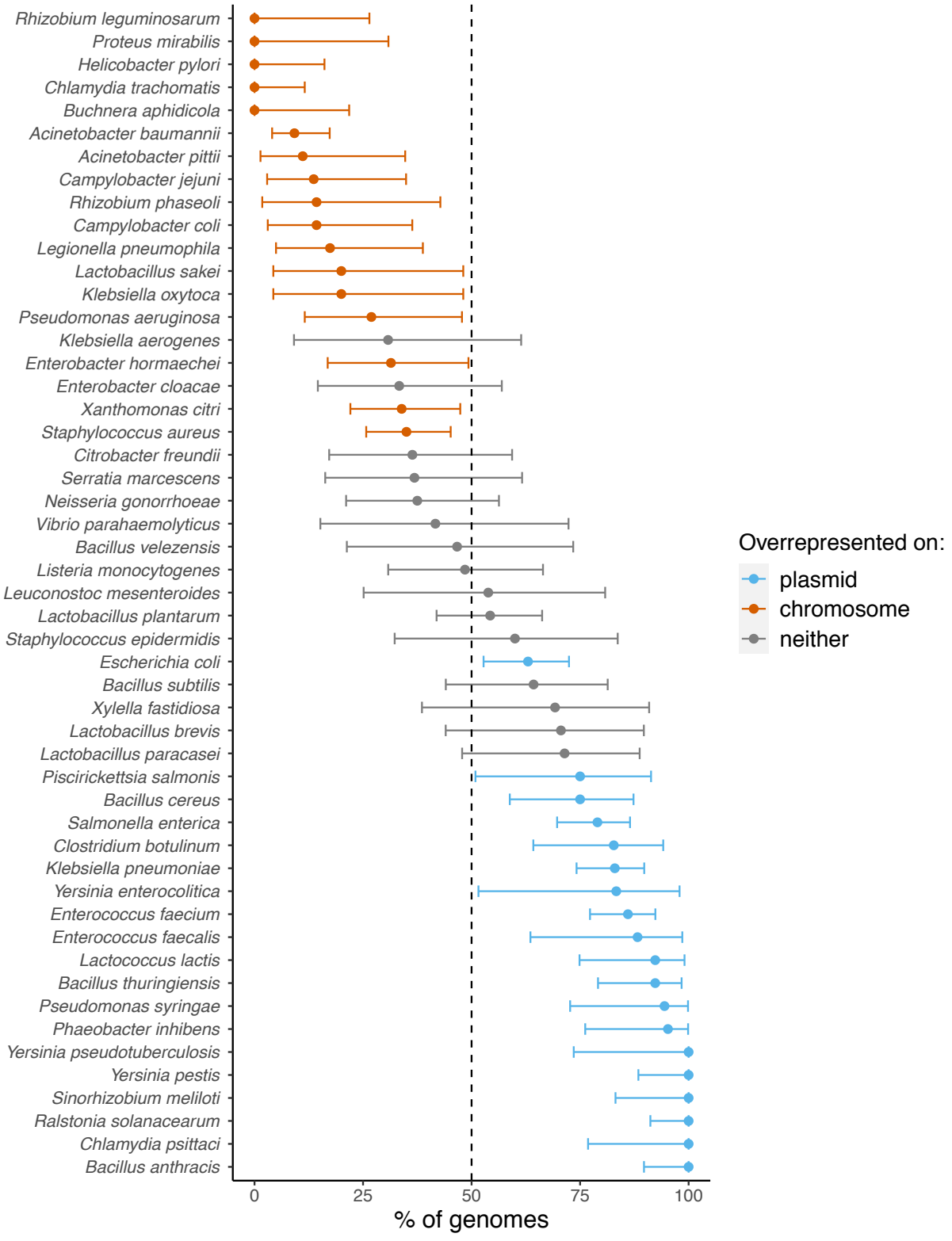
66. Kümmerli, R., Schiessl, K. T., Waldvogel, T., McNeill, K. & Ackermann, M. Habitat structure and the evolution of diffusible siderophores in bacteria. *Ecol. Lett.* **17**, 1536–1544 (2014).
67. Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L. & Brüssow, H. Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.* **6**, 417–424 (2003).
68. Burrus, V. & Waldor, M. K. Shaping bacterial genomes with integrative and conjugative elements. *Res. Microbiol.* **155**, 376–386 (2004).
69. 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).
70. Rodríguez-Rubio, L. *et al.* Extensive antimicrobial resistance mobilization via multicopy plasmid encapsidation mediated by temperate phages. *J. Antimicrob. Chemother.* **75**, 3173–3180 (2020).
71. Ramsay, J. P. & Firth, N. Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Curr. Opin. Microbiol.* **38**, 1–9 (2017).
72. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 3801–3806 (1999).
73. 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).
74. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* **46**, e5 (2018).
75. Gardy, J. L. & Brinkman, F. S. L. Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* **4**, 741–751 (2006).
76. 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).
77. 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).
78. Hadfield, J. D. MCMCglmm Course Notes. Available at cran.us.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf. (2019).
79. Nakagawa, S. & Schielzeth, H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **4**, 133–142 (2013).

80. Nakagawa, S., Johnson, P. C. D. & Schielzeth, H. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J R Soc Interface* 11 (2017).
81. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinforma. Oxf. Engl.* **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).

Extended Data Figures



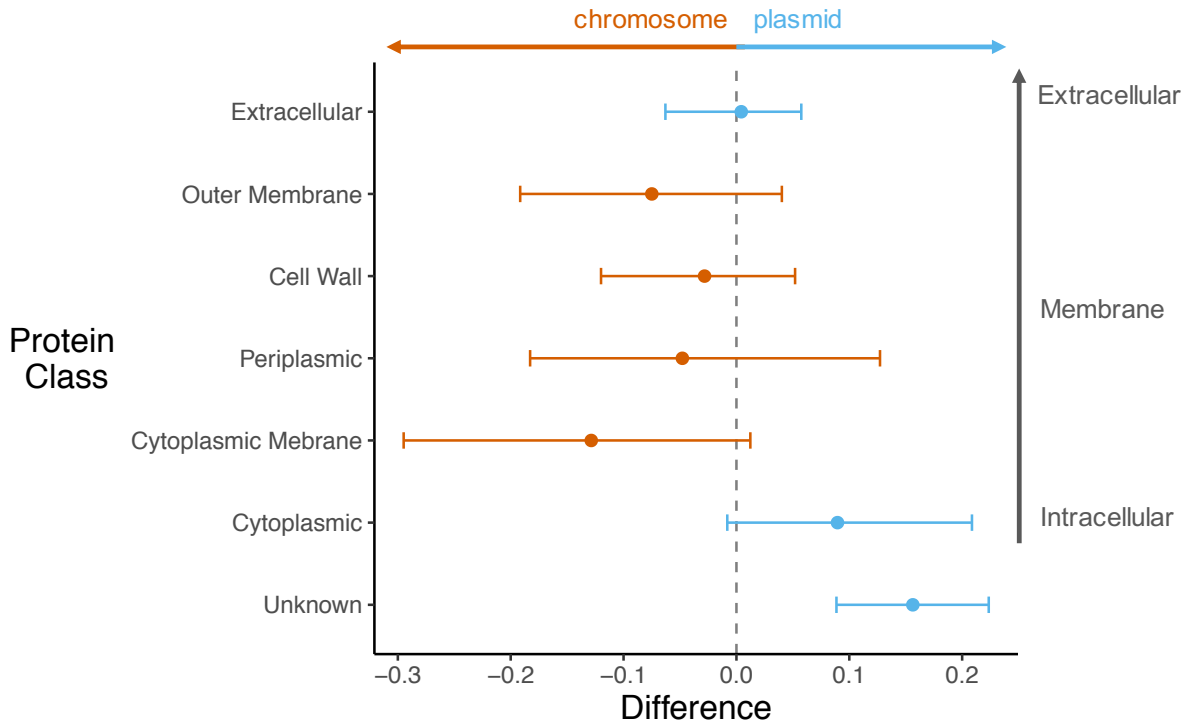
Extended Data Figure 1. Protein subcellular localisations.
Visualisation of all possible subcellular locations predicted by PSORTb. The left panel shows a cross-section of a typical Gram-negative bacterium and the right panel shows the equivalent for a Gram-positive bacterium. Both kinds of bacteria 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 Figure 2. Substantial variation within and between species in the genomic location of extracellular proteins.

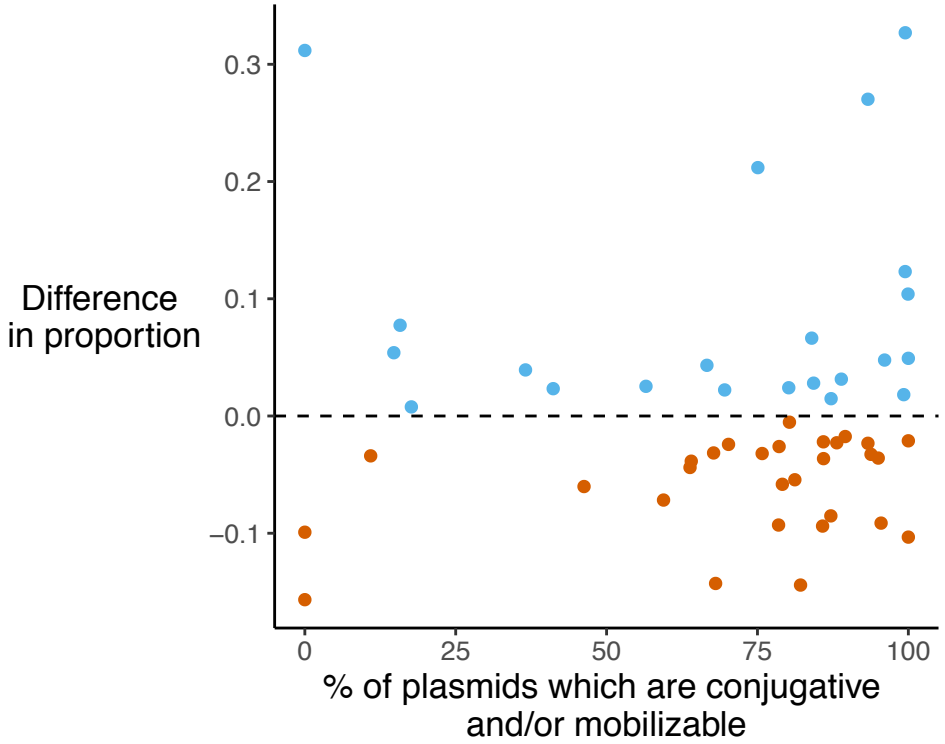
The x-axis is the % of genomes in each species where the proportion of plasmid proteins predicted as extracellular is greater than the proportion of chromosome proteins predicted as extracellular. Crucially, this considers only whether the plasmid proportion is greater than the

chromosome proportion for each genome, rather than also considering the magnitude of the difference (Figure 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 overrepresented on plasmids. Species in red have <50% of genomes where plasmid > chromosome extracellular proportion, meaning extracellular proteins are significantly overrepresented on chromosomes. Species in grey have a 95% CI which overlaps 50%, so extracellular proteins are not significantly overrepresented on either plasmids or chromosomes in these species.



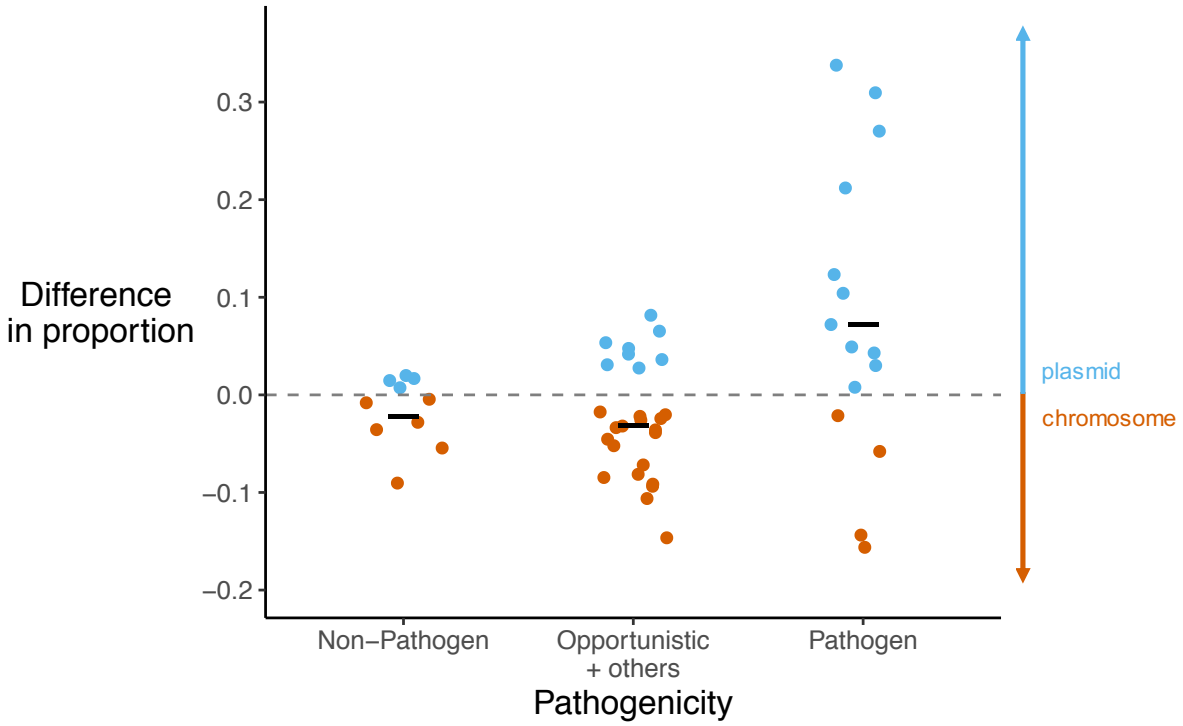
Extended Data Figure 3. Difference in plasmid and chromosome proportion for all protein classes predicted by PSORTb.

The x-axis is the difference in plasmid and chromosome extracellular proportions, as in Figure 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 on the difference in plasmid and chromosome proportion across all species, accounting for phylogeny and sample size. The only proteins significantly overrepresented in either direction are unknown proteins, which make up a higher proportion of plasmid proteins in all species we analysed.



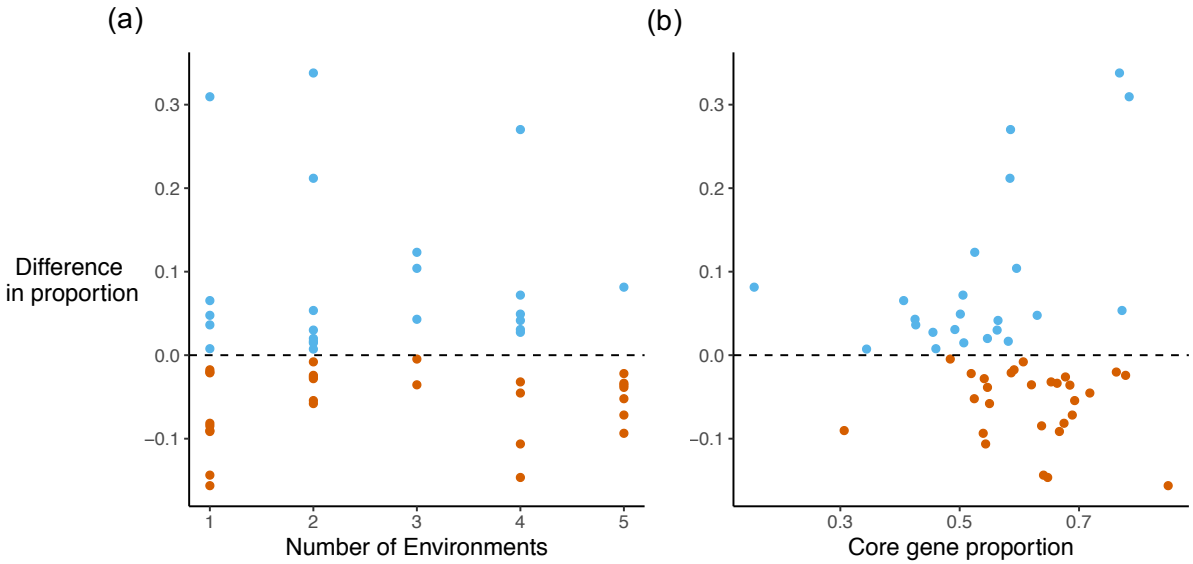
Extended Data Figure 4. No effect of plasmid mobility on the difference in plasmid and chromosome proportion of genes coding for extracellular proteins.

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 Figure 2. Each dot is the mean for all genomes in a species. Species in blue are those with genes coding for extracellular proteins overrepresented on plasmids, while species in red have genes coding for extracellular proteins overrepresented on chromosomes.



Extended Data Figure 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 overrepresented on plasmids, while species in red have genes coding for extracellular proteins overrepresented on chromosomes. Species were categorised 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 Figure 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 overrepresented on plasmids, while species in red are those with extracellular proteins overrepresented on chromosomes. For both these measures, we found no significant correlation with the genomic location of genes coding for extracellular proteins across species.

Chapter 3: Plasmid size, mobility and range

Abstract

Many characteristics of bacterial plasmids are highly variable, both within and between species. First, the size of plasmids can vary by several orders of magnitude. Second, many plasmids are capable of transferring to other cells via a process called conjugation, while others are incapable of transferring via this process. Third, some plasmids are unique to a bacterial species, while others have been sequenced across multiple bacterial classes. Each of these three characteristics could be seen as analogous to the 'life-history' traits often studied in animals and plants, which describe different aspects of a species' lifecycle. Understanding whether and how life-history traits may correlate with one another can help us understand the evolution of these traits. Here, we analysed how the size, mobility and range of 3522 plasmids from 51 diverse species of bacteria correlate with one another. We found that plasmid mobility is positively correlated with plasmid range. Additionally, we found that plasmids of different mobilities had different directions of correlation between plasmid range and plasmid size. Together, our analyses provide a comprehensive study of plasmid variation, providing a basis for future work.

Introduction

Plasmids are semi-autonomous, usually circular segments of DNA capable of replicating independently from their host cell's chromosome(s) (Stewart & Levin 1977; Levin *et al.* 1979; Dietel *et al.* 2018). Plasmids are widespread across bacteria and archaea, and have even been found inside some eukaryotic cells, such as yeast (Broach *et al.* 1982; Rodríguez-Beltrán *et al.* 2021). In bacteria, plasmids are often defined as carrying only 'accessory' genes (Tazzyman & Bonhoeffer 2015). These are genes not necessary to survival of the cell, but still may provide useful functions. Many plasmids are capable of transferring to neighbouring cells in a process called conjugation (Smillie *et al.* 2010). This is a form of horizontal gene transfer, where genes transfer to other individuals within the same generation, rather than vertically via descendants. This conjugation process is largely controlled by genes on the plasmids themselves (Smillie *et al.* 2010; Rodríguez-Beltrán *et al.* 2021).

Not all plasmids are capable of transferring via conjugation, and of those that are, not all have complete control of their own transfer. In general, we can group plasmids into three classes based on their mobility (Smillie *et al.* 2010). First, conjugative plasmids carry all the genes

necessary for the conjugation process, and can transfer between cells independently. Second, mobilizable plasmids are also capable of transferring via conjugation, but cannot do this alone. Instead, they code for only a subset of conjugation genes, and so rely on the presence of a conjugative plasmid to code for the rest of the process. Third, non-mobilizable plasmids cannot transfer via conjugation at all. They lack the crucial *oriT* gene, which allows a plasmid to be pulled through the conjugative tube called the pili. Without this gene, they are incapable of being transferred via this process, regardless of the presence of other plasmids. Therefore, these differences mean that mobility can vary substantially between plasmids (Smillie *et al.* 2010).

Plasmids differ not just in mobility, but also in their potential range of bacterial hosts. While some plasmids have been sequenced in only a single species of bacteria, others are capable of transferring to a variety of other species (Redondo-Salvo *et al.* 2020). For example, the pPCP1 plasmid is unique to *Yersinia pestis*, and carries several genes important for the species' virulence (Rajanna *et al.* 2010). One the other hand, some IncP plasmids have been sequenced in a diverse range of species across the Proteobacteria phyla (Klümper *et al.* 2015). Additionally, some of these can even jump to other phyla, with an analysis of bacteria in a soil sample indicating plasmid transfer between Gram-positive Firmicutes and Gram-negative Actinobacteria (Klümper *et al.* 2015).

Plasmids also vary in size by several orders of magnitude (Smillie *et al.* 2010; Shintani *et al.* 2015; Rodríguez-Beltrán *et al.* 2021) (Figure 1). Some plasmids are so small that they carry only the genes necessary for them to replicate. In contrast, other plasmids can be huge, reaching up to a third of the length of their hosts' chromosome. These very large plasmids are often called 'megaplasmids', though at what size a plasmid becomes a megaplasmid, and a megaplasmid becomes a second chromosome, is arbitrary and inconsistent across species. Additionally, the term 'chromid' has been coined for sequences which appear plasmid in origin, but which now function more as a secondary chromosome (Harrison *et al.* 2010) (Figure 1).

Clearly, there are several characteristics of plasmids that can vary substantially. Plasmid mobility, range and size each describe a feature of the 'life history' of plasmids. These could be considered analogous to the life history traits that evolutionary biologists study in animals, such as lifespan, offspring size and age at maturity (Stearns 1992). A key question is whether and how life history traits correlate with one another (Stearns 1983). Life history traits can be

positively correlated if natural selection selects for a high value of one trait while indirectly selecting for a high value of another trait. Alternatively, life history traits can be negatively correlated if there is a trade-off between an organism's ability to maximise both (Stearns 1989). Consequently, we may also expect that the mobility, range and size of plasmids could be correlated.

Understanding the presence and direction of these correlations can give insights into how selection may be acting on different life history traits (Stearns 2000). Therefore, exploring correlations between different characteristics of plasmids could also be useful for understanding plasmid evolution. Some studies have examined how plasmid size varies with respect to plasmid mobility, and others have explored how the range of potential hosts varies across plasmids (Smillie *et al.* 2010; Shintani *et al.* 2015; Rodríguez-Beltrán *et al.* 2021). However, how these three 'life-history' traits correlate with one another has not been explored together and in detail, especially across a wide diversity of plasmid sequences.

Here, we used a dataset of 3522 plasmid sequences from 51 diverse bacterial species to test how plasmid size, plasmid mobility and plasmid range vary with respect to each other. We also considered three other characteristics which may be expected to correlate with one or more these traits: the number of protein coding genes, the number of genes coding for extracellular proteins and the lifestyle of the plasmids' host species. Together, these analyses provide an initial step into how different characteristics of plasmids are correlated with one another, potentially providing insights into how selection acts on these candidate life-history traits of plasmids.

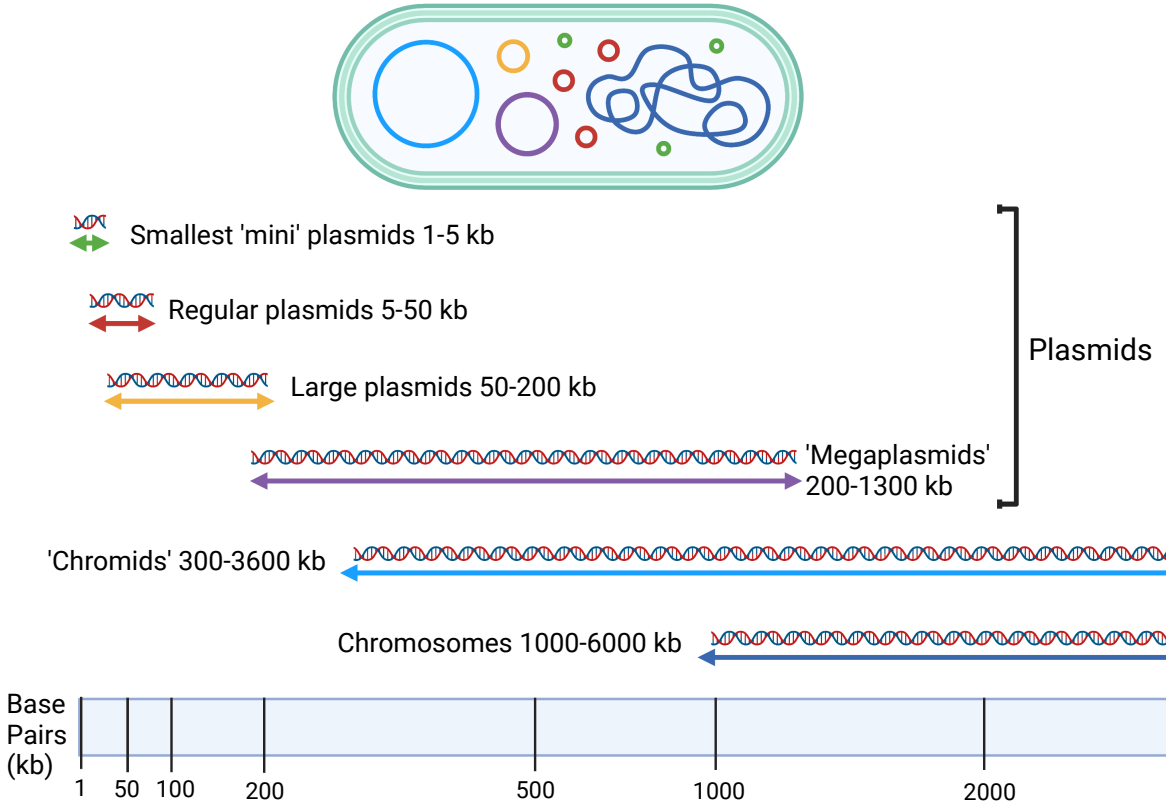


Figure 1. Plasmids, megaplasmids, chromids and chromosomes.

Illustration of the range of sizes for different genomic elements in bacteria. All numbers refer to the number of base pairs. Units are kilobases (kb), meaning 1 kb is 1000 bases. Plasmids range from 1 to 1300 kb, and includes 'megaplasmids' which are >200kb. 'Chromids' are between 300 and 3600 kb, and function as secondary chromosomes. Bacterial chromosomes range from 1000 to 6000 kb. Made using Biorender.com.

Method

Dataset

We used the dataset of chromosomes and plasmids from 51 species analysed in Chapter 2. Specifically, we used the 3522 plasmid sequences for which we were able to obtain mobility estimates using MOB-suite (Robertson & Nash 2018). Plasmids were categorised into one of three levels of mobility: (i) non-mobilizable plasmids, which cannot be transferred via conjugation; (ii) mobilizable plasmids, which can be transferred via conjugation but require the machinery of a conjugative plasmid; (iii) conjugative plasmids, which code for all the machinery required for conjugation and can therefore transfer independently.

In addition to providing a mobility prediction, MOB-suite also provided information on how widely a plasmid is distributed, which we will refer to as plasmid range. Plasmid range is a measure of the breadth of the different bacterial hosts a plasmid is carried in. Specifically, it is defined as the highest taxonomic rank of the genomes in which a plasmid (or plasmids very similar to it) is found. For example, a plasmid found only in genomes of *Yersina sp.* would have a plasmid range of 'genus', while a plasmid found in a number of Gammaproteobacteria species would have a plasmid range of 'class'. In general, the higher the taxonomic rank of genomes carrying the plasmid, the larger the plasmid's range. Each plasmid was assigned one of seven plasmid ranges: species, genus, family, order, class, phylum and domain.

Statistical analysis

In this chapter, we have analysed the data using two approaches: (i) considering each plasmid as an independent data point; (ii) controlling for both the phylogeny of the plasmids' host species and the number of plasmids per species. Which of these approaches is more appropriate can depend on the question being asked.

In Chapter 2, when we examined whether more mobile plasmids carried proportionally more genes coding for extracellular proteins than less mobile plasmids, we analysed whether this pattern was consistent across bacterial species. Therefore, we first calculated the correlation within species before comparing these across species, and then controlled for the phylogenetic non-independence between species using a phylogeny. This approach ensured that any results were not biased by an artificially larger number of similar plasmids from a few species.

However, rather than use plasmids as replicates to understand patterns across bacterial species, here we are considering patterns across plasmids themselves. We do not have a phylogeny for plasmids, and they are unlikely have a common single origin. Consequently, we are unable to control for shared plasmid history, although we would like to. Bacterial phylogeny does not necessarily reflect plasmid evolutionary history, especially as many plasmids in the dataset may exist in species other than that of the genome they were sequenced in here. Therefore, our alternative approach is to consider plasmids as individual data points, with the option of further controlling for species phylogeny and sample size as random effects in the model.

However, considering each of the 3522 plasmids as an individual data point could cause misleading significant results. Our results in Chapter 2 showed how an analysis using individual plasmids as data points could produce statistically significant differences, despite the fixed effect of the model only explaining 1.5% of the variance in the response variable. This suggests the importance of examining effect sizes as well as p-values, particularly when analysing very large datasets.

For the analyses in this chapter, we have used a mix of these approaches. We have considered plasmids as independent data points, both with and without controlling for their host species' phylogeny and plasmid range. We have also reported and discussed the effect sizes of our statistical models to better explore the biological importance of any significant results. For consistency, we have used the R^2 value to report the effect size, which is generally defined as the proportion of variance in the dependent variable explained by the independent variable. As a proportion, it is bound between 0 and 1, though it is often expressed as a percentage. As discussed in Chapter 2, a minimum of 5-10% of variance explained is reasonable for many areas of evolutionary biology (Cohen 1988; Jennions & Møller 2003; Crawley 2014).

Results

Plasmid size and protein coding genes

To better understand different aspects of plasmid size, we first examined how plasmid sequence length correlated with the number of protein coding genes. In bacterial genomes, there is very little redundancy in the genome, with the vast majority of base pairs making up part of a gene. This is in contrast to eukaryotic genomes, which frequently have large stretches of DNA between genes (Bobay & Ochman 2017). This means that the size of a plasmid, in terms of the number of base pairs in its sequence, is usually highly correlated with the number of genes the plasmid carries.

This is the case for my dataset, where the number of base pairs was highly correlated with the number of protein coding genes (Figure 2) (ANOVA: estimate=3.23×10^6, t=331.61, $p<0.001$, $R^2=0.969$; MCMCglmm: posterior mean=0.001, pMCMC<0.001; $R^2=0.962$). For the rest of this Chapter, I will use the number of base pairs as a measure of plasmid size.

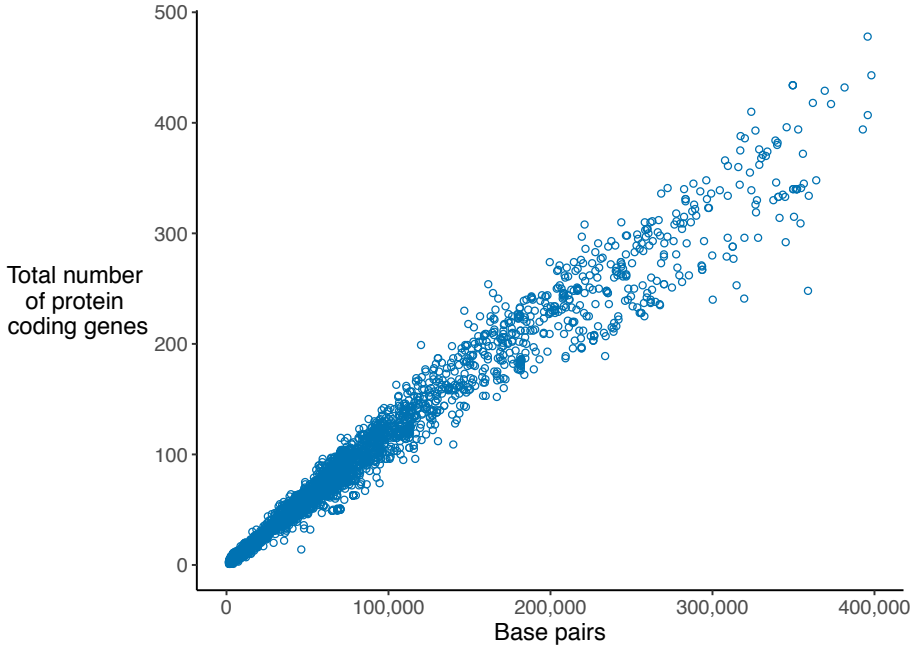


Figure 2. Correlation between two measures of plasmid size.

Each plasmid is represented as a blue circle. The x-axis indicates the number of base pairs making up the plasmid sequence, and the y-axis indicates the number of protein coding genes on the plasmid. The two measures are highly correlated.

Correlation between plasmid mobility and plasmid range

Next, we tested the extent to which plasmid mobility and plasmid range were correlated. Although we may expect this to be the case in general, the two measures are different in what they are estimating, and so may not be perfectly correlated. While plasmid mobility is a proxy for whether, and if so how frequently, a plasmid can be transferred via conjugation, the range of a plasmid indicates how far it has actually spread across different bacteria species.

An example of when these two measures may differ is if a non-mobilizable plasmid had been present in the ancestor of a large number of bacterial species, and many of these species retained the plasmid. Similarly, a conjugative plasmid could have transferred rapidly within a species, but never end up in other species, either due to the lifestyle of the species or systems like restriction modification preventing its uptake. Finally, there is the possibility that some plasmids predicted as non-mobilizable are capable of transferring to other cells by mechanisms other than conjugation.

However, despite these caveats, plasmid mobility and plasmid range were positively correlated (Figure 3) (Chi Sq; $\chi^2 = 538$, df=12, $p < 0.001$). Additionally, when comparing the range of the mobility categories to one another, conjugative plasmids have a broader range than non-mobilizable plasmids (T-test; Conjugative = 3.42, Non-mobilizable = 2.63, t=12.75, df = 2255, $p<0.001$, $R^2=0.067$), and mobilizable plasmids also have a broader range than non-mobilizable plasmids (T-test; Mobilizable = 3.58, Non-mobilizable = 2.63, t=13.8, df = 2350, $p<0.001$, $R^2=0.076$). In contrast, there was no real difference between the range of conjugative and mobilizable plasmids, since less than 0.3% of the variance in host-range was explained by whether a plasmid was conjugative or mobilizable (T-test; Conjugative = 3.42, Mobilizable = 3.58, t=-2.44, df=2254, p=0.015, $R^2=0.0026$). Together, these results suggest that the ability to conjugate, even if only occasionally in the case of mobilizable plasmids, increases the range of genomes in which a plasmid is found.

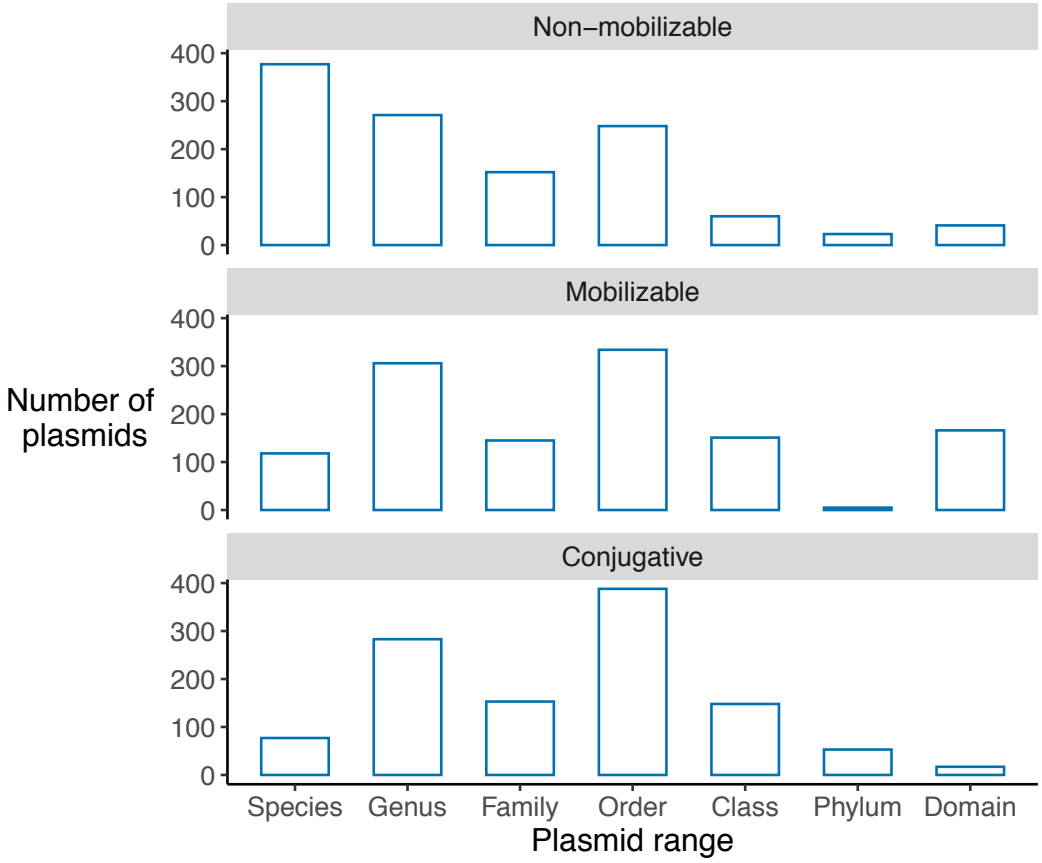


Figure 3. Plasmid range and plasmid mobility.

The range of plasmids increases along the x-axis, and the panels indicate the three types of plasmid mobility. The bars indicate the number of plasmids in

each plasmid range and mobility combination. Overall, plasmid range is broader with increasing mobility of plasmids.

How does plasmid size vary with respect to plasmid mobility and plasmid range?

Figure 4 shows the distribution of plasmid size across all 3522 plasmids in our dataset. At just 1506 base pairs, a non-mobilizable *Escherichia coli* was the smallest plasmid. In contrast, a conjugative *Pseudomonas aeruginosa* plasmid was the largest, at 398,807 base pairs. The mean plasmid size, indicated by a dashed vertical line in Figure 4, was 73,407 base pairs.

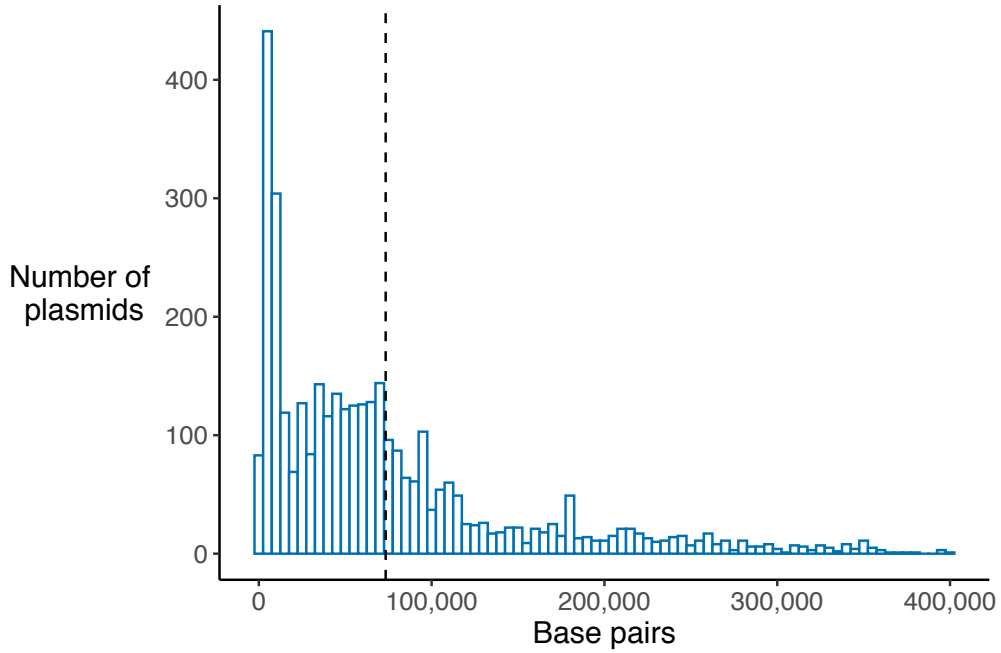


Figure 4. Frequency distribution of plasmid size.
The x-axis is the number of base pairs in each plasmid sequence, split into groups which increase by 5000 base pairs along the axis. The y-axis is the number of plasmids which are in each of these groups. The dotted line is the mean number of base pairs across all plasmids in the dataset.

We then examined how plasmid size varied with plasmid mobility (Figure 5). The mobility of plasmids had a significant effect on their size (ANOVA on three groups; F=341.8, df=3485, $p<0.001$, $R^2= 0.163$). Specifically, conjugative plasmids were significantly larger than non-mobilizable plasmids, while mobilizable plasmids were significantly smaller than non-

mobilizable plasmids (ANOVA on three groups, R^2=0.164; Conj. compared to non-mob.: estimate=43150, t=14.97, p<0.001; Mobilizable compared to non-mobilizable: estimate=-32465, t=-11.52, p<0.001). We found the same results when controlling for species' phylogeny and number of plasmids per species (MCMCglmm on three groups, R^2=0.078; Conj. compared to non-mob: posterior mean = 45990, 95% CI=40125 to 51241, pMCMC<0.001; Mob. compared to non-mob.: posterior mean = -20458, 95% CI = -25434 to -15240, pMCMC<0.001).

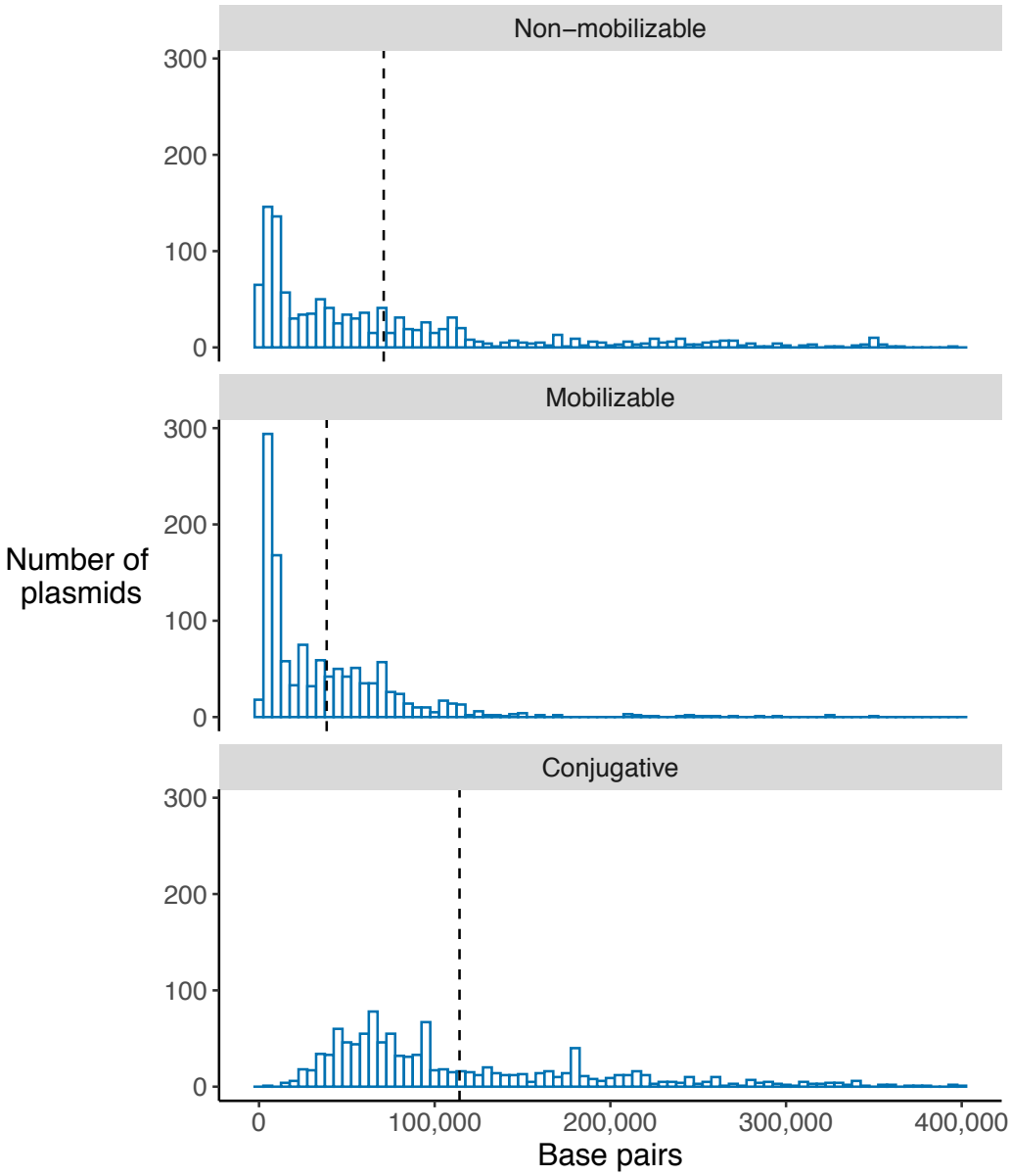


Figure 5. Plasmid size and plasmid mobility

We categorised plasmids into one of three categories of mobility: non-mobilizable, mobilizable and conjugative, with a panel for each. The x-axis is the number of base pairs in each plasmid sequence, split into groups which increase by 5000 base pairs along the axis. The y-axis is the number of plasmids in our dataset which are in each of these groups. The dotted line is the mean number of base pairs across all plasmids for each category of mobility.

Next, we examined how plasmid size correlated with plasmid range (Figures 6 & 7). When analysing all plasmids, regardless of mobility, and considering plasmid range as a continuous variable, there was no correlation between plasmid range and plasmid size (Figure 6) (ANOVA: slope estimate=-527, t=-0.69, p=0.49, R^2=<0.001; MCMCglmm: posterior mean = -1282, 95% CI=-2879 to 312, pMCMC=0.136, R^2<0.001).

Next, to explore the correlation between plasmid size and range with respect to mobility, we analysed the three mobility categories separately (Figure 7). For non-mobilizable plasmids only, there was no correlation between plasmid size and plasmid range (Figure 7) (ANOVA: slope estimate=-493, t=-0.33, p=0.74; MCMCglmm: posterior mean=5191, 95% CI= 1952 to 8122, pMCMC=0.002, R^2 <0.01). However, for both mobilizable and conjugative plasmids, there was a significant correlation between plasmid size and plasmid range (Figure 7). The correlation was negative for mobilizable plasmids and positive for conjugative plasmids; however, the effect sizes of these correlations were quite small, and were further reduced once we controlled for host species bacterial phylogeny (ANOVAs; Mobilizable: slope estimate=-6166, t=-9.28, p<0.001, R^2=0.065; Conjugative: slope estimate=12848, t=7.89, p<0.001, R^2=0.052) (MCMCglmm; Mobilizable: posterior mean=-5469, 95% CI=-7043 to -4133, pMCMC<0.001, R^2=0.029; Conjugative: posterior mean=14207, 95% CI = 10444 to 18184, pMCMC<0.001, R^2=0.030).

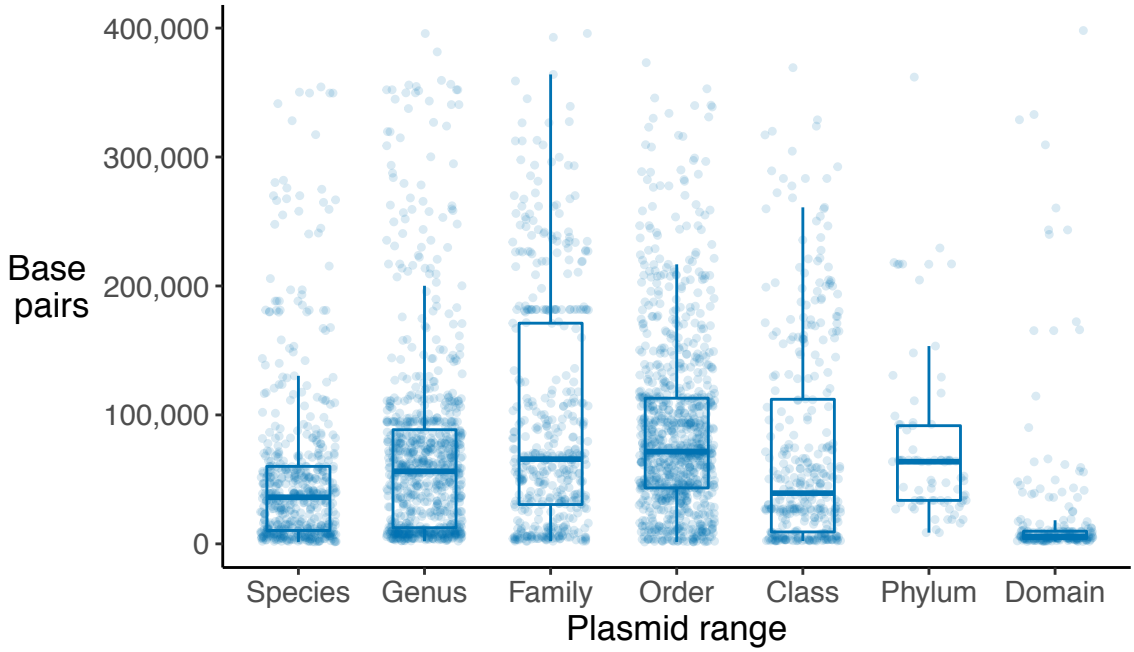


Figure 6. Plasmid range and plasmid size.

Each plasmid is represented by a circle. Plasmids are split into one of seven taxonomic ranges, which is increasingly broad along the x-axis. The y-axis is the number of bases in each plasmid sequence. Overall, there was no correlation between plasmid size and plasmid range.

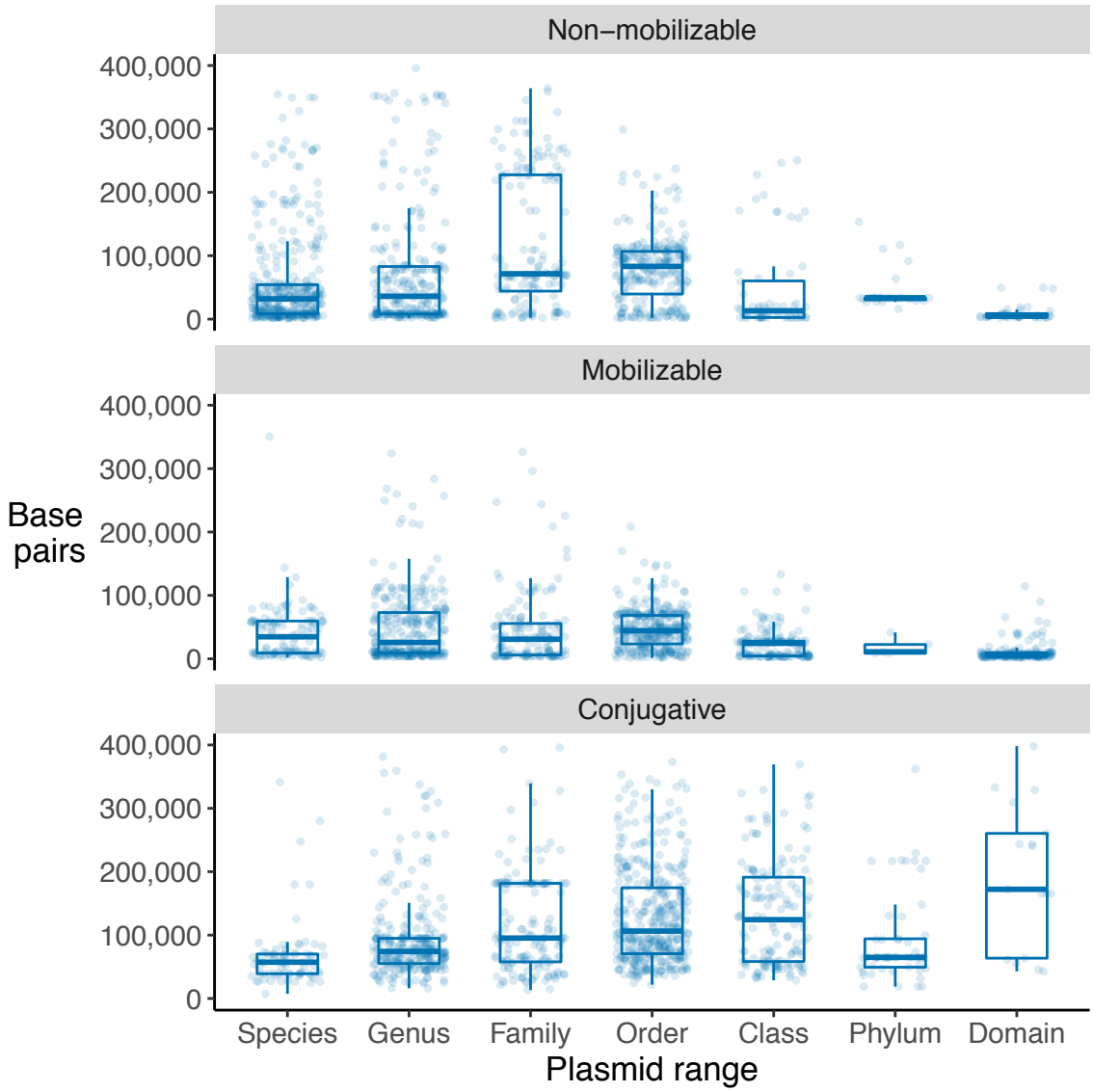


Figure 7. Plasmid range, mobility and size.

Each plasmid is represented by a circle. Data is shown as in Figure 5, but with plasmids additionally categorised into one of three mobility categories, with one panel for each. Mobilizable plasmids were smaller with increasing range, while conjugative plasmids were larger with increasing range.

Does plasmid size correlate with species lifestyles?

We then considered how plasmid size varied across species with different pathogenicity and host-ranges, rather than only examining individual plasmids themselves. In Chapter 2, we found that pathogenic species with a broad host-range had plasmids with a higher proportion of genes coding for extracellular proteins, relative to their chromosome(s). We also found that this was not due to differences in the mobility of plasmids of broad host-range species.

Therefore, in this Chapter, we instead examined whether these species differed in the size of their plasmids.

When controlling for host species phylogeny and number of plasmids per species, we found no difference in the size of species' plasmids between the three pathogenicity and host-range groups (Figure 8) (MCMCglmm of three groups, R^2=0.038; Narrow host-range compared to broad host-range pathogens: posterior mean=-24126, 95% CI=-76485 to 31509, pMCMC=0.368; Non-pathogens compared to broad host-range pathogens: posterior mean=-46698, 95% CI=-96408 to 8181, pMCMC = 0.09).

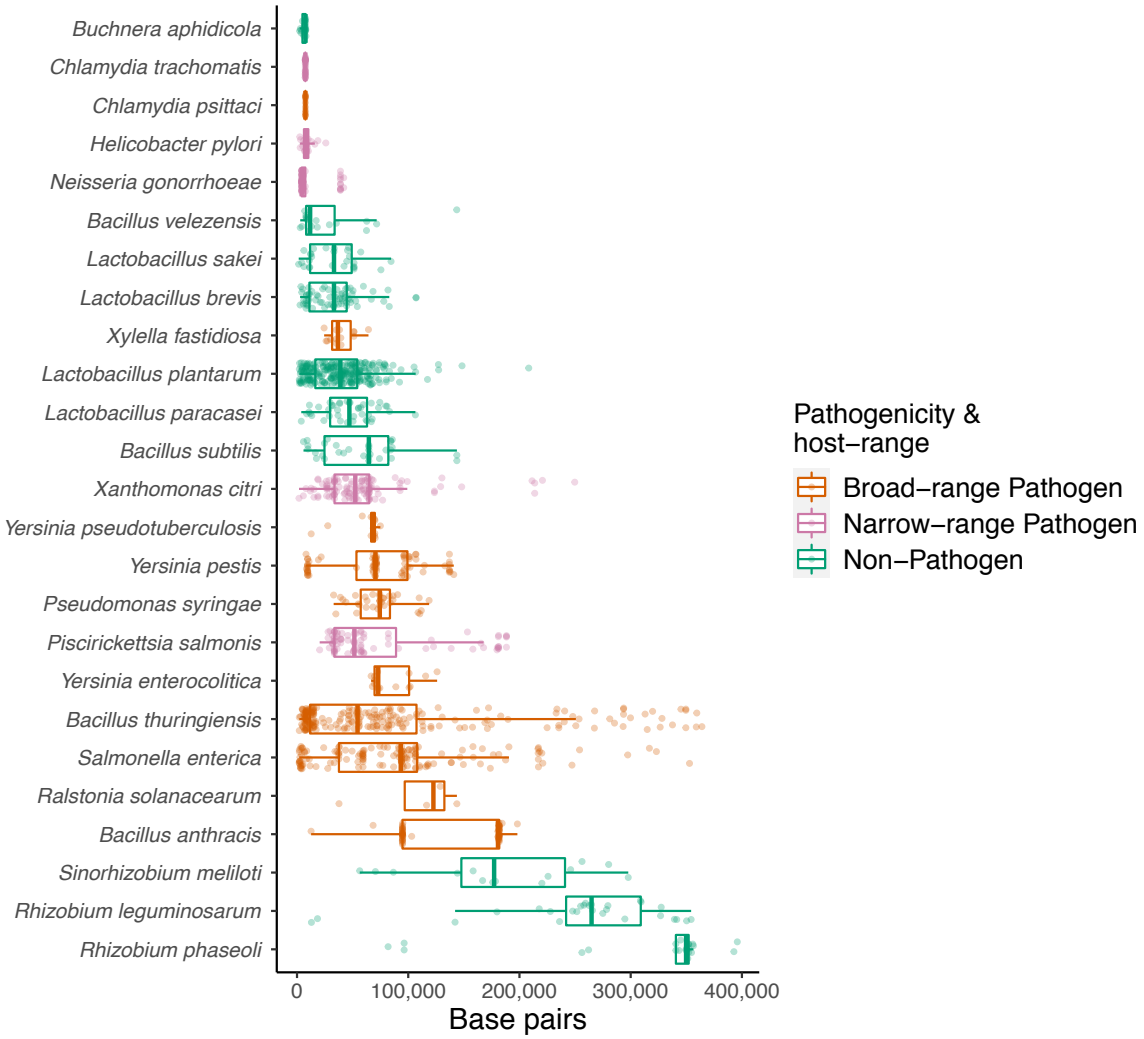


Figure 8. Plasmid size and species' pathogenicity and host-range.
Each plasmid is represented by a circle. The x-axis indicates the number of base pairs in the sequence of each plasmid. Plasmids are grouped into the species they were sequenced in, labelled on the y-axis. Boxplots indicate the

distribution of plasmid sizes for each species. Circles and boxplots are coloured by the pathogenicity and host-range of the species. Only species we could be sure were either pathogens or non-pathogens are shown. Pathogens are further split into broad or narrow host-range species.

Plasmid size and genes coding for extracellular proteins

We also examined whether there was a correlation between plasmid size and the proportion of plasmid genes coding for extracellular proteins. We found no significant correlation between plasmid size and the proportion of plasmid genes that coded for extracellular proteins (ANOVA: slope estimate = 1.9×10^{-8}, t=1.56, p=0.12; MCMCglmm: $R^2<0.01$) (Figure 9). While some very small plasmids do appear to have particularly high proportions of extracellular proteins in Figure 9, this is likely an artefact of calculating proportions of a very small total, rather than an actual effect.

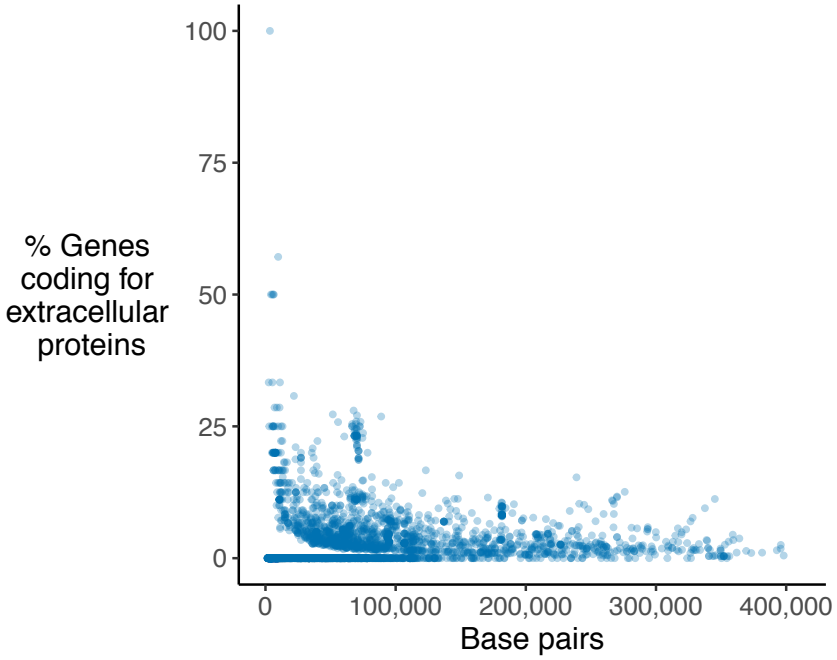


Figure 9. Plasmid size and extracellular proteins.

Each plasmid is represented as a circle. The x-axis is the number of base pairs in each plasmid sequence. The y-axis is the proportion of genes which code for extracellular proteins of each plasmid. Overall, there is no correlation.

Discussion

We found a number of correlations between the different characteristics of plasmids we examined. First, we found that plasmid mobility and plasmid range were positively correlated, suggesting that the ability to conjugate does increase a plasmids' potential range of bacterial hosts. Second, we also found that plasmids within each of the three classes of mobility differed significantly in both their size and range. While we found no significant correlation between the size and host-range across all plasmids, we did find that the size of mobilizable plasmids was negatively correlated with range, while the size of conjugative plasmids was positively correlated with range. However, we also found a number of characteristics that were not correlated with one another. Specifically, we found no correlation between the size of plasmids and either the lifestyle of their host species or the proportion of their genes coding for extracellular proteins. Overall, our results suggest that selection may act differently on the size and range of plasmids from the three mobility classes.

For conjugative plasmids, the positive correlation of plasmid size with plasmid range could be driven by two potential factors (Figure 7). First, conjugative plasmids with a broader range may be under stronger selection to carry multiple genes, in order to provide benefits to multiple potential hosts. Second, selection on these plasmids to lose genes, so as to reduce their cost to hosts, could be weaker. This is because their fitness will be less dependent on the fitness of their host cell, and more dependent on ability to transfer horizontally. Whether this positive correlation between plasmid size and plasmid range is due to stronger selection for gene gain, or weaker selection for gene loss, will require further analysis.

While both conjugative and mobilizable plasmids are capable of transferring, our results suggest that plasmid size is likely to be under different selection pressures between the two. We found that mobilizable plasmids were generally much smaller than conjugative plasmids (Figure 5). One explanation for this is simply that they do not carry genes for conjugation. However, we found that the plasmid size of mobilizable plasmids was correlated with plasmid range, but in the opposite direction to conjugative plasmids (Figure 7). This suggests that rather than simply not carrying conjugation genes, mobilizable plasmids could be under different selection pressures related to their size.

Mobilizable plasmids have been described as 'hijacking' the machinery of conjugative plasmids (Gérard Guédon *et al.* 2017). This description suggests that they may act as largely selfish mobile elements, potentially spreading between cells at the expense of conjugative plasmids, and providing little benefits to the host. In this case, we may expect that mobilizable plasmids with the broadest range may be under the strongest selection to lose any genes that are not required for transfer, and so these would be the most 'parasitic' of the mobilizable plasmids. In contrast, those which only spread between a narrow range of hosts could be under weaker selection to lose genes. This is because vertical inheritance may be a more important aspect of their fitness, and so carrying some genes which are beneficial to the host could make them more likely to be maintained. Overall, this suggests that conjugative and mobilizable plasmids of a similar range may have different selection pressures acting upon their size.

For non-mobilizable plasmids, the lack of correlation between plasmid size and plasmid range is perhaps unsurprising (Figure 7). This is because a non-mobilizable plasmid with a broad range is still unable to transfer, and so here plasmid range is not a measure of potential mobility for these plasmids. What it means for a non-mobilizable plasmid to have a narrow versus a broad range is unclear. Figure 3 shows that the majority of non-mobilizable plasmids have a comparatively narrow range. Those with a broader range could be due to a distant species acquiring the plasmid through another mechanism of horizontal gene transfer. Alternatively, those plasmids could have been maintained in an entire bacterial lineage. Regardless, the lack of correlation suggests that selection on the size of non-mobilizable plasmids is unrelated to the potential for plasmids to transfer.

Limited evidence for a correlation of species lifestyle with plasmid size

We found relatively little evidence that a species' lifestyle correlated with the size of their plasmids. Specifically, we found that whether a species was pathogenic, and the host-range of the species, did not correlate with the size of their plasmids (Figure 8). This suggests that the environmental variability of species has little effect on relative selection for gene gain and loss in plasmids, at least within pathogens. However, in Figure 8, the three species with the largest plasmids all live inside the root nodules of plants. In contrast, the five species with the smallest plasmids all live inside the cells of other organisms, with some acting as pathogens and others as symbionts. Therefore, while plasmid size does not correlate with pathogenicity or pathogen host-range, this may suggest that other aspects of a species' lifestyle may exert similar pressures on the size of plasmids. This will require further analysis.

How could copy number affect selection on plasmid size?

An additional feature of plasmids, which we have so far not considered here, is that plasmids can exist in multiple copies per cell (San Millan *et al.* 2017; Rodríguez-Beltrán *et al.* 2020). This could be important for selection on plasmid size, because smaller plasmids have been shown to generally exist in higher numbers of copies (Zhong *et al.* 2011). High copy number could be beneficial to hosts, since it could allow high expression of the useful genes carried on plasmids. Therefore, plasmids could be under selection to reduce their size to increase their copy number. On the other hand, multiple copies of plasmids require more resources for plasmid replication and translation, suggesting that selection on the copy number of plasmids could further complicate the direction of potential selection on plasmid size.

Conclusions

Overall, our analyses provide an initial step in understanding how different candidate 'life-history' traits of plasmids correlate with one another. We found that plasmid mobility and plasmid range are positively correlated, suggesting that the ability to conjugate increases a plasmid's potential range of hosts. While we found no correlation between plasmid range and plasmid size across all plasmids, we did find that the size of conjugative plasmids was positively correlated with range, while the size of mobilizable plasmids was negatively correlated with plasmid range. Although the effect sizes of these correlations are small, they potentially suggest that the size of conjugative and mobilizable plasmids may be under different selection pressures depending on how aligned their fitness is with their host. To further explore these potential correlations, empirical analyses would be particularly helpful. For example, altering the size of plasmids experimentally could test explicitly how variation affects plasmid mobility and plasmid maintenance.

References

Bobay, L.-M. & Ochman, H. (2017). The Evolution of Bacterial Genome Architecture. *Front. Genet.*, 0.

Broach, J.R., Guarascio, V.R. & Jayaram, M. (1982). Recombination within the yeast plasmid 2mu circle is site-specific. *Cell*, 29, 227–234.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd edn. Routledge.

Crawley, M.J. (2014). *Statistics: An Introduction Using R*. John Wiley & Sons.

Dietel, A.-K., Kaltenpoth, M. & Kost, C. (2018). Convergent Evolution in Intracellular Elements: Plasmids as Model Endosymbionts. *Trends in Microbiology*, 26, 755–768.

Gérard Guédon, Virginie Libante, Charles Coluzzi, Sophie Payot, & Nathalie Leblond-Bourget. (2017). The Obscure World of Integrative and Mobilizable Elements, Highly Widespread Elements that Pirate Bacterial Conjugative Systems. *Genes*, 8, 337.

Harrison, P.W., Lower, R.P.J., Kim, N.K.D. & Young, J.P.W. (2010). Introducing the bacterial "chromid": not a chromosome, not a plasmid. *Trends Microbiol*, 18, 141–148.

Jennions, M.D. & Møller, A.P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology*, 14, 438–445.

Klümper, U., Riber, L., Dechesne, A., Sannazzarro, A., Hansen, L.H., Sørensen, S.J., *et al.* (2015). Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J*, 9, 934–945.

Levin, B.R., Stewart, F.M. & Rice, V.A. (1979). The kinetics of conjugative plasmid transmission: Fit of a simple mass action model. *Plasmid*, 2, 247–260.

Rajanna, C., Revazishvili, T., Rashid, M.H., Chubinidze, S., Bakanidze, L., Tsanava, S., *et al.* (2010). Characterization of pPCP1 Plasmids in Yersinia pestis Strains Isolated from the Former Soviet Union. *International Journal of Microbiology*, 2010, e760819.

Redondo-Salvo, S., Fernández-López, R., Ruiz, R., Vielva, L., de Toro, M., Rocha, E.P.C., *et al.* (2020). Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat Commun*, 11, 3602.

Robertson, J. & Nash, J.H.E. (2018). MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom*, 4.

Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R.C. & San Millán, Á. (2021). Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology*, 1–13.

Rodríguez-Beltrán, J., Sørum, V., Toll-Riera, M., Vega, C. de la, Peña-Miller, R. & Millán, Á.S. (2020). Genetic dominance governs the evolution and spread of mobile genetic elements in bacteria. *PNAS*, 117, 15755–15762.

San Millan, A., Escudero, J.A., Gifford, D.R., Mazel, D. & MacLean, R.C. (2017). Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat Ecol Evol*, 1, 0010.

Shintani, M., Sanchez, Z.K. & Kimbara, K. (2015). Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Front. Microbiol.*, 6.

Smillie, C., Garcillan-Barcia, M.P., Francia, M.V., Rocha, E.P.C. & de la Cruz, F. (2010). Mobility of Plasmids. *Microbiology and Molecular Biology Reviews*, 74, 434–452.

Stearns, S.C. (1983). The Influence of Size and Phylogeny on Patterns of Covariation among Life-History Traits in the Mammals. *Oikos*, 41, 173–187.

Stearns, S.C. (1989). Trade-Offs in Life-History Evolution. *Functional Ecology*, 3, 259.

Stearns, S.C. (1992). *The evolution of life histories*.

Stearns, S.C. (2000). Life history evolution: successes, limitations, and prospects. *Naturwissenschaften*, 87, 476–486.

Stewart, F.M. & Levin, B.R. (1977). The Population Biology of Bacterial Plasmids: A Priori Conditions for the Existence of Conjugationally Transmitted Factors. *Genetics*, 87, 209–228.

Tazzyman, S.J. & Bonhoeffer, S. (2015). Why There Are No Essential Genes on Plasmids. *Mol Biol Evol*, 32, 3079–3088.

Zhong, C., Peng, D., Ye, W., Chai, L., Qi, J., Yu, Z., *et al.* (2011). Determination of Plasmid Copy Number Reveals the Total Plasmid DNA Amount Is Greater than the Chromosomal DNA Amount in Bacillus thuringiensis YBT-1520. *PLoS ONE*, 6, e16025.

Chapter 4. Why do plasmids have an AT-bias?

Abstract

Plasmids are found in genomes across the bacterial tree of life. They are semi-autonomous segments of DNA, many of which are capable of transferring between different bacterial hosts. Plasmid sequences appear to have a high percentage of nucleotide bases which are A and T, relative to bacterial chromosomes. However, the reason for this AT-bias is unclear. A and T nucleotides are less costly to produce and present in higher quantities in bacterial cells than G and C bases. Therefore, AT-bias in plasmids could be an adaptation to reduce their cost. Alternatively, plasmid AT-bias could instead be a product of increased genetic drift and weaker purifying selection in plasmids leading to the accumulation of mutations, which are biased towards A and T bases, in plasmid sequences. Here, we used a dataset of 3522 plasmids to test key predictions arising from these two hypotheses. Overall, we found little support for the adaptation hypothesis, and comparatively better support for the hypothesis that increased genetic drift and weaker purifying selection in plasmids drives their consistent AT-bias via the accumulation of AT-biased mutations. Future work could look at other intracellular elements that display AT-bias, such as endosymbionts, to see if similar patterns are found. Additionally, we suggest that examining signatures of selection and/or drift in plasmid sequences could provide more insights than the correlational analyses presented here.

Introduction

Found in almost every bacterial species that has been sequenced, plasmids are circular segments of DNA which appear to play an important role in bacterial evolution. Although they carry genes which are usually considered as 'accessory' to their host's chromosome, many bacteria rely on plasmids for key parts of their lifestyle. Examples include the virulence plasmids identified in a diverse range of pathogens, and the pSym plasmids that carry genes for nodulation in Rhizobia species (Hale 1991; Cornelis *et al.* 1998; Ding & Hynes 2009). Additionally, many plasmids can transfer between cells in process known as conjugation, which is a form of horizontal gene transfer. Such plasmids either code for all the genes necessary for conjugation, and are referred to as conjugative plasmids, or carry only a subset of genes, and are referred to as mobilizable plasmids. Other plasmids cannot be transferred at all via this process, and are known as non-mobilizable plasmids.

Plasmid sequences have been observed to be particularly enriched with A and T bases, compared to G and C bases (Nishida 2012; Dietel *et al.* 2018). The relative proportion of AT bases compared to GC bases, often referred to as base content, is usually very stable across genomes of the same bacterial species, and across genes within a bacterial genome. While the apparent AT-bias of plasmid sequences has been widely observed and discussed, the reason for this is still unclear (Nishida 2012; Bohlin *et al.* 2017; Dietel *et al.* 2018, 2019). Here, we consider two major hypotheses for why plasmids appear to be consistently enriched with AT bases compared to their bacterial host's chromosome.

First, AT-bias in plasmids could be an adaptation to be less costly to their bacterial hosts (Figure 1). AT bases are generally less costly to produce than GC bases, and also present at higher concentrations inside cells (Dietel *et al.* 2019). Recently, an experimental study showed that plasmids which had a higher AT-content than their hosts' chromosome were less costly than those with similar or lower AT-content (Dietel *et al.* 2019). If plasmids are too costly relative to their benefit, hosts would be under selection to lose such plasmids. Therefore, increased AT-content of plasmids could reduce the cost of plasmids, allowing them to be maintained in the long-term. An analogous hypothesis has also been suggested for why many bacterial endosymbionts have AT-rich genomes, compared to their eukaryotic hosts (Rocha & Danchin 2002; Dietel *et al.* 2018).

Second, AT-bias could instead be the product of increased genetic drift, weaker purifying selection and higher rates of mutation in plasmids (Figure 1). Across bacteria, mutations are more likely to result in AT sites compared to GC sites (Hershberg & Petrov 2010; Hildebrand *et al.* 2010). This appears to be the case even in bacterial species with a low baseline AT-content. If mutation is higher in plasmids, and such mutations have a higher fixation rate, this could lead to plasmids having a higher AT-content relative to bacterial chromosomes.

There are several reasons why plasmids could have a higher mutation rate compared to chromosomes. Plasmids frequently exist in multiple copies per cell, which has been suggested to generate more opportunities for mutation compared to single copy plasmids (Rodríguez-Beltrán *et al.* 2020, 2021). Additionally, due to the potential for plasmid loss, many plasmids exist in only a subset of cells at any one time, potentially reducing their population size relative to their hosts. This could mean a weaker effect of purifying selection in plasmids of nearly-neutral deleterious mutations, and together with genetic drift lead to the accumulation of

mutations in plasmids (Rodríguez-Beltrán *et al.* 2021). Therefore, if mutations occur and are fixed at a higher rate in plasmids compared to chromosomes, this could explain the AT-bias of plasmids.

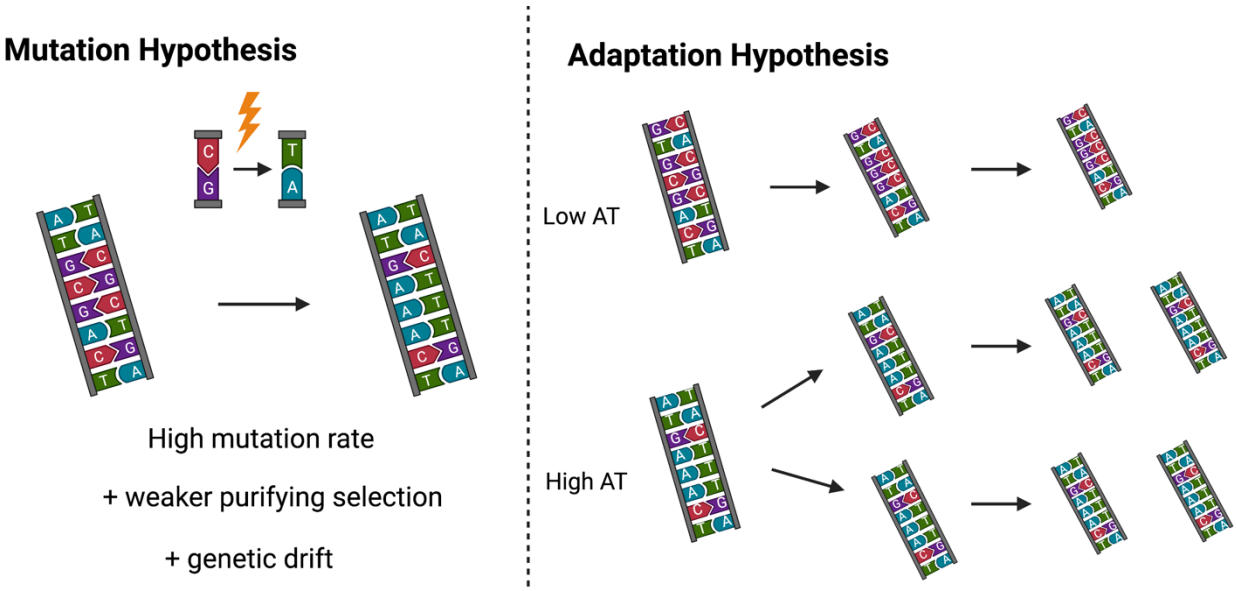


Figure 1. Two hypotheses for AT-bias in plasmids.
Segments of plasmid DNA sequences are illustrated, with AT shown in green/blue, and GC shown in purple/red. Mutation hypothesis: plasmid sequences become enriched with AT because genetic drift and weaker purifying selection leads to more mutations being fixed, which are AT-biased. Adaptation hypothesis: plasmids with a high AT-content are less costly to their hosts than low AT-content plasmids, meaning they will spread more throughout the population, and over time plasmids will become increasingly AT-biased. Created using Biorender.com.

These two hypotheses give different predictions for how AT-content would be expected to vary with respect to plasmid transferability. Plasmid mobility and range capture two different aspects of plasmid transferability: their ability to transfer via conjugation and the range of bacterial species they are found in, respectively. As expected, in Chapter 3 we found that plasmid mobility and range were positively correlated. In this Chapter, we will use plasmid mobility and range as two different ways to estimate plasmid transferability.

Plasmids with a lower mobility and narrower range are unlikely to spread much to other cells, in comparison with broad range conjugative plasmids. These low mobility and narrow-range

plasmids will therefore be more reliant on the fitness of their bacterial host for their own fitness, because their only way of getting into the next generation is vertical inheritance via their hosts' daughter cell.

Therefore, if AT-bias was an adaptation of plasmids to reduce their cost to their bacterial host, we would expect plasmids with lower transferability to be those under strongest selection to increase their AT-content.
Consequently, we can predict that:

(1) Plasmid AT-content should be negatively correlated with plasmid mobility.
    - The fitness of lower mobility plasmids is more aligned with their host, increasing selection to reduce their cost by increasing their AT-content.
(2) Plasmid AT-content should be negatively correlated with the range of hosts that carry the plasmid.
    - The fitness of narrower range plasmids is more aligned with their host, increasing selection to reduce their cost by increasing their AT-content.

In contrast, if AT-bias is instead due to mutations accumulating in plasmid sequences, this would be the case for all plasmids, regardless of their transferability. This would mean the effects of mutation, genetic drift and weaker purifying selection on the AT-content of plasmids would be similar, regardless of transferability.
Consequently, we can predict that:

(1) Plasmid AT-content should be uncorrelated with plasmid mobility.
    - Plasmids with different mobilities accumulate mutations at a similar rate.
(2) Plasmid AT-content should be uncorrelated with the range of hosts that carry the plasmid.
    - Plasmids with different ranges accumulate mutations at a similar rate.

To test the different predictions of these two hypotheses, we used a dataset of 3522 plasmids from 51 diverse bacterial species. We examined how both the AT-content of plasmids, and the AT-content of plasmids compared to their species' chromosomal AT-content, varied with respect to plasmid mobility and plasmid range.

Method

We used the dataset of chromosomes and plasmids from 51 species from Chapters 2 and 3. Specifically, we used the 3522 plasmid sequences for which we were able to obtain mobility and plasmid range predictions using MOB-suite (Robertson & Nash 2018).

AT-content of chromosomes

To control for variation in plasmid AT-content arising from variation in their host species' AT-content, we also compared plasmid AT-content to the baseline AT-content of their species. The AT-content is very similar across chromosomes of the same species, usually with a range of less than half a percent. Therefore, we collected the AT content of the chromosome(s) of each species' representative genome listed on NCBI. For the analyses in this chapter, we have compared plasmids from each species to this value.

Statistical Analysis

In Chapters 2 and 3, we discussed whether genomes and species can be considered as independent from one another. We concluded that because of shared ancestry, phylogenetic history of species needs to be controlled for in statistical analyses – as is common in evolutionary analyses.

However, when considering analyses using individual plasmids as data points, how to control for phylogenetic history becomes more difficult. This is because plasmids are themselves independent entities, and may have a different evolutionary history from their host cell. In Chapter 3, we analysed plasmid data in two different ways, either by considering plasmids as independent, or by controlling for the phylogeny and number of plasmids of the species each plasmid was sequenced in. We also discussed the importance of considering effect sizes, not just significance values, especially when analysing very large datasets.

In this Chapter, we will use the same two approaches as the previous chapter, noting any analyses in which the results are different, and discussing the potential reasons for this. The analyses on plasmids in this Chapter have an additional confounding variable, which is the baseline AT-content of the species they were sequenced in. Therefore, when testing for patterns of AT-content across plasmid mobility and/or range, we have analysed both the AT-content of

plasmids themselves, and also compared to the AT-content of the species' representative chromosome.

Results

Are plasmids AT-rich compared to their chromosomes?

To check that the plasmids in our dataset were indeed enriched with A and T bases, we first examined the extent to which plasmids in our dataset exhibited AT-bias. Figure 2 shows the distribution of the AT-content of each species' representative chromosome and all plasmids in our dataset. Overall, there was considerable variation in the AT-content of species' chromosomes. *Ralstonia solanacearum* had the lowest AT-content with 33.3%, while *Buchnera aphidicola* had the highest AT-content with 74.7%. There was similarly large variation in the AT-content of plasmids. The plasmid with the lowest AT-content in our dataset was a non-mobilizable plasmid sequenced in a *Xanthomonas citri* genome, with an AT-content of 31.4%, while a non-mobilizable *Buchnera aphidicola* plasmid had the highest AT-content at 76.3%.

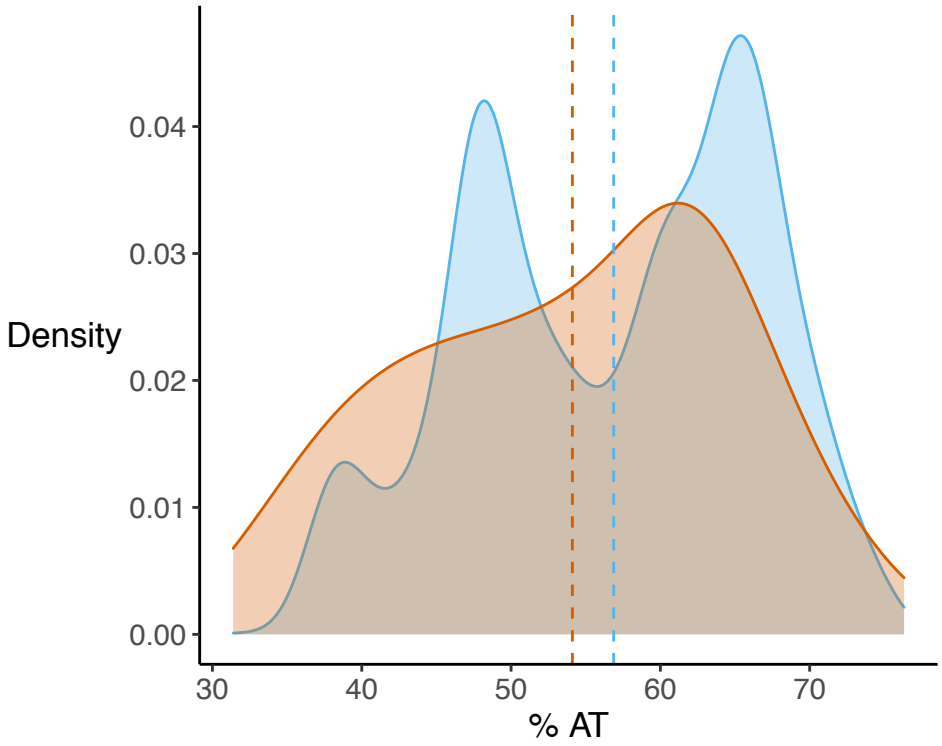


Figure 2. AT-content distribution of plasmids and chromosomes.
Density plots of the distribution of the percentage of bases that are A or T for plasmid sequences (blue) and chromosomes (red). The greater the density,

the more plasmids or chromosomes there are with that value of % AT. In this way a density plot is effectively a smoothed histogram. The dotted lines indicate the mean of each distribution. Overall, the mean of the % of AT bases is higher on plasmids than chromosomes, though there is much variation.

The mean plasmid percentage of A or T bases was on average 3.0% higher than the representative chromosomes (Unpaired t-test; mean plasmid % AT = 56.9, mean chromosome % AT = 53.9, $t=12.7$, $df=6972$, $p<0.01$, $R^2=0.023$). However, we also need to compare plasmids to the AT-content of the species they were sequenced in. This is required because, when considering plasmids as independent, 88% of the variance in plasmid AT-content was explained by the AT-content of their species' chromosome (ANOVA; slope estimate = 0.92, $t=159.9$, $p<0.01$, $R^2=0.88$) Similarly, when controlling for species phylogeny, the AT-content of their species' chromosome explained 74.6% of the variation in plasmid AT-content (MCMCglmm: posterior mean=0.780, 95% CI=0.706 to 0.855, pMCMC<0.001, $R^2=0.746$). Therefore, when testing the predictions of the two hypotheses, we needed to consider if plasmids had a higher AT-content than the chromosome, in addition to whether plasmids have a high AT-content in general.

Figure 3 shows the AT-content of every plasmid in each species and the AT-content of the representative chromosome for each species. Plasmids (blue circles) had consistently higher AT-content than their species' chromosomes (red dots). This result holds irrespective of whether we analysed with (a): a paired t-test of the species' chromosome AT-content compared to each plasmid's AT-content (Paired T-test; mean difference = 2.99, 95% CI = 2.88 to 3.11, $t=51.23$, $df=3487$, $p<0.001$, $R^2=0.428$); or (b) the difference in % AT-content of every plasmid compared to its species' representative chromosome, when controlling for host species' phylogeny and number of plasmids per species (Figure 4; MCMCglmm: posterior mean=2.76, 95% CI=1.826 to 3.614, pMCMC=0.002). Overall, plasmids displayed consistent AT-bias relative to their species' chromosomes, as has been observed and discussed previously (Nishida 2012; Dietel *et al.* 2018, 2019).

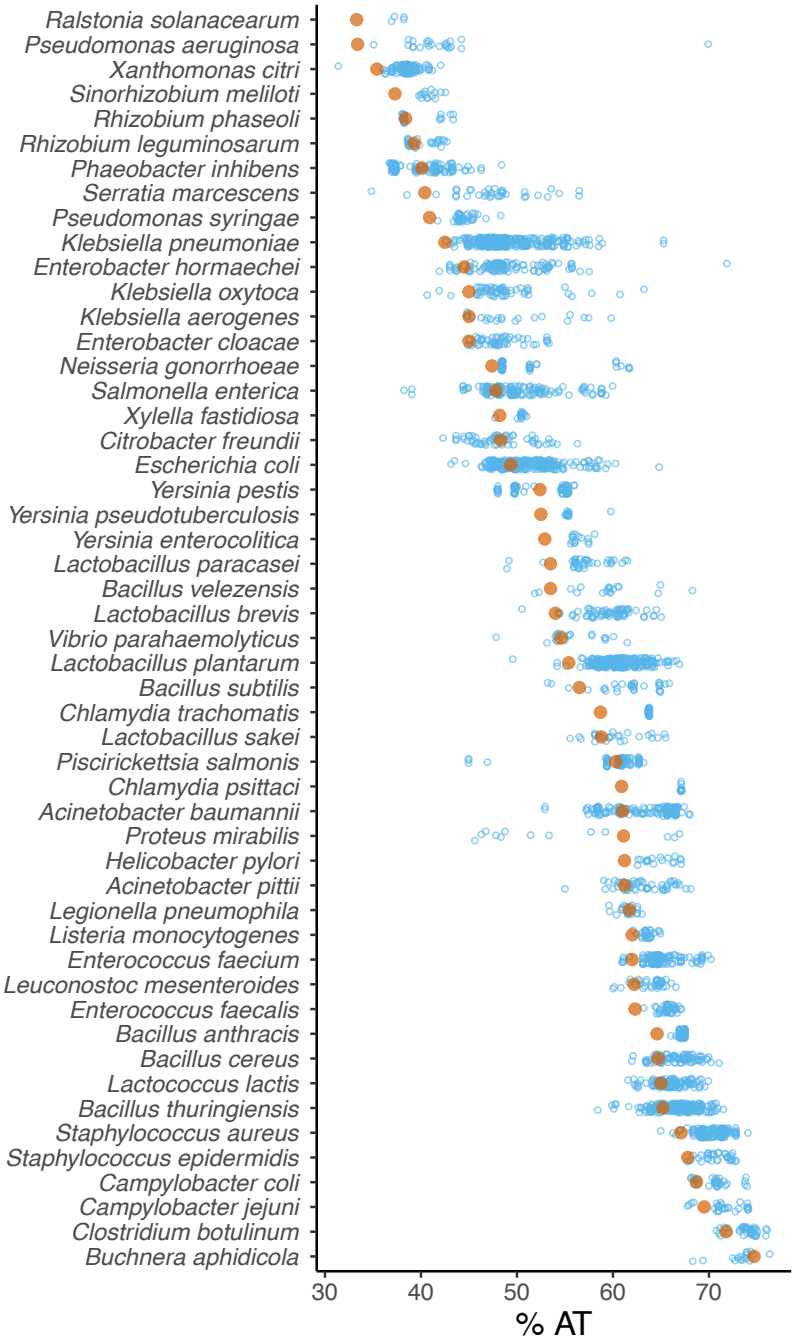


Figure 3. Plasmid AT-compared to their species' chromosome.

The y-axis shows all 51 species, and the x-axis is the % of bases that are A or T. Red dots are the % AT of each species' representative chromosome. Blue circles are individual plasmids. Very few plasmids are to the left of their species' red dot, indicating that plasmids consistently have a higher AT-content than the chromosome. This is true both for species with very low and very high chromosomal AT-content.

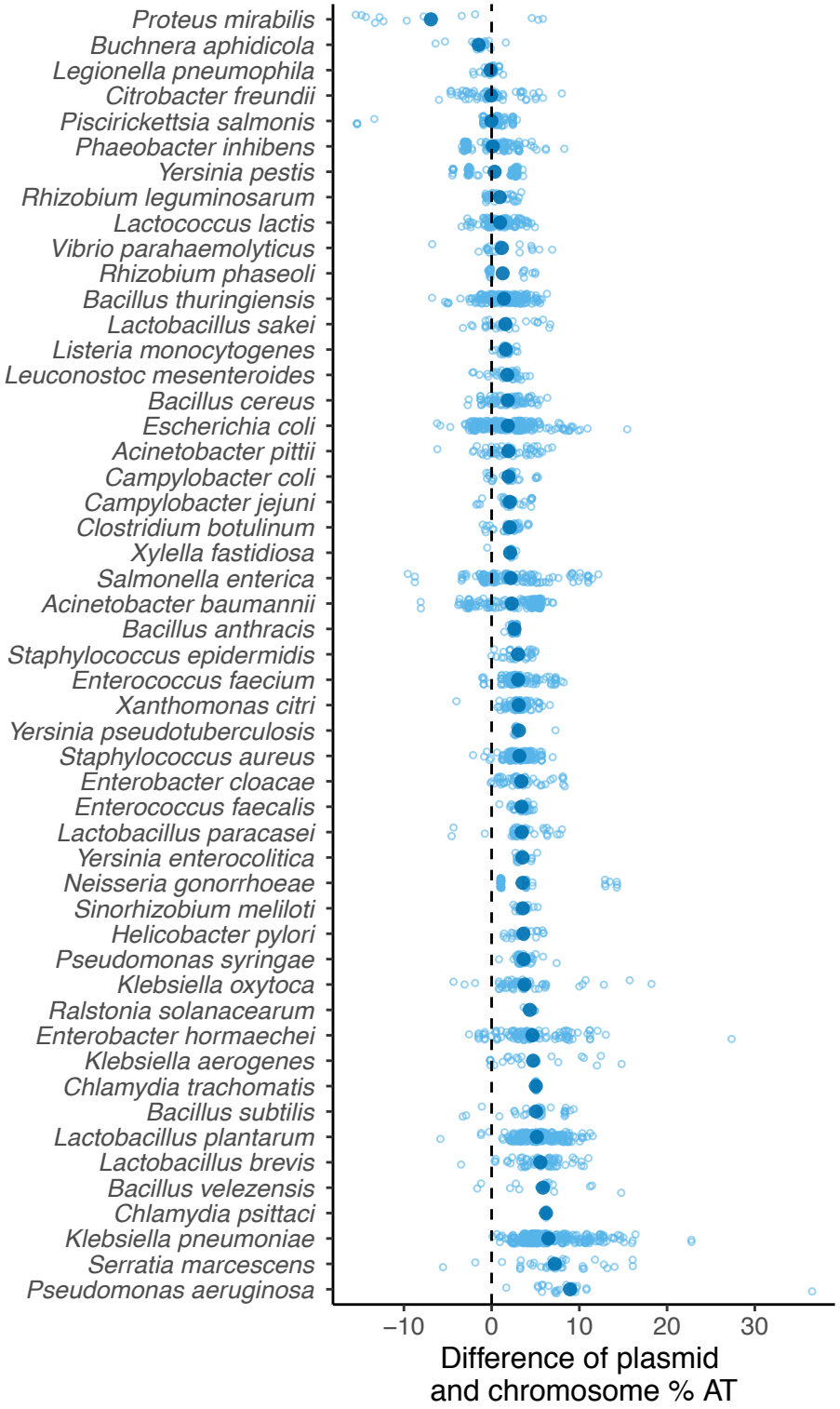


Figure 4. Difference in % AT of plasmids compared to chromosomes.

The y-axis shows all 51 species, and the x-axis is the difference in the percentage of bases that are A or T for each plasmid compared to its species' representative chromosome. For example, a plasmid with an AT-content of 50% from a species with a chromosome of 40% would have a value of 10 on

this graph. Each blue circle is a plasmid, and the dark blue dot is the mean difference for each species. Almost all species have a mean difference of above 0, meaning their plasmids have a higher AT-content than their chromosome.

How does plasmid AT-content vary with plasmid mobility and plasmid range?

To test the predictions of the two hypotheses for why this plasmid AT-bias is such a consistent feature of plasmids, we examined how plasmid AT-content varied with respect to plasmid mobility and plasmid range.

First, to test the prediction of the adaptation hypothesis that more mobile plasmids should have a lower AT-content, we examined how AT-content varied across the three classes of plasmid mobility: non-mobilizable, mobilizable and conjugative plasmids (Figure 5). When we considered plasmids as independent from one another, plasmid mobility explained approximately 5.5% of the variance in the percentage of bases that were A or T across plasmids (ANOVA with three groups; $F=103.2$, $df=3485$, $p<0.001$, $R^2=0.055$). However, when controlling for host species' phylogeny and number of plasmids per species, plasmid mobility explained less than 0.01% of plasmid AT-content (MCMCglmm with three groups, $R^2<0.01$; mob compared to non-mob: posterior mean= 0.024, pMCMC=0.856; conj compared to non-mob: posterior mean=0.065, pMCMC=0.610).

We found a significant but weak negative correlation between AT-content and plasmid mobility, when we considered plasmids as independent, consistent with the prediction of the adaptation hypothesis (ANOVA; slope estimate=-1.77, $t=-8.78$, $p<0.001$, $R^2=0.021$). However, when we controlled for host species phylogeny and plasmid number, this significant correlation disappeared (MCMCglmm; posterior mean = 0.033, 95% CI = -0.10 to 0.16, pMCMC=0.622, $R^2<0.001$).

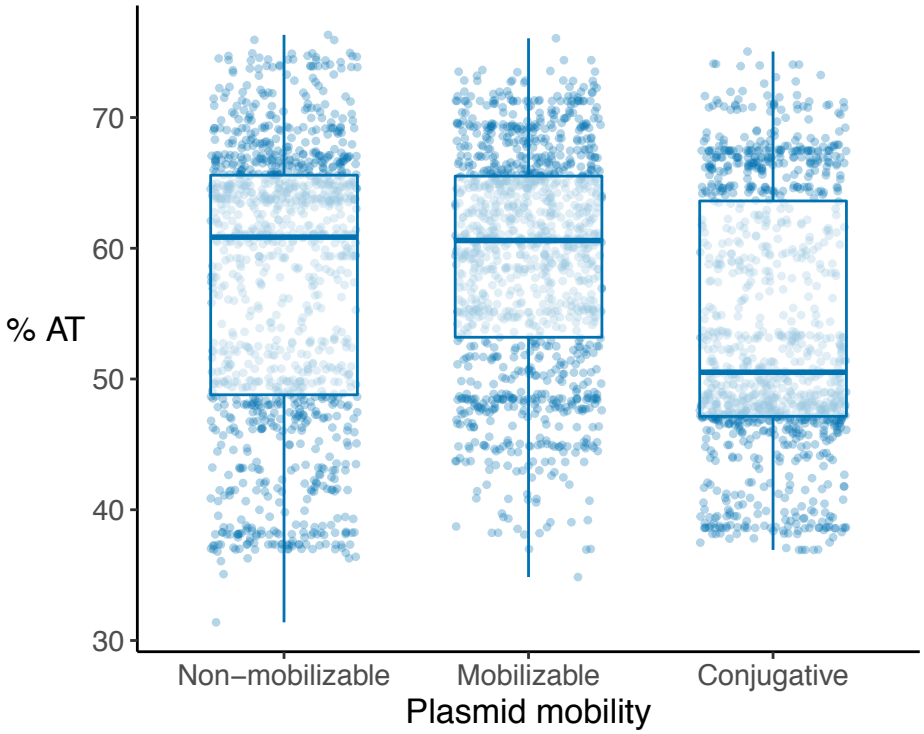


Figure 5. AT-content and plasmid mobility.

Plasmids, each shown as a blue circle, are categorised into one of three mobility classes: non-mobilizable, mobilizable and conjugative. The relative mobility of these classes increases along the x-axis. The y-axis shows the percentage of bases that are A or T. Overall, there is no difference between the three mobility classes.

To further test this prediction, while controlling for the species' baseline AT-contents, we also examined the effect of plasmid mobility on the difference in AT-content of plasmids compared to their species' chromosome (Figure 6). We found that plasmid mobility explained virtually none of the variance in the difference between plasmid and chromosome AT-content, at less than 0.1% (ANOVA: R^2 <0.01; MCMCglmm: R^2 <0.01).

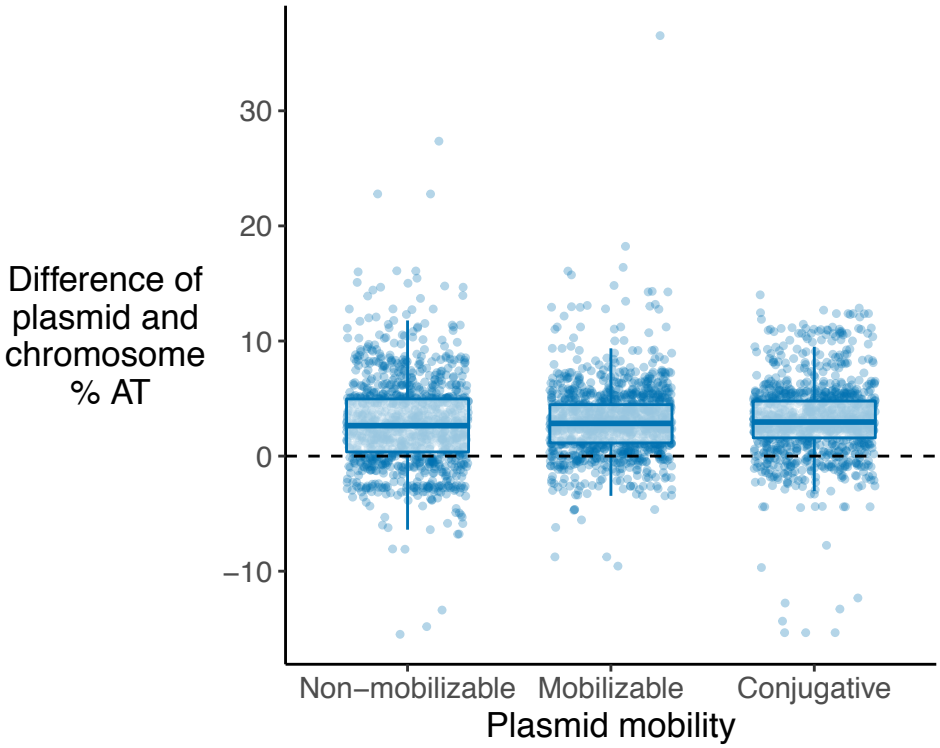


Figure 6. Difference in plasmid and chromosome AT-content and plasmid mobility

Plasmids, each shown as a blue circle, are categorised into one of three mobility classes: non-mobilizable, mobilizable and conjugative. The relative mobility of these classes increases along the x-axis. The y-axis shows the difference in the percentage of bases that are A or T for plasmids compared to their species' representative chromosome. Overall, there is no difference between the three mobility classes.

Next, to test the second prediction of the adaptation hypothesis, we examined whether plasmid AT-content was negatively correlated with plasmid range. If plasmid AT-bias is an adaptation to be less costly, plasmids with a lower transferability would be predicted to be those with the highest AT-content, because their fitness would be more aligned with the fitness of their host. Therefore, they would be under the strongest selection to reduce their cost to the host, to maximise their success via vertical inheritance. In addition to plasmid mobility, we used plasmid range as a way of estimating plasmid transferability.

Consistent with this prediction, when considering plasmids as independent data points, we found a significant negative correlation between plasmid range and the percentage of bases

which were A or T (Figure 7) (ANOVA; slope estimate = -1.74, t=-18.13, p<0.001; R^2=0.086). However, plasmid range only explained 8.6% of variation in plasmid AT-content, and this fell to less than 0.1% when we controlled for host species' phylogeny and number of plasmids per species (MCMCglmm; posterior mean = -0.06, 95% CI=0.007 to -0.142, pMCMC=0.106, R^2<0.001). The negative correlation between AT-content and plasmid range was also no longer significant (MCMCglmm; posterior mean=-0.061, 95%CI = -0.14 to 0.007, pMCMC=0.106, R^2<0.001).

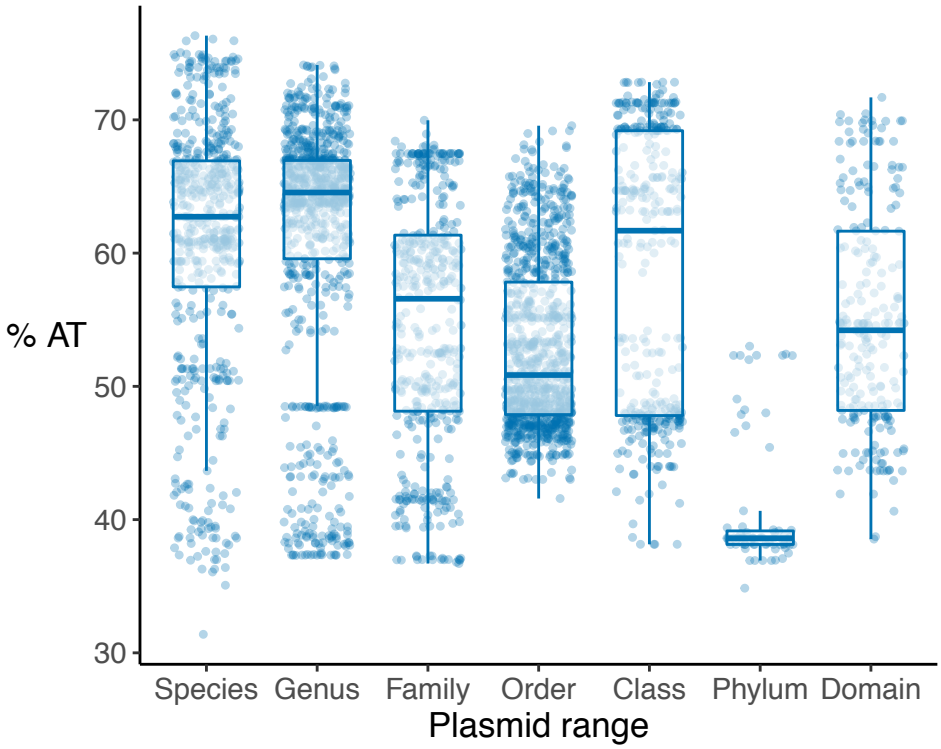


Figure 7. AT -ontent and plasmid range.

Plasmids, each shown as a blue circle, are categorised into one of seven plasmid ranges, depending on the taxonomic rank of the common ancestor of all genomes which contain a copy of that plasmid. The y-axis is the percentage of based which are either A or T. Overall, there is a significant but very weak negative correlation between % AT and plasmid range.

To further test this second prediction, we also examined each of the three categories of plasmid mobility separately (Figure 8). When we considered plasmids as independent data points, we found a significant negative correlation between the percentage of A or T bases and plasmid range for all three classes of plasmid mobility (ANOVA; Non-mobilizable: slope estimate =

-1.78, t=-9.54, p<0.001, R^2=0.072; Mobilizable: slope estimate=-0.90, t=-6.74, p<0.001, R^2=0.04; Conjugative: slope estimate=-3.68, t=-20.94, p<0.001, R^2=0.28). However, when we controlled for host species' phylogeny and number of plasmids per species, the negative correlation for mobilizable plasmids was no longer significant, and the correlation for non-mobilizable plasmids became positive (MCMCglmm; Non-mobilizable: posterior mean=0.16, 95 % CI=0.03 to 0.29, pMCMC=0.018, R^2<0.001; Mobilizable: posterior mean=-0.040, 95% CI=-0.15 to 0.07, pMCMC=0.478, R^2<0.001; Conjugative: posterior mean=-0.441, 95% CI=-0.57 to -0.30, pMCMC<0.001, R^2<0.001). Only conjugative plasmids still had a significant negative correlation between host-range and AT-content, but the effect size was extremely small. These results suggest there was no meaningful correlation for any of the plasmid mobility categories when plasmids were not considered independent.

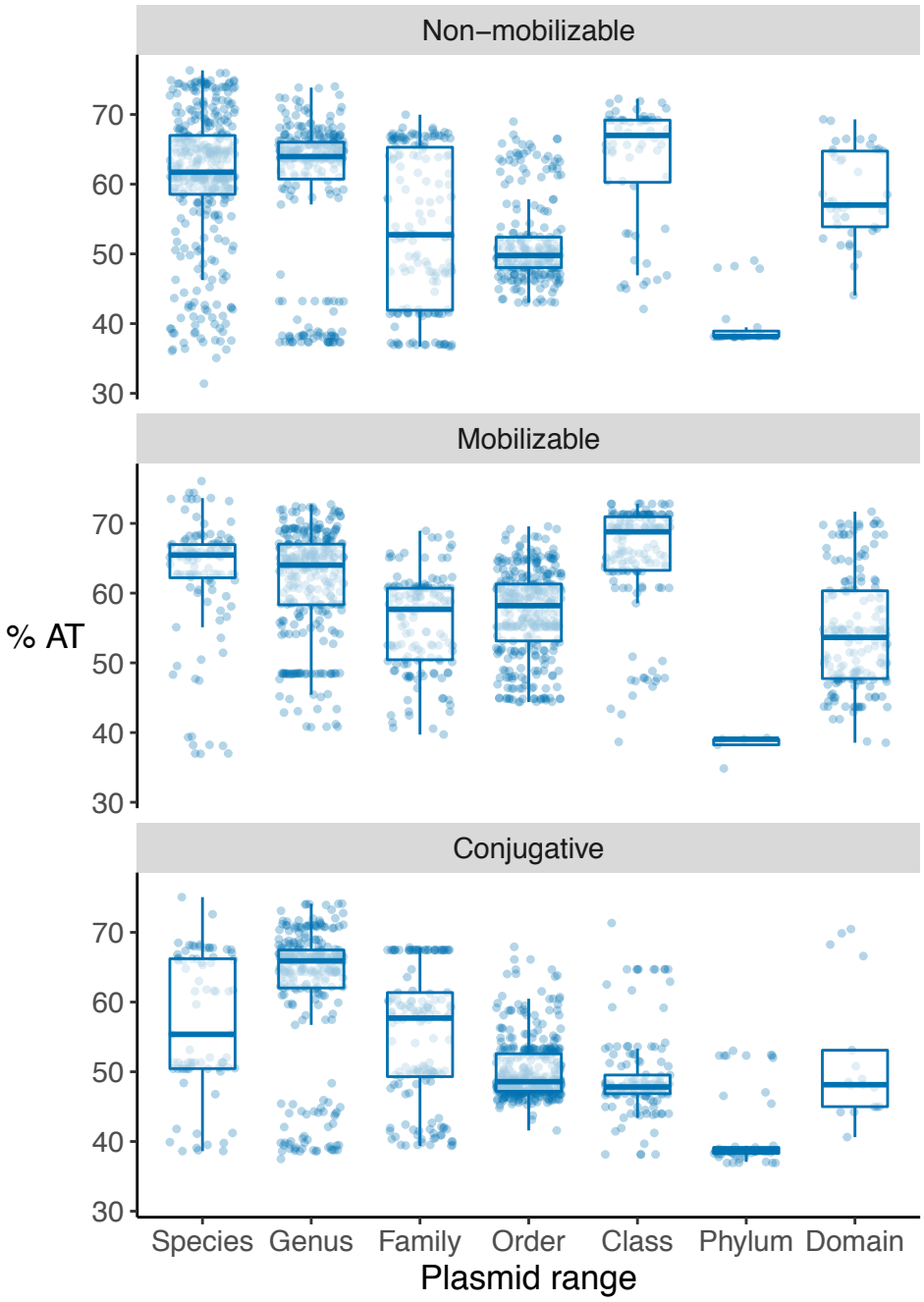


Figure 8. AT-content, plasmid range and plasmid mobility.

Plasmids, each shown as a blue circle, are categorised into one of seven plasmid ranges, depending on the taxonomic rank of the common ancestor of all genomes which contain a copy of that plasmid. Additionally, plasmids are categorised as one of three different mobilities: non-mobilizable, mobilizable and conjugative, with one panel for each of these mobilities. The y-axis is the percentage of bases which are either A or T. Overall, there was mixed, but generally weak, evidence for a correlation in each mobility class.

We also considered how the difference in plasmid AT-content compared to their species' chromosomes, in addition to the AT-content alone, correlated with plasmid range (Figures 9 and 10). In contrast to the previous results, we found a significant, but very weak, positive correlation between the difference in plasmid and chromosome AT-content and plasmid range when we considered plasmids as independent, and no correlation when we controlled for phylogeny and number of plasmids (Figure 9) (ANOVA: slope estimate = 0.26, t=7.38, $p<0.001$, $R^2=0.015$; MCMCglmm: posterior mean=-0.05, 95% CI = -0.13 to 0.02, pMCMC=0.204, $R^2<0.001$).

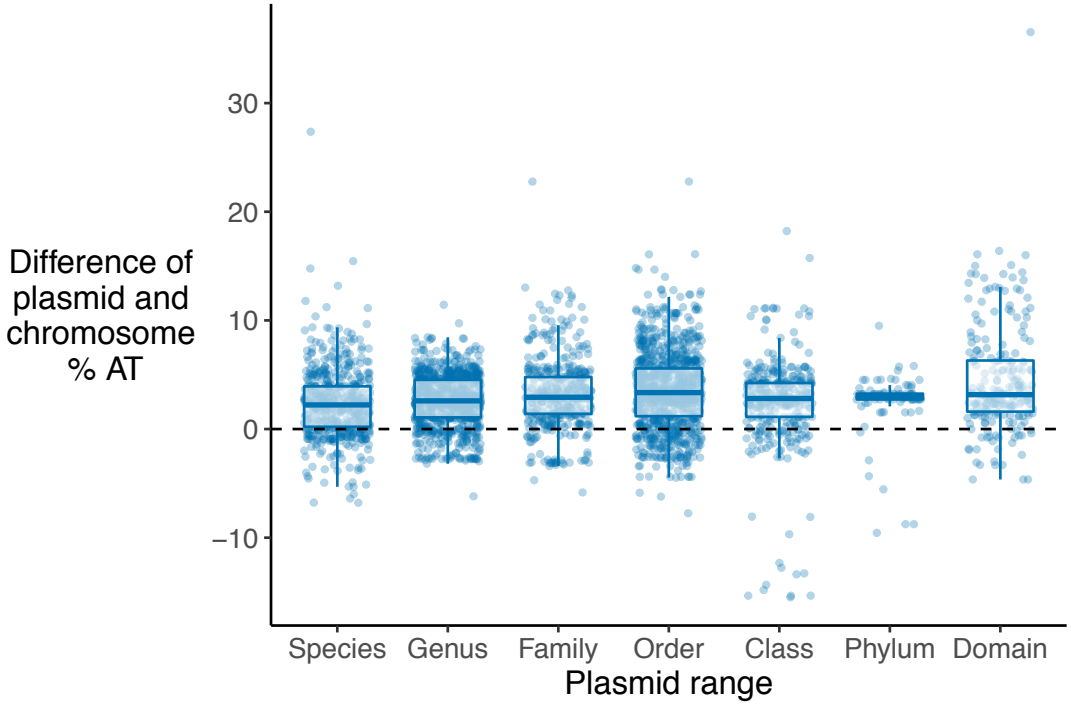


Figure 9. Difference in plasmid and chromosome AT-content and plasmid range.

Plasmids, each shown as a blue circle, are categorised into one of seven plasmid ranges, depending on the taxonomic rank of the common ancestor of all genomes which contain a copy of that plasmid. The y-axis shows the difference in the percentage of bases that are A or T for plasmids compared to their species' representative chromosome. Overall, there seems to be no effect on the AT-content, relative to chromosomes, of plasmid range.

As above, we then considered each of the three plasmid mobility categories separately (Figure 10). For non-mobilizable plasmids, we found a significant, but very weak, positive correlation between plasmid range and the difference between plasmid and chromosome AT-content

(ANOVA: slope estimate=0.54, t=7.694, p<0.001, R^2=0.048; MCMCglmm: posterior mean = 0.15, 95% CI=0.01 to 0.28, pMCMC=0.044, R^2=0.003). This was also the case for mobilizable plasmids when plasmids were considered independent, but this became non-significant when we controlled for phylogeny and plasmid number (ANOVA: slope estimate=0.20, t=3.89, p<0.001, R^2=0.011; MCMCglmm: posterior mean=0.005, 95% CI = -0.11 to 0.11, pMCM=0.93, R^2<0.001). For conjugative plasmids, the R^2 for both analyses was extremely small, and the estimate of the correlation was only significant when controlling for species' phylogeny and plasmid number, suggesting no, or very little, correlation overall (ANOVA: slope estimate=-0.028, t=-0.39, p=0.697, R^2<0.001; MCMCglmm: posterior mean=-0.408, 95% CI=-0.55 to -0.27, pMCMC<0.001, R^2<0.001).

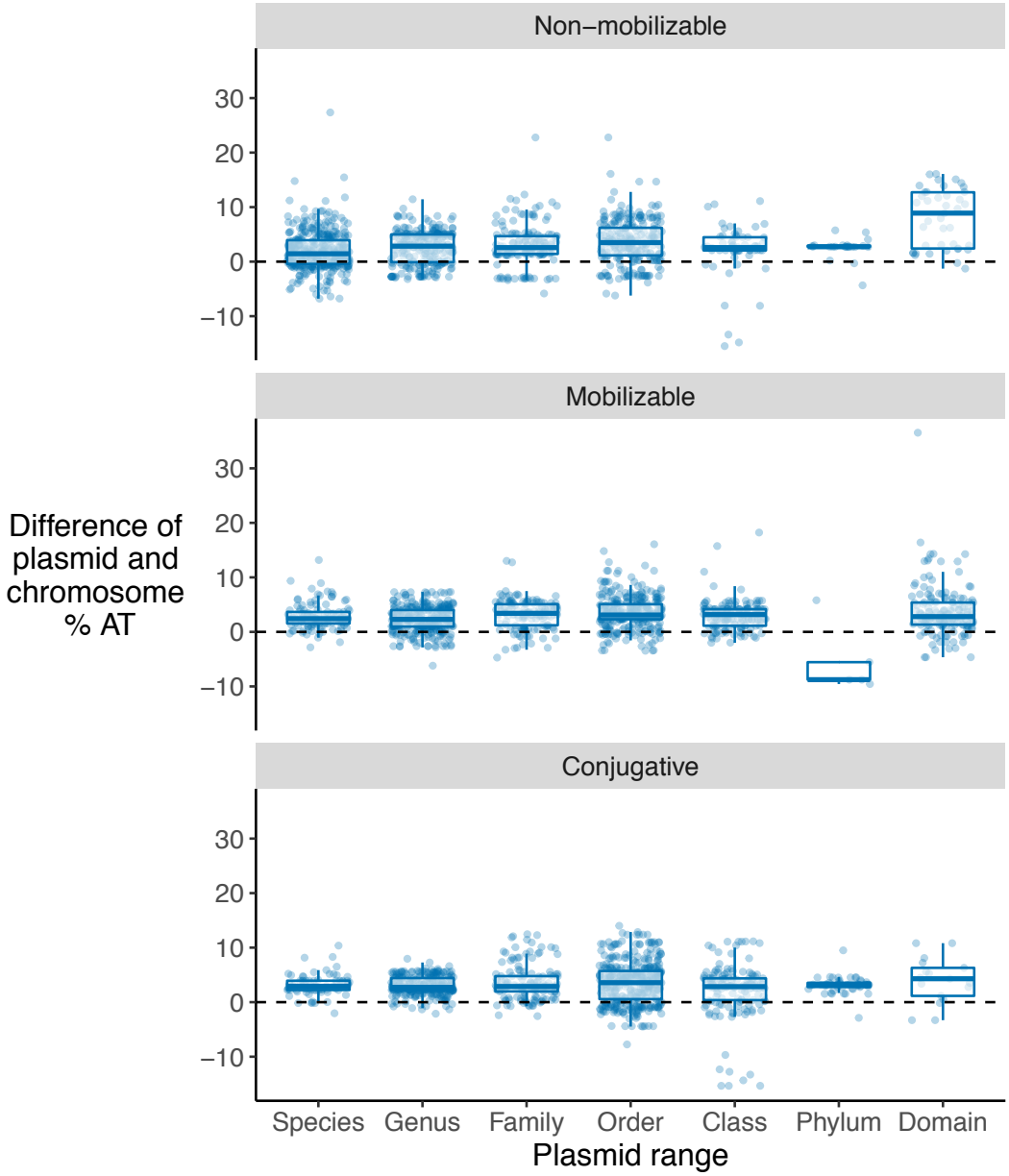


Figure 10. Difference in plasmid and chromosome AT-content, plasmid range and plasmid mobility.

Plasmids, each shown as a blue circle, are categorised into one of seven plasmid ranges, increasing in range along the x-axis, and one of three plasmid mobilities, with one panel for each of these mobilities. The y-axis shows the difference in the percentage of bases that are A or T for plasmids compared to their species' chromosome(s). Overall, there were no meaningful significant correlations.

Discussion

As expected, we found that plasmids have a consistently higher AT-content than their species' chromosome AT-content (Figures 3 & 4). We then tested the predictions of two hypotheses for why this plasmid AT-bias exists by examining how plasmid AT-content correlated with range and mobility. Overall, we found little evidence for a negative correlation between plasmid AT-content and both mobility and range, as predicted by the adaptation hypothesis (Figures 5 & 7). This was true when considering the percentage of A and T bases alone, and also when calculating the difference between plasmid and chromosome AT-content (Figures 6 & 9). While we did initially find negative correlations between AT-content and plasmid range for all three plasmid mobility classes, these correlations largely disappeared and/or had very small effect sizes when controlling for phylogeny, number of plasmids per species and the AT-content of the species' chromosomes (Figure 8 & 10). Taken together, these results do not support the adaptation hypothesis, and are therefore more consistent with the hypothesis that plasmid AT-bias is due to accumulation of mutations in plasmid sequences.

We found that plasmids are consistently enriched with A and T bases compared to chromosomes, and that this was the case for almost all of the species we analysed (Figure 3). We also found that the AT-content of plasmids is highly correlated with the AT-content of their species' chromosome. Specifically, around 75-88% of the variance in plasmid AT-content was explained by the AT-content of their species' chromosome. This is despite the AT-content of chromosomes varying considerably across species, from 33-75%. The strong correlation between plasmid and chromosome AT-content requires explanation, since it suggests that plasmids have more similar AT-contents to their host chromosomes than we may expect. Some authors have suggested this could be evidence of selection on plasmid AT-content (Dietel *et al.* 2018, 2019). While we find no evidence of this in our analyses here, understanding the reasons for this correlation is a key question for future work.

We found weak or no support for the hypothesis that AT-bias is an adaptation of plasmids to reduce their costs to their host. When analysing plasmids as independent, we found that plasmid AT-content was significantly negatively correlated with: (a) plasmid mobility, consistent with the first prediction of the adaptation hypothesis (Figure 5); (b) plasmid range, both for all plasmids together and when we considered each of the three mobility classes separately, consistent with the second prediction of the adaptation hypothesis (Figures 7 & 8). However,

when we controlled for species' host phylogeny, number of plasmids and the AT-content of host species' chromosomes, these correlations were no longer significant and/or negative (Figures 6, 9 & 10). Additionally, the very small effect sizes for all of these results suggests that plasmid mobility and plasmid range have very little influence on the AT-content of plasmids, in contrast to the two predictions of the adaptation hypothesis. Instead, these results are more consistent with the two predictions of the mutation hypothesis.

A caveat here is that as the range of a plasmid becomes broader, it becomes less meaningful to compare the plasmid AT-content to its host chromosome, since this is only one of the many species it has been sequenced in. This makes assessing support for the second prediction of the adaptation hypothesis particularly difficult. Nevertheless, it seems that while there are some significant correlations in the direction expected, these are weak, and overall suggest that plasmid range has a limited effect on the AT-content of plasmids.

Additionally, another caveat is that the population size of plasmids could potentially be positively correlated with their transferability (both mobility and range). Larger population sizes would reduce the power of genetic drift and increase the strength of purifying selection, together leading to fewer mutations accumulating in these plasmid sequences. When we stated that the Mutation hypothesis predicted no correlation between plasmid AT-content and plasmid mobility or range, this assumed AT-biased mutations would accumulate at the same rate in all plasmids, regardless of transferability. However, a positive correlation between plasmid population size and transferability might instead mean that the least transferrable plasmids would accumulate mutations at the fastest rate. Therefore, rather than predict no correlation between plasmid AT-content and mobility and/or range, the mutation hypothesis may instead predict a negative correlation, the same prediction as the adaptation hypothesis. This would limit our ability to distinguish between these hypotheses using correlational analyses such as those presented in this Chapter.

Overall, we found little support for the adaptation hypothesis explaining why plasmids have consistent AT-bias. Instead, our results suggest that the mutation hypothesis may instead be more important for driving plasmid AT-bias. Future work could examine signatures of selection and/or drift within plasmid sequences to further examine these two hypotheses, which would provide additional insights beyond the correlational analyses we have presented here. Additionally, calculating the AT-content of each chromosome, rather than using the species'

representative chromosome as we have done here, would allow better comparisons between plasmids and their hosts.

Plasmids are not the only intracellular element which display AT-bias. Bacterial endosymbionts have also been consistently observed to be AT-rich, both compared to their free-living ancestors and to their eukaryotic hosts (Rocha & Danchin 2002; Dietel *et al.* 2018). The suggested reasons for this bias are largely analogous to those we tested for plasmids: either living inside a host cell increases the rate at which AT-biased mutations are fixed, or endosymbionts evolve AT-rich genomes to reduce their cost to the host cell (Moran 1996; Wernegreen & Moran 1999; Rocha & Danchin 2002; Wernegreen 2002). Therefore, similar analyses but on endosymbiont genomes could help further assess which of these hypotheses matters more for intracellular elements in general. It could also help to consider whether both plasmids and endosymbionts contain signatures within their genome that are more consistent with adaptation or genetic drift.

References

Bohlin, J., Eldholm, V., Pettersson, J.H.O., Brynildsrud, O. & Snipen, L. (2017). The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics*, 18, 151.

Cornelis, G.R., Boland, A., Boyd, A.P., Geuijen, C., Iriarte, M., Neyt, C., *et al.* (1998). The Virulence Plasmid of Yersinia, an Antihost Genome. *Microbiol Mol Biol Rev*, 62, 1315–1352.

Dietel, A.-K., Kaltenpoth, M. & Kost, C. (2018). Convergent Evolution in Intracellular Elements: Plasmids as Model Endosymbionts. *Trends in Microbiology*, 26, 755–768.

Dietel, A.-K., Merker, H., Kaltenpoth, M. & Kost, C. (2019). Selective advantages favour high genomic AT-contents in intracellular elements. *PLOS Genetics*, 15, e1007778.

Ding, H. & Hynes, M.F. (2009). Plasmid transfer systems in the rhizobia. *Can J Microbiol*, 55, 917–927.

Hale, T.L. (1991). Genetic basis of virulence in Shigella species. *Microbiol. Rev.*, 55, 206–224.

Hershberg, R. & Petrov, D.A. (2010). Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLOS Genetics*, 6, e1001115.

Hildebrand, F., Meyer, A. & Eyre-Walker, A. (2010). Evidence of Selection upon Genomic GC-Content in Bacteria. *PLoS Genet*, 6, e1001107.

Moran, N.A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *PNAS*, 93, 2873–2878.

Nishida, H. (2012). Comparative Analyses of Base Compositions, DNA Sizes, and Dinucleotide Frequency Profiles in Archaeal and Bacterial Chromosomes and Plasmids. *International Journal of Evolutionary Biology*, 2012, e342482.

Robertson, J. & Nash, J.H.E. (2018). MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom*, 4.

Rocha, E.P.C. & Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. *Trends in Genetics*, 18, 291–294.

Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R.C. & San Millán, Á. (2021). Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology*, 1–13.

Rodríguez-Beltrán, J., Sørum, V., Toll-Riera, M., Vega, C. de la, Peña-Miller, R. & Millán, Á.S. (2020). Genetic dominance governs the evolution and spread of mobile genetic elements in bacteria. *PNAS*, 117, 15755–15762.

Wernegreen, J.J. (2002). Genome evolution in bacterial endosymbionts of insects. *Nature Reviews Genetics*, 3, 850–861.

Wernegreen, J.J. & Moran, N.A. (1999). Evidence for genetic drift in endosymbionts (Buchnera): analyses of protein-coding genes. *Molecular Biology and Evolution*, 16, 83–97.

Chapter 5: Environmental variability and the structure of bacterial pangenomes

Abstract

Pangenomes, defined as all the genes that have been sequenced in a species, vary substantially in structure across bacteria. This variation is usually captured by comparing the percentage of genes which are core, those found in all or the majority of genomes, with the percentage that are accessory, those found in only a subset of genomes. From a mechanistic perspective, pangenome structure is likely to be driven by differences in gene gain and loss across genomes of the same species. In contrast, *why* this seems to vary so much between species is less clear. It has been observed that species which encounter more variable environments tend to have more variable pangenomes. However, explicit evidence for this correlation is limited. Here, we repeat, simplify and extend previous studies to assess the current evidence for a correlation between pangenome and environmental variability. We examine whether the pangenome structure of 126 bacterial species correlates with two measures of environmental variability, and find mixed support. We use these results to identify limitations with the current approach, especially with the measures of environmental variability. We then suggest future analyses that incorporate measures which better reflect bacterial lifestyles. These future analyses will help to better explore what aspects of bacterial lifestyle affect pangenome structure, and potentially address whether this pangenome variation is due to neutral or adaptive processes.

Introduction

The term 'pangenome' refers to all of the genes carried by individuals of a certain group. While usually used in the context of prokaryotic species' genomes, a pangenome could refer to all the genes found in any group, including animal and plant species, and even humans (McInerney *et al.* 2017b).

However, the structure of eukaryote and prokaryote pangenomes appears to be very different. In most eukaryotic species, the genomes of individuals usually differ predominantly at the allelic level. For example, all humans usually carry the LCT gene to produce lactase, the enzyme that allows the digestion of milk (Ingram *et al.* 2009). Whether this enzyme continues to be produced into adulthood depends on which allele of the gene a person carries. In contrast, individuals of the same bacterial species tend to vary much more at the genetic level. For example, one individual may carry a suite of genes that allow them to break down certain food

sources, invade certain hosts and survive in certain toxic environments (Goyal 2018). Another individual of the same species may carry very few of these genes, but instead have its own suite of genes which provide similar functions but for different food sources, hosts and toxic environments.

As such, bacterial pangenomes usually consist of genes which are found in all members of a species, called 'core' genes, and genes found in only a subset of individuals, called 'accessory' genes. As more and more bacterial genomes have been sequenced, it has become apparent that the relative size and proportion of the core and accessory genome can vary substantially between different bacterial species. What causes so much variation between species is not fully understood.

From a mechanistic perspective, the core and accessory structure of bacterial pangenomes is produced by differential gain and loss of genes (Puigbò *et al.* 2014; Domingo-Sananes & McInerney 2021). The gain of new genes occurs frequently in bacteria, particularly through horizontal gene transfer (HGT). This is a process whereby individual bacteria can acquire genes from other individuals within the same generation. HGT can occur between different species, allowing individuals to acquire genes from a potentially unlimited pool. With all this gene gain, individual bacterial genomes could potentially become huge. Therefore, there must also be frequent loss of genes to maintain genomes of the same species at similar sizes (McInerney *et al.* 2017b; Domingo-Sananes & McInerney 2021). Together, differences in the gain and loss of genes could explain how pangenome structure varies between species.

It has been observed that in general, species with more variable environments and lifestyles tend to have more variable pangenomes (McInerney *et al.* 2017b, a; Maistrenko *et al.* 2020). This has been suggested to be due to the relative diversity of genes gained, and different environments selecting for different genes to be lost. First, the variation of genes that individuals from one bacterial species acquire is likely to depend on the diversity of other bacterial species that they regularly encounter. This diversity of encountered species probably correlates with the number and variety of environments individuals of a species regularly live in. A model by Niehus et al. (2015) showed that HGT could maintain and even increase genetic diversity, but only when migration of new genotypes into the population was permitted (Niehus *et al.* 2015). Therefore, more variable environments would expand the number and variety of

genes that bacteria could potentially acquire, which would then be reflected in a species' pangenome, as part of a largely neutral process (Andreani *et al.* 2017).

Second, if individuals only encounter one environment out of a large number that are possible for that species, the fitness of such individuals would only be under selection for that particular environment (Polz *et al.* 2013; McInerney *et al.* 2017b; Goyal 2018). Consequently, any genes they carry that are useful in other environments will no longer be useful, and will likely be lost. The more environments a species can live in, the more variable selection on gene loss would be across individuals of the same species. These 'Gene-by-environment' interactions could be important in shaping the structure of pangenomes (Domingo-Sananes & McInerney 2021). Therefore, more variable environments could increase the diversity of potential selection on gene loss, leading to a higher proportion of genes that are not present in all genomes of the species.

However, the evidence for a correlation between bacterial lifestyle and pangenome variability is largely observational, with authors noting patterns or comparing certain species (McInerney *et al.* 2017b; Maistrenko *et al.* 2020; Domingo-Sananes & McInerney 2021). Recently, Maistrenko *et al*. (2020) studied how the environmental variability of 155 bacterial species correlated with different measures of their pangenome structure. Of nine measures examined, they found only the size of the core genome was significantly positively correlated with the number of environments a species lived in (Maistrenko *et al.* 2020). This was despite environmental preferences explaining up to 49% of the variance in species' pangenome measures. Their results suggest that while lifestyle and environment variability are likely to be important for determining pangenome structure, the relative importance of the different aspect(s) of a species' lifestyle and/or environment remains unknown.

To understand any potential effect of lifestyle and environment in determining the structure of species' pangenomes, a comparative analysis across bacteria is required that considers multiple aspects of species' lifestyle and environment. In particular, these aspects should have a clear hypothesised influence on the relative balance of gene gain and loss, which would in turn determine pangenome structure. While Maistrenko et al. (2020) considered many measures of variation across pangenomes, their measure of environmental variability, defined as the number of 63 habitats a species' 16S rRNA had been sequenced in, was comparatively simple.

There is the potential for future analyses to consider measures that are more informed by our understanding of bacterial lifestyles. We shall return to this point in the discussion.

Here, we provide an initial step by repeating, simplifying and extending previous analyses to review the current evidence for a potential positive correlation between species' environmental and pangenome variability. Specifically, we analyse whether different measures of pangenome structure correlate with two measures of environmental variability across 126 bacterial species. We use these results to discuss potential caveats and limitations of these kinds of analyses. Consequently, we then suggest potential ways to expand such analyses, with a particular focus on next steps to fully understand what aspects of bacterial environments and/or lifestyles affect species' pangenome structure and why.

Methods

Collection of pangenome data

We collected bacterial pangenome data from panX (https://pangenome.org/) (Ding *et al.* 2018). PanX is a web-based pangenome database that uses a pipeline to break annotated genomes into genes and then cluster them into orthologous groups. To allow identification of orthologous gene clusters and establishment of pan-genomes, panX only includes species that have a minimum of 10 complete genomes in the RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/). We retrieved data on 126 bacterial pan-genomes composed of 6234 genomes. PanX stores data in JSON format and we downloaded this using GNU Wget (https://www.gnu.org/software/wget/).

Analysis of pangenome data

We performed all our analyses of the pangenomes in R (version 4.0.2). We used the R package 'jsonlite' to convert the JSON data files into R objects. The pangenome data we used included information on orthologous genes and the strains in which they were found. We thus categorized genes into core or accessory genes based on this information. There is some inconsistency in how to define core genes in the literature. Some authors define core genes as those present in all genomes, while others choose a high threshold that allows for an occasional genome not to carry that gene (Medini *et al.* 2005; Maistrenko *et al.* 2020; Hall *et al.* 2021; Whelan *et al.* 2021). We decided to analyse our data with a mix of both approaches, and have analysed core genes using three different thresholds. Specifically, we have defined core genes

as: (1) genes present in 100% of genomes; (2) genes present in ≥ 90% of genomes; (3) genes that present in ≥ 80% of genomes.

We were also interested in genes that only exist in small subset of genomes. These are the genes that vary most between genomes, and are therefore potentially most correlated with the environment. Genes found in only a subset of genomes are defined as 'accessory' genes, and some authors refer to accessory genes found in only a small subset of genomes as 'cloud' genes. We will refer to them here as accessory genes for simplicity. We defined these accessory genes with two thresholds: (1) genes present in ≤ 10% of genomes; (2) genes present in ≤ 20% of genomes.

We then calculated the percentage of core and accessory genes for each species as followed:

$$Percentage\ of\ core\ genes\ = \frac{number\ of\ core\ genes}{number\ of\ genes} \times 100\% .$$

$$Percentage\ of\ accessory\ genes\ = \frac{number\ of\ accessory\ genes}{number\ of\ genes} \times 100\% .$$

We did this for all thresholds of core and accessory genes stated above.

In addition, we were also interested in understanding the average percentage of core genes at the individual genome level, rather than the entire pangenome level. Consequently, we calculated the percentage of core genes at the genome level for each species as followed:

$$Percentage\ of\ core\ genes\ = \frac{\sum_i^n \frac{number\ of\ core\ genes\ in\ genome\ i}{number\ of\ genes\ in\ genome\ i}}{number\ of\ genome(n)} \times 100\% .$$

We did this for the three thresholds of core genes, as stated above.

Some genomes did not contain any genes found in <10 or <20% and so this would bias our calculations of the average percentage of accessory genes at the genome level. Therefore, we only looked at the percentage of accessory genes at the pangenome level.

Measures of environmental variability

We used data from two different studies which both estimated the number of environments that different bacterial species lived in. The first measure compared species' 16s rRNA sequences to a diverse range of metagenome datasets, and recorded the number of 63 types of environment

a species was found in (Maistrenko *et al.* 2020). Examples of these 63 environments include: ice, lake, wetland, rock, skin, stomach, blood, ray finned fish, mammal, rhizosphere, forest and mangrove. The second measure also used 16s rRNA datasets from a diverse range of environments, and recorded the number of five broader environments in which a match was found for each species: water, wastewater, sediment, soil and host (Garcia-Garcera & Rocha 2020).

Statistics

We carried out all statistical analysis and graph plotting in R (version 4.0.2). We used the R package MCMCglmm to disentangle the associations between bacterial pangenome features and environmental variability (Hadfield 2010). The evolutionary history of bacteria could mean closely related species have more similar pangenome structure, regardless of their environmental variability. Consequently, we controlled for the phylogenetic relationships between species in our dataset by setting the phylogeny as a random effect in our model. We generated a phylogeny of the 126 species in our dataset using the method described in Chapter 2. We have reported the posterior mean, 95% Credible Intervals (functionally similar to 95% Confidence Intervals), the pMCMC value (used here as 'p-value'), and the R^2 of the fixed effect for each model in Table 1.

Results

Substantial variation in pangenome structure across species

We found extreme variation between species in the structure of their pangenomes (Figure 1). Across all measurements, the species with the most variable pangenome was *Escherichia coli*, while the species with the least variable pangenome was *Rickettsia japonica*. For example, only 1.6% of the genes in the *Escherichia coli* pangenome were core genes found in all genomes, and 78.2% were accessory genes found in less than 10% of genomes. In contrast, 91.4% of genes in the *Rickettsia japonica* pangenome were found in all genomes, and only 0.2% were found in less than 10% of genomes.

We also considered how variable individual genomes of the same species were to one another. Specifically, we counted the number of core genes in each genome, and divided this by the total number of genes in that genome. We then calculated the mean value across all genomes in a species (Figure 2). We did this for all three thresholds of core (100%, 90% and 80%). This

measure was also extremely variable across species. On average, only 9.8% of genes in an average *Escherichia coli* genome were found in 100% of genomes, and 44.1% of genes were found in at least 80% of genomes. In contrast, 94.2% of an average *Rickettsia japonica* genome consists of core genes found in 100% of genomes, with 98.8% of genes found in at least 80% of genomes.

Overall, we found that both pangenomes and individual genomes are extremely variable across species. For some species, a very small proportion of genes were found in all genomes, while in others virtually all genes were found in the majority of genomes. We next looked at whether the environmental variability of species could explain this variation.

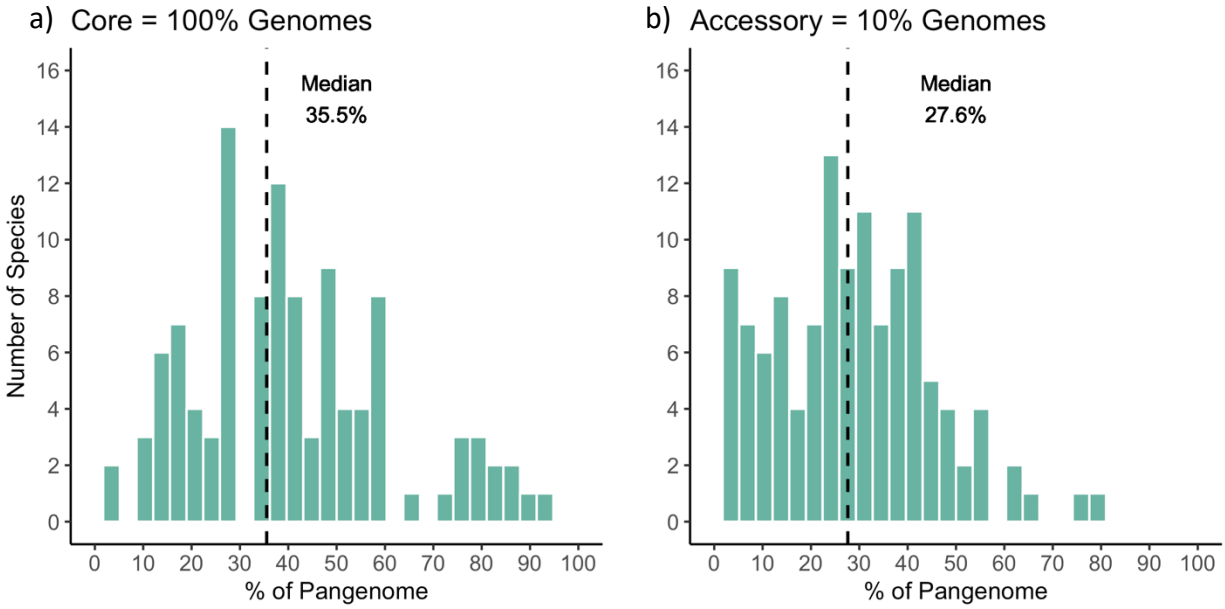


Figure 1. Percentage of core and accessory genes across species' pangenomes.

Histograms showing the distribution of the percentage of genes that are (a) core and (b) accessory, across all 126 species' pangenomes. The horizontal dashed line indicates the median percentage of core and accessory genes, in (a) and (b) respectively. The broad width of the histogram bars indicates that species vary considerably in their pangenome structure.

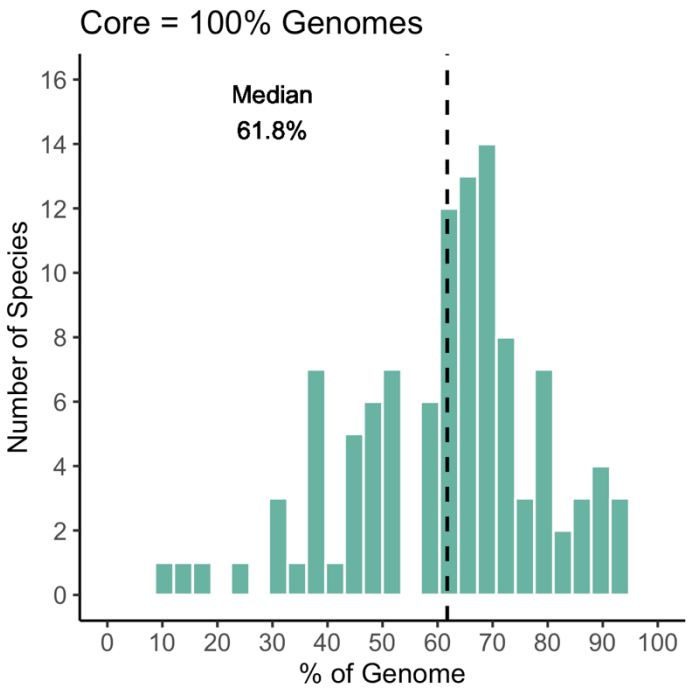


Figure 2. Percentage of core genes in an average genome across species. Histogram showing the distribution of the percentage of genes which are core in an average genome, across 126 species. Here, core is defined as 100% of genomes. The dashed line is the median species for the percentage of genes in an average genome that are found in all genomes.

No correlation between environmental variability and the percentage of core or accessory genes in species' pangenomes

The lifestyle of a species could explain why species differ so much in the structure and variability of their genomes. To explore this, we examined whether the environmental variability of species could explain some of this variation. In our analyses, we used two measures of environmental variability. Both were based on the number of environments that a species' 16S rRNA had been found in.

First, we measured environmental variability using the number of five broad environments each species was found in. We then compared this to the percentage of the pangenome that was core genes, defined as genes present in 100%, 90% or 80% of a species' genomes, and the percentage of the pangenome that was accessory genes, defined as those present in 10% or 20% of a species' genomes.

We found no significant correlation between the number of environments and the percentage of both core and accessory genes in a species' pangenome (Table 1, rows 1-5; Figures 3 & 4). This lack of correlation was true for all three thresholds of how we defined core genes, and both thresholds of how we defined accessory genes (Tables 1, rows 1-5; Figures 3 & 4).

	Threshold	Posterior mean	Lower 95% CI	Upper 95% CI	pMCMC	Sigificance	R^2 for fixed effect
% Core vs Number of Environments (Pangenome)							
1	100%	-3.341	-6.826	0.458	0.082	None	0.019
2	90%	-2.998	-7.338	1.038	0.166	None	0.013
3	80%	-2.871	-6.801	1.210	0.19	None	0.012
% Accessory vs Number of Environments (Pangenome)							
4	10%	2.301	-1.551	5.905	0.238	None	0.010
5	20%	2.903	-1.111	6.916	0.128	None	0.013
% Core vs Number of Environments (Average Genome)							
6	100%	-5.199	-8.463	-1.295	<0.001	***	0.071
7	90%	-4.248	-6.683	-1.837	<0.001	***	0.081
8	80%	-2.870	-4.913	-0.589	0.01	**	0.058
% Core vs Ubiquity (Pangenome)							
9	100%	0.387	-0.135	0.929	0.156	None	0.014
10	90%	0.356	-0.192	0.936	0.212	None	0.010
11	80%	0.339	-0.173	0.948	0.236	None	0.009
% Accessory vs Ubiquity (Pangenome)							
12	10%	-0.229	-0.740	0.250	0.386	None	0.007
13	20%	-0.242	-0.772	0.267	0.356	None	0.006
% Core vs Ubiquity (Average Genome)							
14	100%	0.326	-0.166	0.752	0.15	None	0.018
15	90%	0.118	-0.203	0.447	0.466	None	0.004
16	80%	0.103	-0.161	0.394	0.444	None	0.004

Table 1. MCMCglmm results.

Each row is the result of one of the MCMCglmm models analysed in this Chapter. Analyses are grouped by the measures of pangenome and environmental variability they were analysing for a correlation. Analyses that are identical except for the threshold that was used to define core and/or accessory genes are therefore grouped. The threshold column indicates is the specific percentage of genomes used in that model to define core or accessory genes. All models analyse the pangenomes of 126 species.

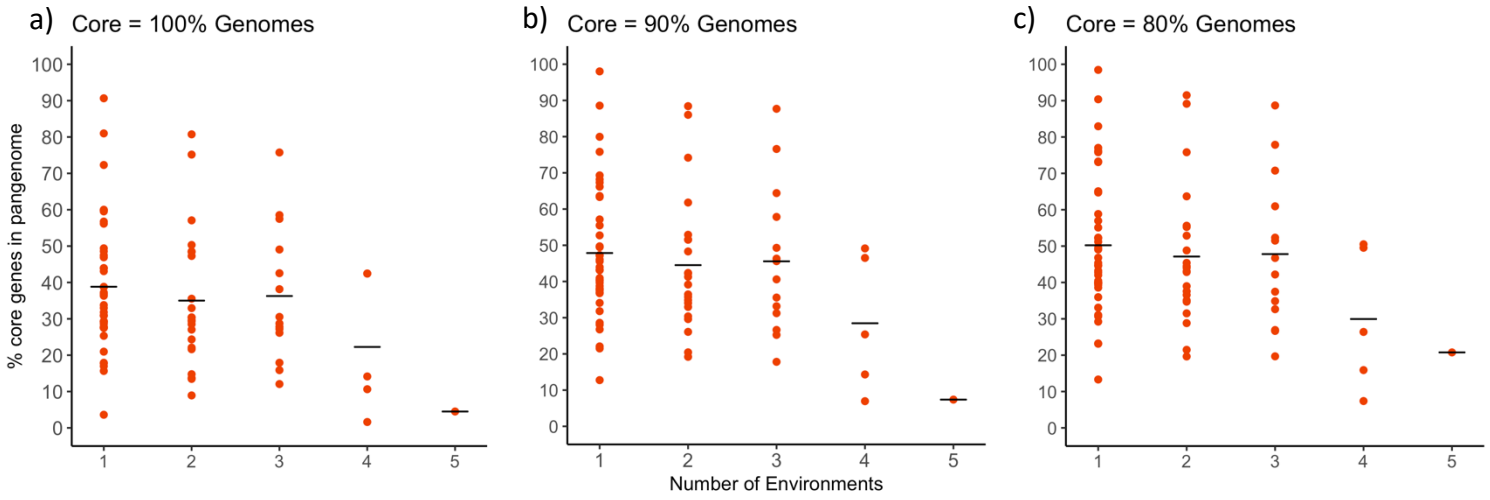


Figure 3. No correlation between the number of environments and the percentage of core genes in species' pangenomes.

The x-axes indicate the number of five broad environments a species' 16S rRNA was sequenced in, and the y-axes indicate the percentage of genes in a species' pangenome which are core genes, defined at three thresholds: (a) 100% of genomes; (b) 90% of genomes; (c) 80% of genomes. Each dot is a species, and the horizontal bars are the mean percentage for all species of those found in the same number of environments.

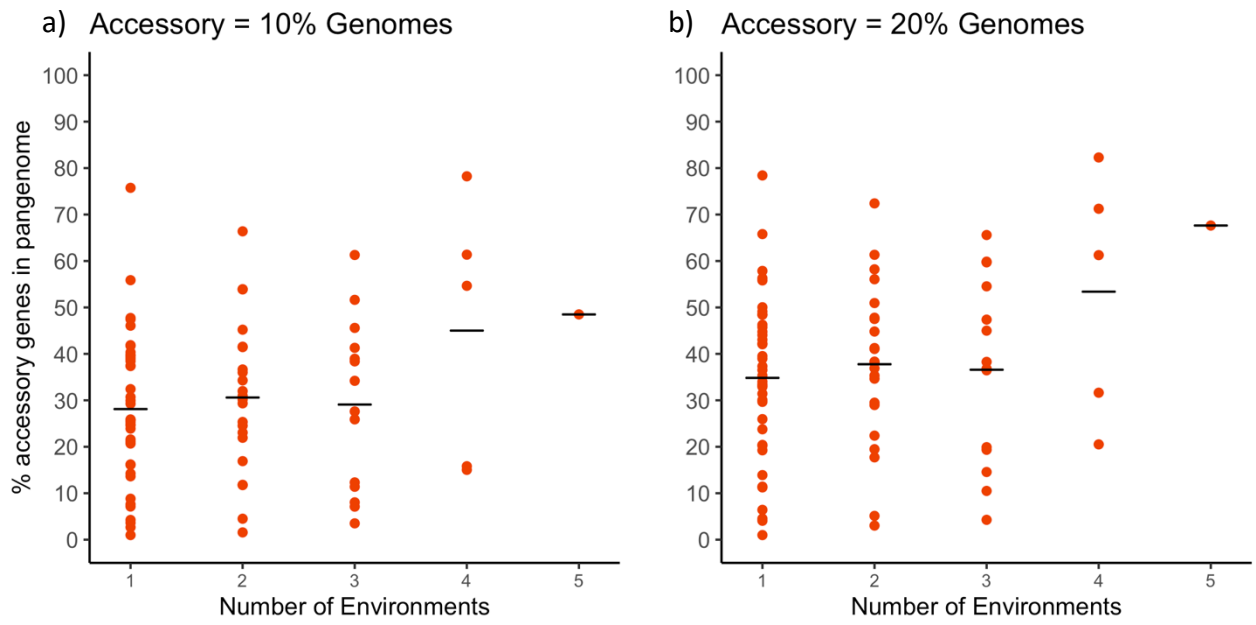


Figure 4. No correlation between the number of environments and the percentage of accessory genes in species' pangenomes.

The x-axes indicate the number of five broad environments a species' 16S rRNA was sequenced in, and the y-axes indicate the percentage of genes in a species' pangenome which are accessory genes, defined at two thresholds: (a) 10% of genomes; (b) 20% of genomes. Each dot is a species, and the horizontal bars are the mean percentage for all species of those found in the same number of environments.

Negative correlation between the number of environments and the percentage of core genes in a species' average genomes

However, we did find a negative correlation between the number of environments and our measure of variability at the genome level. Specifically, this measure considered the percentage of a species' average genome that was core genes. We found a significant negative correlation between this measure and the number of environments for all three thresholds of how we defined core genes (Table 1, rows 6-8; Figures 5).

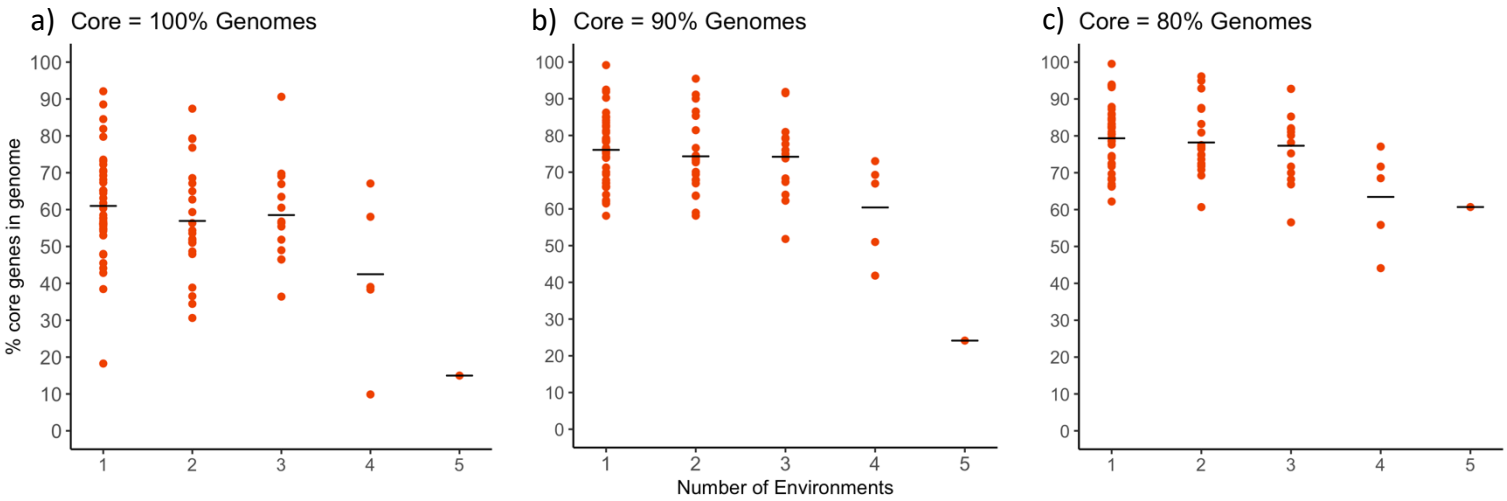


Figure 5. Negative correlation between the number of environments and the percentage of core genes in a species' average genome.
The x-axes indicate the number of five broad environments a species' 16S rRNA was sequenced in, and the y-axes indicate the percentage of genes in a species' average genome which are core genes, defined at three thresholds: (a) 100% of genomes; (b) 90% of genomes; (c) 80% of genomes. Each dot is a species, and the horizontal bars are the mean percentage for all species of those found in the same number of environments.

No correlation between environment ubiquity and the percentage of core and accessory genes in species' pangenomes

Second, we measured environmental variability using the number of 63 more specific environments a species was sequenced in. The authors of the original study which used this measure referred to this as 'environment ubiquity' (Maistrenko *et al.* 2020). We will also use this terminology to make clear that while this is also measure of the number of environments, it is a different measure to the results in Figures 3, 4 and 5.

Consistent with the previous measure of environmental variability, we found no correlation between the environment ubiquity of species and the percentage of core or accessory genes in species' pangenomes (Table 1, rows 9-13; Figures 6 and 7). This was the case for all thresholds of core and accessory genes we considered.

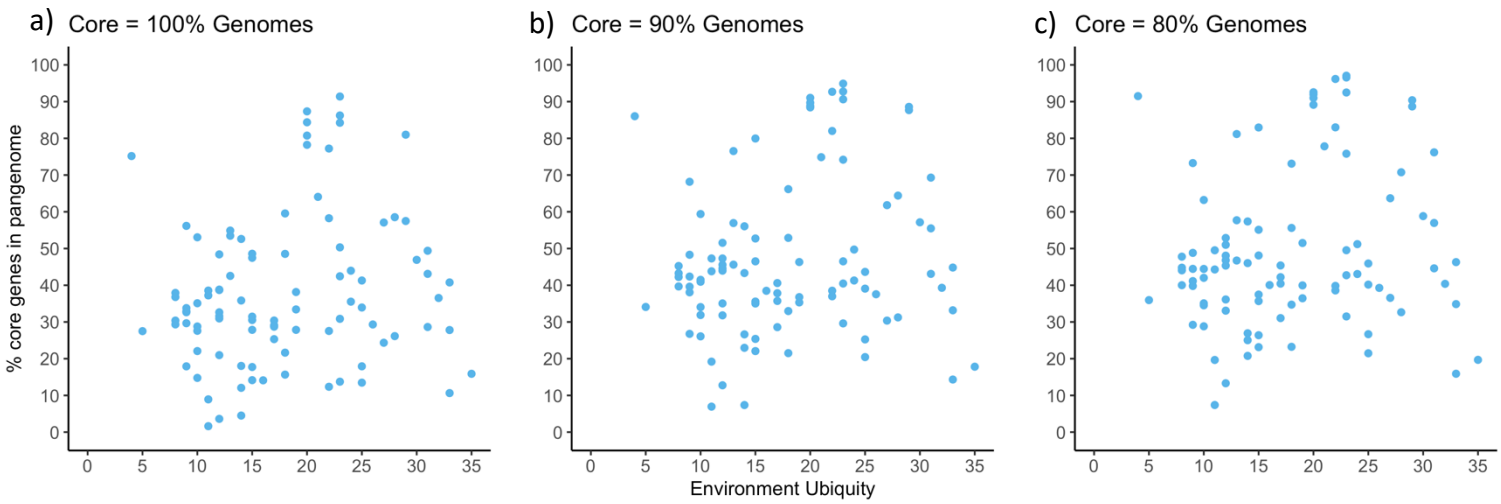


Figure 6. No correlation between environment ubiquity and the percentage of core genes in species' pangenomes.

The x-axes indicate the number of 63 environments a species' 16S rRNA was sequenced in, and the y-axes indicate the percentage of genes in a species' pangenome which are core genes, defined at three thresholds: (a) 100% of genomes; (b) 90% of genomes; (c) 80% of genomes. Each dot is a species.

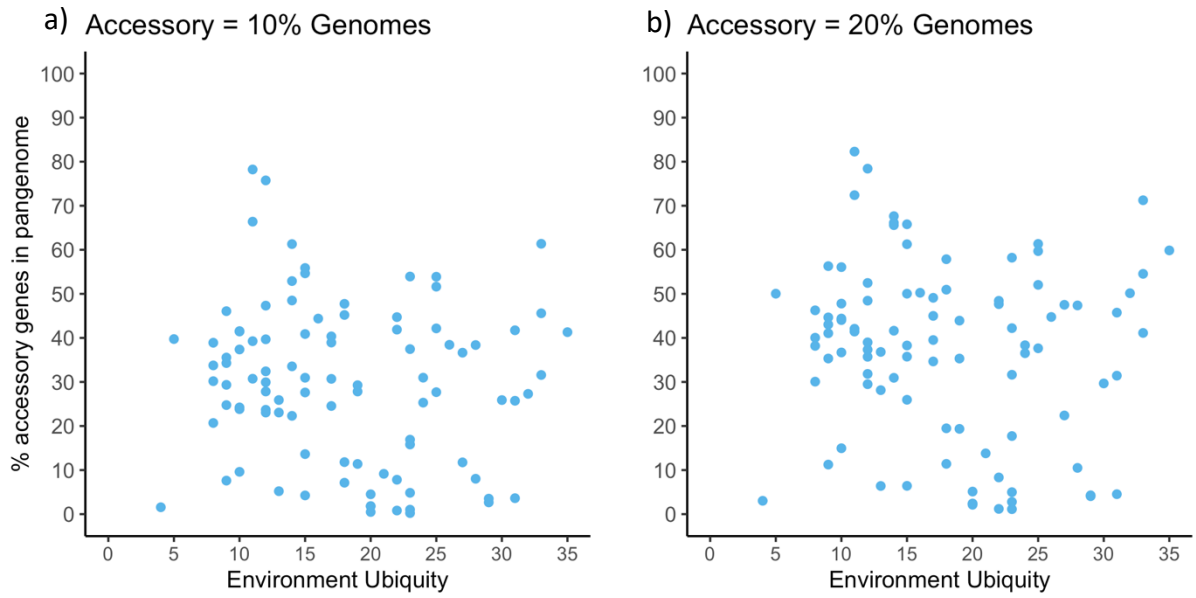


Figure 7. No correlation between environment ubiquity and the percentage of accessory genes in species' pangenomes.

The x-axes indicate the number of 63 environments a species' 16S rRNA was sequenced in, and the y-axes indicate the percentage of genes in a species' pangenome which are accessory genes, defined at two thresholds: (a) 10% of genomes; (b) 20% of genomes. Each dot is a species.

No correlation between environment ubiquity and the percentage of core genes in a species' average genome

We also found no correlation between the environment ubiquity and the percentage of core genes in a species' average genome (Table 1, rows 14-16; Figure 8). This was true for all three thresholds we used to define core genes. These results are in contrast to the significant negative correlations we found between the number of environments and the same measure of variability between species' genomes.

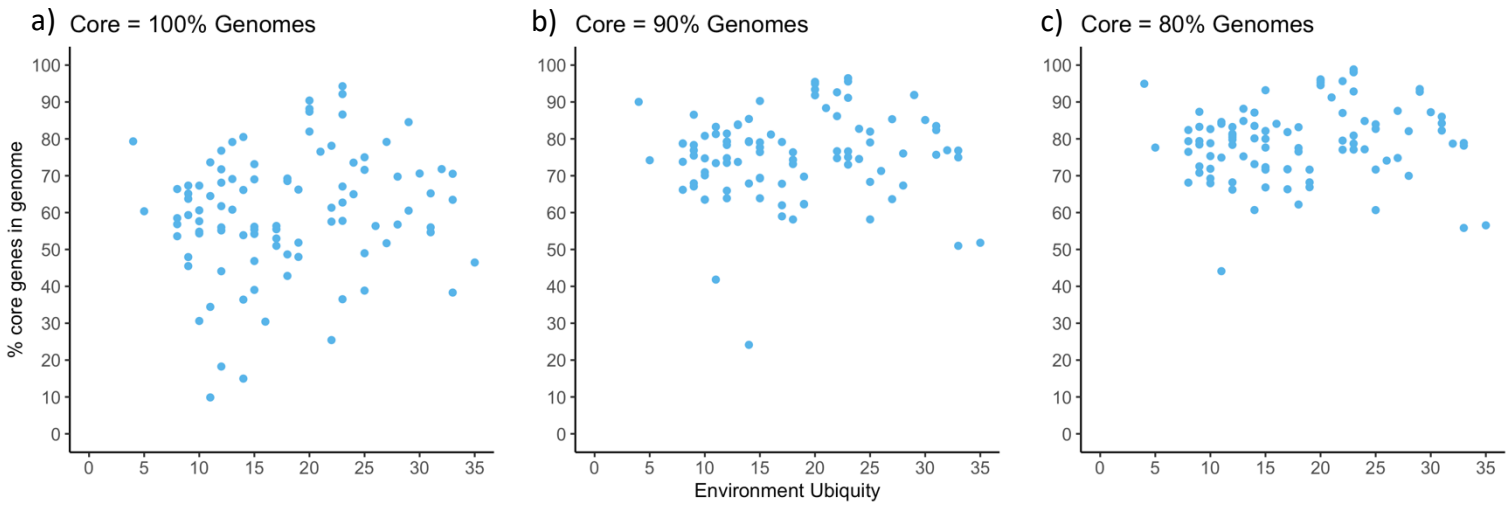


Figure 8. No correlation between environment ubiquity and the percentage of core genes in a species' average genome.

The x-axes indicate the number of 63 environments a species' 16S rRNA was sequenced in, and the y-axes indicate the percentage of genes in a species' average genome which are core genes, defined at three thresholds: (a) 100% of genomes; (b) 90% of genomes; (c) 80% of genomes. Each dot is a species.

Positive correlation between environment ubiquity and the size of the core genome.

Finally, we considered the size of the core genome, defined as the number of genes which are found in all genomes. This is the only measure of pangenome structure that Maistrenko et al. (2020) found was significantly correlated with their measure of environment ubiquity. In agreement with their results, we also found a significant positive correlation between core genome size and the environment ubiquity of species (Figure 9) (MCMCglmm; posterior mean = 63.7, 95% CI = 39.6 to 88.0, pMCMC < 0.001, R^2 of environment ubiquity = 0.19). The environment ubiquity of a species explained almost 20% of the variation in core genome size, once phylogeny was controlled for.

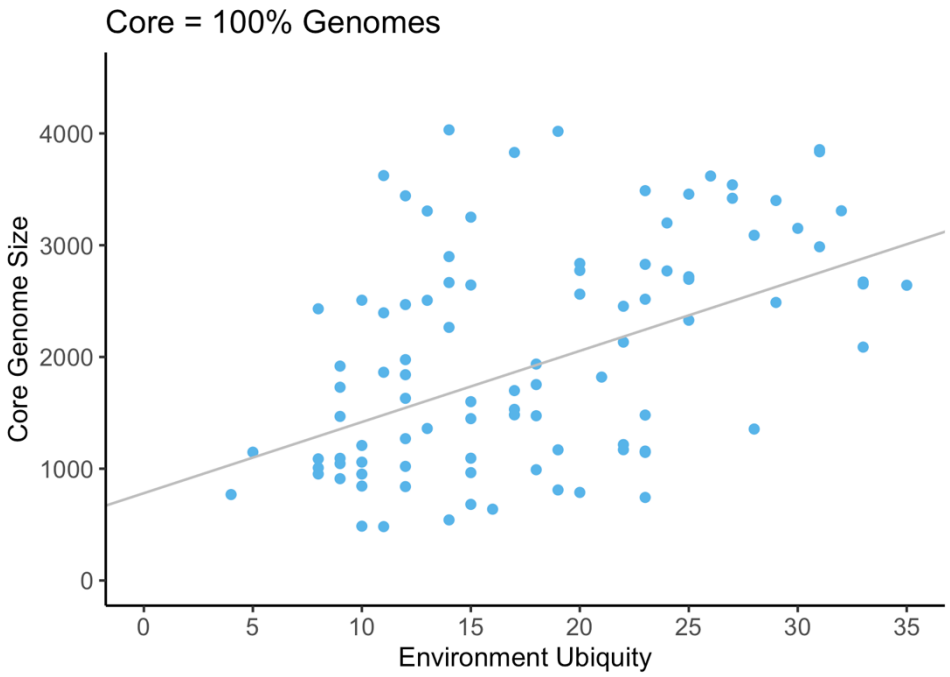


Figure 9. Positive correlation between environment ubiquity and core genome size.

The x-axis indicates the number of 63 environments a species' 16S rRNA was sequenced in, and the y-axis is the number of genes in a species' pangenome found in 100% of genomes. Each dot is a species, and the grey line is the slope estimated by a MCMCglmm analysis controlling for phylogenetic relationships between species.

Discussion

We found mixed support for a correlation between species' environmental and pangenome variability. We found that the number of five broad environments a species' 16s rRNA was sequenced in was negatively correlated with the percentage of genes in an average genome that were core genes (Figure 5). We also found, in agreement with Maistrenko *et al*. (2020), that the number of 63 environments a species' 16s rRNA was sequenced in was positively correlated with the size of their core genome (Figure 9).

However, we found no significant correlations between both measures of environmental variability and the percentage of core or accessory genes in species' pangenomes (Figures 2-4 & 6-8). Considering the percentage of core and accessory genes at the pangenome level, rather than the genome level, could be potentially misleading, because selection acts on individual

genomes, not pangenomes. Therefore, this could explain why we found that only the percentage of core genes at the average genome level was significantly correlated with the first measure of environmental variability.

Furthermore, the positive correlation between core genome size and the second measure of environmental variability also suggests that considering genetic variation at the genome level could be more informative, since again this is an entity which selection will be able to act upon.

Estimating species' environmental variability

The number of environments that a species' 16S rRNA has been sequenced in could potentially be a poor proxy of environmental variability. First, even if the 16s rRNA of a species is sequenced in an environment, this does not necessarily mean that this is part of that species' natural habitat. For example, while gut bacteria such as *Escherchia coli* are often isolated from rivers, this is usually because of sewage contamination, rather than these bacteria naturally living in aquatic habitats long-term.

Second, the number of environments does not consider the potential for variation within these environments. If what matters for pangenome structure is the availability of new genes, some environments may be better able to provide this than others. For example, the gut is likely to have a high turnover of bacteria, due to the constant influx of new bacteria via food intake. Additionally, while soil is considered as one environment in both measures we considered here, soil is a highly variable environment with multiple niches, potentially providing many opportunities to gain and lose genes. In contrast to these, blood, which is usually sterile, is likely to only have one or a few species present. Therefore, while the human gut, soil and blood are three of the potential habitats in the second measure of environmental variability we considered, blood is likely to provide far fewer opportunities to acquire new genes than the gut and soil.

Third, such measures are limited by which environments have had their metagenomes sequenced, and how frequently. For example, species that live in root nodules may only be described as soil specialists if no root nodule metagenomes have been deposited into public databases, even though they may only transiently live in the soil. This is likely also the case for other species which may sometimes live in more niche habitats.

Future directions

When attempting to address whether bacterial lifestyle can explain why pangenomes are so variable, we think it is important to consider actual features of species' lifestyles. Additional measures are required to truly understand what features of pangenomes vary with a species' environment, and we have started carrying out such analyses. In particular, we suggest analyses based on the lifestyle of species, rather than these more passive presence and absence measures. A targeted, species by species approach, would ensure that the habitat and lifestyle of species was more accurately represented. It would also allow better comparisons between species that occupy similar environments, but which may have different potentials to acquire genes.

What is likely to be important is considering the environmental variability that is relevant to individual bacterial cells. Therefore, in future analyses, we will collect data on features of bacterial lifestyles that have clear predictions for how this may affect differences in the gain and loss of genes across genomes. This will also allow different aspects of species' lifestyle to be considered, beyond simply how variable their environments are. For example, we may expect bacteria that live inside other cells to encounter more stable environments, and so categorising species' based on different lifestyle traits would allow comparison of the pangenomes of intracellular vs free-living bacteria. Alternatively, categorising host-living species into those with either broad or narrow ranges, similar to the approach in Chapter 2, could be another feature of bacterial lifestyle that may be predicted to affect the structure of their pangenomes.

Additionally, to understand whether neutral or adaptive processes are more important for determining the structure of species' pangenomes, we propose analysing whether signatures of these processes are present in these genomes. Specifically, we will test whether signatures of positive selection are stronger in accessory genes compared to core genes, which would be predicted by the adaptive hypothesis. We would also analyse whether signatures of evolution in the accessory genome are more consistent with predominantly adaptive or neutral processes. Finally, we could analyse the presence of such adaptive or neutral signatures with respect to species' lifestyles. This would allow us to explore whether having additional accessory genes is always adaptive, but to different extents in certain species, or whether the genomes of some species do not appear to be under selection to have accessory genes at all.

References

Andreani, N.A., Hesse, E. & Vos, M. (2017). Prokaryote genome fluidity is dependent on effective population size. *The ISME Journal*, 11, 1719–1721.

Ding, W., Baumdicker, F. & Neher, R.A. (2018). panX: pan-genome analysis and exploration. *Nucleic Acids Res*, 46, e5.

Domingo-Sananes, M.R. & McInerney, J.O. (2021). Mechanisms That Shape Microbial Pangenomes. *Trends in Microbiology*, 29, 493–503.

Garcia-Garcera, M. & Rocha, E.P.C. (2020). Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. *Nature Communications*, 11, 758.

Goyal, A. (2018). Metabolic adaptations underlying genome flexibility in prokaryotes. *PLoS Genet*, 14, e1007763.

Hadfield, J.D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, 33, 1–22.

Hall, R.J., Whelan, F.J., Cummins, E.A., Connor, C., McNally, A. & McInerney, J.O. (2021). Gene-gene relationships in an Escherichia coli accessory genome are linked to function and mobility. *Microbial Genomics*, 7, 000650.

Ingram, C.J.E., Mulcare, C.A., Itan, Y., Thomas, M.G. & Swallow, D.M. (2009). Lactose digestion and the evolutionary genetics of lactase persistence. *Hum Genet*, 124, 579–591.

Maistrenko, O.M., Mende, D.R., Luetge, M., Hildebrand, F., Schmidt, T.S.B., Li, S.S., *et al.* (2020). Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *The ISME Journal*, 14, 1247–1259.

McInerney, J.O., McNally, A. & O'Connell, M.J. (2017a). Reply to 'The population genetics of pangenomes.' *Nat Microbiol*, 2, 1575–1575.

McInerney, J.O., McNally, A. & O'Connell, M.J. (2017b). Why prokaryotes have pangenomes. *Nat Microbiol*, 2, 17040.

Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, Genomes and evolution, 15, 589–594.

Niehus, R., Mitri, S., Fletcher, A.G. & Foster, K.R. (2015). Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nature Communications*, 6.

Polz, M.F., Alm, E.J. & Hanage, W.P. (2013). Horizontal Gene Transfer and the Evolution of Bacterial and Archaeal Population Structure. *Trends Genet*, 29, 170–175.

Puigbò, P., Lobkovsky, A.E., Kristensen, D.M., Wolf, Y.I. & Koonin, E.V. (2014). Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol*, 12, 66.

Whelan, F.J., Hall, R.J. & McInerney, J.O. (2021). Evidence for selection in the abundant accessory gene content of a prokaryote pangenome. *Molecular Biology and Evolution*.

Chapter 6: Discussion

Here, I first summarise and discuss the key results of my thesis. Each of Chapters 2-5 has their own discussion, and so I use this Chapter to draw links between the themes discussed in my thesis. I also discuss potential future directions, in particular for how we can study bacterial cooperation using comparative analyses across genomes. Finally, I also consider current problems with these kinds of analyses, and suggest how these can be resolved.

Summary and discussion of results

Horizontal gene transfer and bacterial cooperation

In Chapter 2, I used comparative genomics to test the hypothesis that horizontal gene transfer could help stabilise cooperation in bacteria. As in most previous studies examining this hypothesis, I focused on horizontal gene transfer via plasmids. I identified two key predictions of the hypothesis and tested these by comparing the chromosomes and plasmids of multiple genomes from 51 bacterial species. Both of these predictions related to where genes for cooperation would be expected to be located in bacterial genomes, if plasmids were particularly important in stabilising cooperation across species.

However, contrary to these predictions, I found that genes coding for extracellular proteins, which are likely to act as cooperative public goods, did not make up a higher proportion of: (i) plasmids compared to chromosomes; (ii) mobile plasmids compared to non-mobile plasmids. Therefore, across species, plasmids were not enriched with genes for cooperation. Furthermore, the mobility of plasmids has no impact on whether they carry cooperative genes. This suggests that even in cases where cooperative genes are coded for by plasmids, this is not to increase relatedness at the cooperative loci.

What does this mean for the cooperation hypothesis? We cannot say plasmids never stabilise cooperation, either in certain scenarios, or in certain species. Additionally, we did not consider other forms of horizontal gene transfer, and so cannot rule out a potential role of other mobile genetic elements here. However, I predict that analyses on other vectors of horizontal gene transfer, such as bacteriophages and integrative conjugative elements, would yield a similar lack of support for the hypothesis. I am collaborating with others in my research group to further explore whether this is the case.

Overall, Chapter 2 is arguably the most comprehensive test of this cooperation hypothesis to date. Our results suggest that horizontal gene transfer via plasmids is unlikely to play an important consistent role in determining where genes for cooperation are carried in bacterial genomes. Taken together, we found no evidence for a widespread and consistent role of plasmids in stabilising cooperation.

However, there are limitations of the analyses and results in Chapter 2, and I discuss the implications of these within the Chapter. One of the limitations surrounds the method we used to identify genes for cooperation. Later in this Chapter, I further discuss this limitation as part of a section on the future of studying the genetics of cooperation in bacteria. There, I suggest how new tools could help to improve the generality and reliability of future comparative genomic analyses that also focus on the evolution of cooperation.

Beyond plasmid mobility

I also tested the predictions of two alternate hypotheses for why genes for extracellular proteins were found more on plasmids of some species, but not others. First, plasmids could allow genes to be rapidly gained and lost depending on environmental conditions. This would be particularly useful for species which experienced more variable environments. Second, plasmid carriage of genes could provide benefits beyond the potential mobility of plasmids. These benefits may also be important for species with more variable environments, if plasmid carriage conferred benefits such as increased gene expression via high copy number, for example. To test these two hypotheses, I collected data on three different measures of species' environmental variability. I then examined the extent to which these explained variation in how overrepresented genes coding for extracellular proteins were on species' plasmids compared to chromosomes.

Overall, the results of Chapter 2 were most consistent with the hypothesis that genes for extracellular proteins may be carried on plasmids for reasons other than plasmid mobility. Of the three measures of environmental variability, I found that only one was significantly correlated with whether plasmids were enriched with genes coding for extracellular proteins. Specifically, I found that pathogen species with a broad host-range had the highest proportion of genes for extracellular proteins on their plasmids compared to their chromosomes. This was true when comparing these species to both pathogens with a narrow-host range, and non-

pathogen species. This suggests that pathogenic species with more variable environments code for proportionally more genes coding for extracellular proteins on their plasmids.

I then examined the reason for the difference between broad and narrow host-range pathogens. I found that plasmids of broad host-range plasmids were particularly enriched with genes for extracellular proteins involved with pathogenicity. Taken together with our finding that plasmid mobility did not explain why some plasmids carried more genes coding for extracellular proteins, an ability of plasmid transfer to allow hosts to gain and lose genes cannot explain the variation seen across species. Instead, plasmids seem likely to carry these genes for reasons other than plasmid mobility. Furthermore, these benefits seem particularly useful to pathogens.

One feature of plasmids that could provide a benefit to their hosts is copy number. Plasmids frequently exist in multiple copies per cell, which could potentially lead to high expression of plasmid genes relative to those on the chromosome. Moreover, there is also evidence that bacterial hosts may have some control over the copy number of their plasmids (Rodríguez-Beltrán *et al.* 2021). This potential high expression of plasmid genes, coupled with the ability of hosts to reduce expression when no longer required, could provide a major benefit of plasmids to their hosts. Crucially, these potential benefits of copy number would be possible for all plasmids, including plasmids incapable of transferring via conjugation. Therefore, copy number could be a feature of plasmids in general that provides a benefit to their hosts, and maintains genes which benefit from high expression on plasmids in the evolutionary long-term.

Whether the copy number of plasmids determines which genes they carry requires further analysis. Particularly, I think experimental studies would be most helpful, since copy number is not currently easy to estimate using plasmid sequences alone. First, I would experimentally quantify the copy number of plasmids in a wide range of strains and species. I would then examine how the copy number of plasmids varied with respect to key lifestyle characteristics of the different species, such as pathogenicity and host-range. I would expect that the copy number of broad host-range pathogens would, on average, be greater than non-pathogens or narrow host-range pathogens, if copy number is the feature driving the pattern I found in Chapter 2.

Additionally, to explicitly test how copy number may influence selection on genes for extracellular proteins, I would set up experiments comparing identical strains that differ only in the copy number of a plasmid coding for a particular extracellular protein, such as a protease. In a growth media where production of the protein was important for cell growth, a high copy number may be expected to lead to higher growth due to increased gene expression, and consequent protein production. Alternatively, very high copy number may confer additional energy and resource costs on the host, and outweigh any benefits of increased expression of the protein. Together, these experiments would provide insights into the relative benefits and costs of plasmid copy number, and whether this feature of plasmids could explain why they sometimes carry proportionally more genes for extracellular proteins.

Overall, Chapter 2 suggests that bacterial lifestyle and environmental conditions are potentially more important than bacterial sociality for determining whether extracellular proteins are coded for by plasmids.

Plasmid size, range, mobility and base content

In Chapter 3, I explored three key characteristics of plasmids: size, mobility and range. Each of these is highly variable across bacterial plasmids, and I considered how these correlated with one another across 3522 plasmids from 51 bacterial species. I speculated that these could be candidate 'life-history' traits of plasmids (Stearns 1992). Understanding how life history traits correlate with one another is a key question in evolutionary biology, and helps us to understand how natural selection acts upon these traits (Stearns 1989, 2000). Across plasmids, I found that plasmid mobility and range were positively correlated, suggesting that the ability to conjugate increases the range of a plasmid. In agreement with previous studies, I also found that conjugative plasmids were largest in size, while mobilizable plasmids were smallest (Smillie *et al.* 2010; Nishida 2012; Rodríguez-Beltrán *et al.* 2021). Finally, the direction of a correlation between plasmid range and size was different depending on the mobility of plasmids considered. While the effect sizes for these correlations were relatively low, these results suggest some potential interactions between selection on plasmid size, mobility and range.

In Chapter 4, I tested the predictions of two hypotheses for why plasmid sequences are consistently enriched with A and T bases. As expected, I found that plasmid AT-content was highly variable across plasmids, and consistently higher than chromosomes from the same species. I also found that the AT-content of plasmids was highly correlated with the AT-content

of the chromosome(s) of the species it was sequenced in. To test the two hypotheses, I analysed how plasmid AT-content varied with respect to plasmid mobility and range. Overall, my results were more consistent with the hypothesis that AT-bias in plasmids is due to a greater accumulation of mutations, which are biased towards A and T bases, in plasmid compared to chromosome sequences. The alternative hypothesis, that AT-bias is an adaptation to reduce the cost of plasmids to their hosts, was less consistent with my results.

Together, Chapters 3 and 4 provide insights into the various potential selection pressures that plasmid sequences may experience. Selection pressures on plasmids likely occur at multiple levels: from individual genes on plasmids, to plasmids as entities, and to selection on the benefits and costs of plasmids to their bacterial host. Selection at these different levels will not necessarily be in the same direction, which could potentially mask some of their effects. For example, for potential selection pressures on plasmid size, an individual gene will be under selection to be consistently gained by plasmids, regardless of whether this gain is beneficial to the plasmid, or the host the plasmid is found in. However, if selection at the plasmid and host level may be to lose a particular gene, this will be in conflict with selection on the gene itself. Therefore, the relative influence of selection at these multiple levels is also important to consider for understanding the evolution of plasmids, and their many potential roles in bacterial lifestyles.

Are plasmids analogous to endosymbionts?

The selection pressures that plasmids may be under could be indicative of life inside cells more generally. It has been suggested that plasmids and endosymbiotic bacteria have much in common (Dietel *et al.* 2018). Both live within the cytoplasm of their host, and as such may predicted to undergo similar selection pressures. For example, both will likely be under selection to maximize the likelihood of their retention within the cell, and also their own transmission (either vertically or horizontally). Similarly, both have sequences which are consistently AT-biased, suggesting that the same reasons may drive this phenomenon across intracellular elements more generally (Dietel *et al.* 2019). The extent to which these comparisons are useful for our understanding of the evolution of both plasmids and endosymbionts has yet to be fully addressed and tested. Still, a better understanding of how plasmids and endosymbionts adapt to life inside cells will help to identify how such relationships evolve, and why they remain stable. I suggest that future work should further explore similarities between plasmids and endosymbionts, including whether and to what

extent the study of their evolution can be considered under the same theoretical framework. Outstanding questions include:

(1) To what extent does selection act in the same way on plasmids and endosymbionts?

(2) How does the host-range and mobility of plasmids and endosymbionts affect selection on their benefits and costs to hosts?

(3) Are the signatures of selection and/or drift analogous in plasmid and endosymbiont sequences?

To answer these, I suggest a combination of comparative, theoretical and experimental approaches. First, I would use a comparative genomics approach to compare AT-bias among mutualistic and parasitic endosymbiont species. I would collect data on multiple endosymbiont species, including their AT-content and effect on host, from online databases. To control for potential confounding variables, I would also collect data on transmission (horizontal vs vertical), the duration of the endosymbiotic relationship, and genome size. I would use statistical methods that allow for phylogenetic relationships to be controlled for. This kind of project would be analogous to the plasmid analyses presented in Chapter 4 of this thesis. I have actually already begun initial data collection for this project, which I plan to expand in the future.

Horizontal gene transfer and the structure of bacterial genomes

While horizontal gene transfer does not appear to favour cooperation, it is still likely to affect bacterial evolution in other ways. In Chapter 5, we studied how the structure of bacterial pangenomes, defined as all the genes sequenced in a group of genomes, varied across species with different lifestyles. We used two measures of environmental variability to test the often-mentioned observation that species with more variable environments are those with more variable pangenomes (McInerney *et al.* 2017; Maistrenko *et al.* 2020).

We found that most measures of pangenome variability were not positively correlated with the number of environments a species was found in. However, we did find that the size of the core genome was positively correlated with environmental variability, in agreement with a previous study (Maistrenko *et al.* 2020). Overall, this suggests that while environmental variability is important for certain aspects of the structure of pangenomes, this may not be the case for all features. Future analyses should help to pull apart why only some measures of pangenome

structure were correlated with environmental variability. This is something I am currently working on, and plan to continue in the future.

When attempting to address whether bacterial lifestyle can explain why pangenomes are so variable, I think it is important to consider actual features of species' lifestyles. Measures such as the number of environments a species' 16S rRNA has been sequenced in may not capture the aspects of bacterial lifestyle that actually determine how selection and/or other evolutionary processes act on species' genomes. In Chapter 5, I discussed how I plan to collect data on features of bacterial lifestyles that have clear predictions for how they may affect differences in the gain and loss of genes across genomes. Analysing how these correlate with measures of pangenome variability will help to provide a better and more comprehensive answer to the question of how bacterial lifestyle could affect variation in species' pangenomes, and the role horizontal gene transfer may play in this.

The future of studying the genetics of cooperation in bacteria

Finally, I will now discuss two major outstanding questions for how we study cooperation in bacteria. First, how do we define a cooperative behaviour in bacteria, and how can we identify the genes which code for these behaviours? Second, how can comparative analyses across bacterial genomes improve our understanding of cooperation in bacteria, and how can potential issues of these analyses be resolved?

What does it mean to be a cooperative gene?

In Chapter 2, I explored where genes for cooperation were located in bacterial genomes. If we are to better understand the genetic basis of cooperation in bacteria, it is crucial that we can objectively define a cooperative gene. In general, a social behaviour is one which affects the fitness of an individual other than the actor. If the effect on another individual is positive, this is called a cooperative behaviour. Therefore, a cooperative gene is simply a gene which codes for a cooperative behaviour. Classic examples of cooperative behaviours in animals include sharing food, calling to warn of danger, and helping others to rear young.

However, what constitutes a cooperative behaviour in bacteria can be more difficult to comprehend. Bacteria are so small and simple that most of their actions can simply be condensed down into whether they produce certain molecules. Consequently, a large portion

of the field of microbiology is concerned with profiling the metabolic capacities of different strains and species of bacteria, and looking for ways of utilising this for various applications.

Similarly, cooperative behaviours in bacteria are often based upon the production of certain molecules. Instead of sharing food resources, bacteria produce enzymes to help break food down into smaller pieces (Allison 2005). Instead of calling to warn for danger, bacteria may produce signalling molecules upon their death to warn others against potential threats (LeRoux *et al.* 2015). Instead of helping others to rear young, bacteria secrete molecules that bind them together into biofilms, which may protect against external stresses and promote growth (Nocelli *et al.* 2016).

While the costs of these behaviours are usually proportional to the energy required make the molecule, the benefits are more complex. In the simplest case, the producing cell would be the only individual to benefit from its production of the molecule. This is the case for most molecules which remain inside bacterial cells, with any effect on growth or survival often referred to as 'private'. Therefore, this 'private' production would not be classed as cooperative, or indeed social, behaviours. Alternatively, others could benefit from an individual's production, which would make the behaviour a cooperative one. This is often the case for extracellular molecules, since they may diffuse away from the producing cell. However, how the benefits to oneself and others are partitioned can be unclear. First, any benefit to others could simply be a by-product of benefits directed to oneself. In this case, we would not need to invoke social evolution theory to explain the evolution of this behaviour. Second, rather than being a by-product, the benefit to other cells could instead itself be under positive selection. Behaviours that are under such selection because of their effect on others are those which require social evolution theory to explain.

Defining what constitutes a cooperative, and more generally a social, gene matters for how we ask questions about their evolution in bacteria. Due to the reasons discussed above, detailed experimental studies would be required to determine whether a candidate gene provides benefits to others, and is under selection because of its cooperative effect. However, it is usually unfeasible to achieve the same confidence in a cooperative effect when asking broad questions in comparative studies across species. Clearly, it would be inconceivable to have examined every gene in all 51 species for a potential cooperative effect on other cells, and then further consider how this effect on others compares to the effects on itself. Therefore, proxies of

cooperative genes are required if comparative genomics is to offer any insights into bacterial cooperation. In Chapter 2, we used the bioinformatics tool PSORTb to predict the subcellular location of the protein coded by each gene (Yu *et al.* 2010). Like previous studies, we defined a cooperative gene as any which coded for proteins that were secreted into the extracellular space (Nogueira *et al.* 2009).

Using genes coding for extracellular proteins as a proxy for cooperative genes has some definite advantages. First, it removes any subjectivity when it comes to defining a cooperative gene; instead the program simply searches for signal peptides, which are highly conserved sequences that allow proteins to be properly packaged for secretion. Second, the high fidelity of signal peptide sequences across species means we can estimate which genes are cooperative, even if no experimental study has ever considered cooperation in that species. This increases the number of species able to be included in such studies, thus increasing the phylogenetic breadth for asking questions across bacteria. Third, it takes the program PSORTb approximately 40 minutes to predict the subcellular location of each of the 5000 genes on an *Escherichia coli* chromosome. This rapid pace again means many more genomes and species can be included, in contrast to if genes needed to be assessed as cooperative or not by hand.

However, using such proxies also has some drawbacks. First, not all genes coding for extracellular proteins will necessarily be involved in bacterial cooperation. While many likely do act as some form of public good, some may simply be part of extracellular structures such as flagella, or instead tethered to the membrane by an additional protein. Second, even if genes do provide cooperative benefits, using this kind of proxy gives no information about how much of the benefits go to others compared to the producer. If a bacterium had evolved in a highly structured, viscous medium, many of the molecules it secreted could stay near to the producing cell. This would mean that much of the benefit would still be received by the producer. Third, there are many cooperative traits that could be missed by this type of proxy. This is because it only picks up secreted proteins, rather than all secreted molecules. Several well studied examples of cooperative molecules, such as iron-scavenging siderophores, are actually the product of multiple genes that work together inside the cell to produce the final siderophore product. This kind of molecule is referred to as a secondary metabolite. Therefore, while the proteins produced by such genes are intracellular, their function is to produce cooperative extracellular molecules. This means these kinds of cooperative genes are not classed as extracellular by the program PSORTb.

As new tools become available, our ability to identify cooperative genes will improve, both in speed and accuracy. A recent study by Simonet and McNally identified genes for cooperative public goods by using the program PANNZER, which annotates genes with predicted gene-ontology terms based on signatures in their sequence, in addition to PSORTb (Koskinen *et al.* 2015; Simonet & McNally 2021). The authors then searched these annotations for key gene-ontology terms associated with cooperative behaviours, such as quorum sensing and antibiotic degradation, and labelled those with any positive matches as 'cooperative'. This approach allowed for secondary metabolite genes to be included in the analyses, something not possible when using PSORTb alone.

Additionally, the lead author of the Belcher *et al.* paper*,* which can be found in the appendix of this thesis, manually separated genes for certain behaviours into those conferring either public or private benefits, in genomes of the species *Pseudomonas aeruginosa*. This approach allowed the direct comparison of signatures of selection between genes that were expected to be under kin selection and those that were not, while controlling for potentially confounding effects of differential gene expression. Together, these two approaches emphasise that how cooperative genes are identified for analysis should depend on the scope and purpose of the study.

Furthermore, the study of bacterial cooperation using experiments remains crucial to further understand how the benefits of behaviours are partitioned between the producer and any recipients. Experiments that identify whether behaviours are cooperative, and in what conditions these are selected for, would be particularly useful for bacterial species which have so far received less research attention in relation to their cooperative behaviours. These would help to inform and improve future attempts to identify the genes for cooperation in bacteria.

Comparative genomics and evolutionary questions

The techniques used in this thesis are a combination of genome bioinformatics and comparative analyses. Together, the term comparative genomics has been used to describe these types of studies.

Comparative analyses across species have helped to shape the way we view much of evolutionary biology. For example, thanks to comparative studies, we now know that: (i)

cooperative breeding in birds is associated with low promiscuity; (ii) paternal care in fish may be due to sexual selection through female preference; (iii) inbreeding avoidance only evolves when there is a risk of both inbreeding depression and related mates encountering one another (Cornwallis *et al.* 2010; Goldberg *et al.* 2020; Pike *et al.* 2021). However, there are relatively few comparative studies on bacteria that aim to ask evolutionary questions. This is potentially because so much of the lifestyle and environments of bacteria are still relatively unknown to us, making them a difficult candidate for broad questions across species.

While most of the comparative work on animals has focused on phenotypic traits, bacteria offer an opportunity to ask broad questions at the genotypic level. As of August 2021, there were 24,761 complete bacterial genomes on the RefSeq/ NCBI public database. That is three times as many as the number of animal genomes, including those at all levels of completion. Therefore, the huge amount of genomic data currently available across bacterial species offers an opportunity to understand the evolution of cooperation at the genetic level, in a way not currently possible for animals.

However, larger quantities of data come with increasing challenges in distinguishing artefacts from real biological effects. This is partly because many of the statistical techniques used most frequently in biology were designed for experimental studies, with expected sample sizes relatively low. As a result, large datasets comprising thousands of data points are more likely to produce arbitrary significant results when analysed using these statistical tests. This is not to say significant results in large datasets are always uninformative. Indeed, a large number of data points can be useful for distinguishing small but real effects that would otherwise not be picked up in a smaller-scale analysis.

Instead, statistics on large datasets requires moving beyond considering p-values and significance alone; the size of the effect becomes just as important as the significance for these kind of analyses (Crawley 2014). With large enough datasets, statistical tests will almost always produce a significant result, unless the effect size is exactly zero (Sullivan & Feinn 2012). This is not to say that datasets should be reduced in size to reduce the risk of false positives – quite the opposite. The ability of comparative analyses to draw conclusions across species relies on datasets that are large enough to sufficiently capture the diversity present in nature. Instead, when analysing a large dataset, a p-value of less than 0.05 should not be used

as a reason to conclude the presence of an effect, but should be treated as a 'flag' for where the size of the effect should be investigated, and robustness analyses conducted if necessary.

Additionally, genomic datasets pose further problems. Most statistical tests assume that data points are independent from one another. As discussed throughout this thesis, this is clearly not the case for the genomes available in public databases. Thus, controlling for bacterial phylogeny and number of genomes per species using statistical techniques such as MCMCglmm is crucial for identifying whether a result is a real biological effect, or an artefact of a biased dataset.

Perhaps more difficult to address is the problem that genomic databases are largely unrepresentative of bacteria as a whole. Currently, there are more than 10 times as many *Escherichia coli* genomes available than of the whole Cyanobacteria phylum, despite the latter being hugely diverse and found in almost every aquatic environment on Earth. This suggests that much of the distribution of genomes currently available is due to factors entirely different to distribution and prevalence on Earth. If entire clades of bacteria are missing from the original dataset, this limits our ability to draw broad conclusions across the bacterial domain. This should improve as sequencing becomes cheaper and more feasible for lesser-studied species. To accelerate this improvement, I suggest focusing sequencing efforts particularly on bacterial species currently underrepresented in genomic databases. In the meantime, I would emphasise the importance of careful consideration when compiling datasets for multi-species comparative genomics studies, and identifying and controlling for any biases present.

In general, comparative genomics has the potential to give new and exciting insights into bacterial cooperation and evolution. It can help us test broad ideas and hypotheses across species, as I have done in this thesis. However, there are limitations of these kinds of analyses, and many factors that need to be considered and controlled for. Otherwise, there is the risk of overinterpreting results which may instead be artefacts of current genomic datasets.

References

Allison, S.D. (2005). Cheaters, diffusion and nutrients constrain decomposition by microbial enzymes in spatially structured environments. *Ecology Letters*, 8, 626–635.

Cornwallis, C.K., West, S.A., Davis, K.E. & Griffin, A.S. (2010). Promiscuity and the evolutionary transition to complex societies. *Nature*, 466, 969–972.

Crawley, M.J. (2014). *Statistics: An Introduction Using R*. John Wiley & Sons.

Dietel, A.K., Kaltenpoth, M. & Kost, C. (2018). Convergent Evolution in Intracellular Elements: Plasmids as Model Endosymbionts. *Trends in Microbiology*, 26, 755–768.

Dietel, A.-K., Merker, H., Kaltenpoth, M. & Kost, C. (2019). Selective advantages favour high genomic AT-contents in intracellular elements. *PLOS Genetics*, 15, e1007778.

Goldberg, R.L., Downing, P.A., Griffin, A.S. & Green, J.P. (2020). The costs and benefits of paternal care in fish: a meta-analysis. *Proc. R. Soc. B.*, 287, 20201759.

Koskinen, P., Törönen, P., Nokso-Koivisto, J. & Holm, L. (2015). PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics*, 31, 1544–1552.

LeRoux, M., Kirkpatrick, R.L., Montauti, E.I., Tran, B.Q., Peterson, S.B., Harding, B.N., *et al.* (2015). Kin cell lysis is a danger signal that activates antibacterial pathways of Pseudomonas aeruginosa. *eLife*, 4, e05701.

Maistrenko, O.M., Mende, D.R., Luetge, M., Hildebrand, F., Schmidt, T.S.B., Li, S.S., *et al.* (2020). Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *The ISME Journal*, 14, 1247–1259.

McInerney, J.O., McNally, A. & O'Connell, M.J. (2017). Why prokaryotes have pangenomes. *Nat Microbiol*, 2, 17040.

Nishida, H. (2012). Comparative Analyses of Base Compositions, DNA Sizes, and Dinucleotide Frequency Profiles in Archaeal and Bacterial Chromosomes and Plasmids. *International Journal of Evolutionary Biology*, 2012, e342482.

Nocelli, N., Bogino, P.C., Banchio, E. & Giordano, W. (2016). Roles of Extracellular Polysaccharides and Biofilm Formation in Heavy Metal Resistance of Rhizobia. *Materials*, 9, 418.

Nogueira, T., Rankin, D.J., Touchon, M., Taddei, F., Brown, S.P. & Rocha, E.P.C. (2009). Horizontal Gene Transfer of the Secretome Drives the Evolution of Bacterial Cooperation and Virulence. *Current Biology*, 19, 1683–1691.

Pike, V.L., Cornwallis, C.K. & Griffin, A.S. (2021). Why don't all animals avoid inbreeding? *Proceedings of the Royal Society B: Biological Sciences*, 288, 20211045.

Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R.C. & San Millán, Á. (2021). Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology*, 1–13.

Simonet, C. & McNally, L. (2021). Kin selection explains the evolution of cooperation in the gut microbiota. *PNAS*, 118.

Smillie, C., Garcillan-Barcia, M.P., Francia, M.V., Rocha, E.P.C. & de la Cruz, F. (2010). Mobility of Plasmids. *Microbiology and Molecular Biology Reviews*, 74, 434–452.

Stearns, S.C. (1989). Trade-Offs in Life-History Evolution. *Functional Ecology*, 3, 259.

Stearns, S.C. (1992). *The evolution of life histories*.

Stearns, S.C. (2000). Life history evolution: successes, limitations, and prospects. *Naturwissenschaften*, 87, 476–486.

Sullivan, G.M. & Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. *J Grad Med Educ*, 4, 279–282.

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., *et al.* (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26, 1608–1615.

S1 – Supplementary Results for Genomic Analyses

Plasmids with higher mobility do not carry more genes for extracellular proteins.

We found no difference in the proportion of genes coding for extracellular proteins across the three plasmid mobility types when we compared the means of each mobility type of each species (MCMCglmm; Table S2, row 12).

We also found no significant difference when: (a) carrying out a regression between the proportion of genes coding for extracellular proteins and plasmid 'mobility' treated as a continuous variable (MCMCglmm; Table S2, row 13); (b) testing for a correlation between the proportion of a species' plasmids which can transfer (are either conjugative or mobilizable) and the proportion of plasmid genes coding for extracellular proteins (Fig S5) (MCMCglmm; Table S2, rows 18 and 19); (c) testing for a correlation between the proportion of a species' plasmids which can transfer and how overrepresented or underrepresented extracellular proteins are on plasmids compared to chromosomes (Extended Data Figure 4) (MCMCglmm; Table S2, rows 16 and 17).

As discussed in the previous section, if non-independence is not controlled for, then there is the potential for misleading analyses and spurious significant results. This is especially a problem with analyses on large datasets. Consequently, it is important to examine biological effect sizes, and not just p-values[1]. For example, when we assumed that all 3522 plasmids, were independent data points, we found that 1.8% of conjugative plasmid genes code for extracellular proteins, compared to 1.4% of non-mobilizable plasmid genes. This means that for every 100 plasmid genes, conjugative plasmids carry less than half an additional extracellular protein-coding gene compared to non-mobilizable plasmids. Despite this marginal effect, a MCMCglmm model on this data produced significant pMCMC values for comparisons of the three plasmid mobility types, even though mobility only explains 1.5% of the variation in the proportion of genes coding for extracellular proteins (MCMCglmm; Table S2, rows 14 and 15).

Transfer rates of conjugative, mobilizable and non-mobilizable plasmids.

We have considered the relative rates of transfer among the three mobility types, where conjugative plasmids transfer at faster rates than mobilizable, and mobilizable transfer at faster rates than non-mobilizable[2]. However, the variation in transfer rates within plasmids of the

same mobility type is likely to be large, and mobilization via mechanisms other than conjugation, such as phage transfer, is possible[2–5].

Additionally, if mobilizable plasmids almost always co-occur with conjugative plasmids, they would transfer at a similar rate as conjugative plasmid(s), or potentially even faster if they were smaller and could replicate faster. We examined how frequently the mobilizable plasmids in our dataset co-occurred with conjugative plasmids. There were 727 genomes which carried at least one mobilizable plasmid, comprising 46 species. Of these, 40% (293/676) also carried a conjugative plasmid, while 60% (434/727) did not. This may be biased by a few species with a large number of genomes, so we also analysed the data at the species level to control for this. For each species, we grouped the genomes with mobilizable plasmids into those with and without a conjugative plasmid. We found that 37% of species (17/46) had a majority of genomes which also carried a conjugative plasmid, while 61% (28/46) of species had a majority of genomes which did not carry a conjugative plasmid. One species, Campylobacter coli, had only two genomes which carried a mobilizable plasmid, one of which carried a conjugative plasmid and the other did not.

This suggests that mobilizable plasmids frequently, and potentially more often than not, occur without a conjugative plasmid. This frequent absence of transferability for mobilizable plasmids is likely to lead to a lower transfer rate compared to conjugative plasmids. This supports the use of 'mobility type' as a proxy for transfer rate, specifically that mobilizable plasmids will transfer at a lower rate than conjugative plasmids, on average. However, the variation in transfer rates within plasmids of the same mobility type is likely to be large, and mobilization via mechanisms other than conjugation, such as phage transfer, is possible[2–5]. Quantitative estimates of plasmid transfer rates would help to address these added complications[6], and further examine any potential effect of plasmid mobility on the kinds of genes plasmids carry.

Mobilizable plasmids do not code for more extracellular proteins when they co-occur with conjugative plasmids.

We also examined whether mobilizable plasmids which co-occurred with conjugative plasmids had a greater % of genes that coded for extracellular proteins than those without a conjugative plasmid. This would be expected under the cooperation hypothesis, which suggests that

plasmid mobility is the key driver of whether a cooperative gene should be located on plasmids. We compared genomes with mobilizable plasmids within each species, considering only species which had at least one genome both with and without a conjugative plasmid. We found that for 43% (15/36) of species, mobilizable plasmids that co-occurred with a conjugative plasmid(s) had a greater % of genes coding for extracellular proteins than those without, while for 40% (14/36) of species, mobilizable plasmids that co-occurred with a conjugative plasmid(s) had a lower % of genes coding for extracellular than those without a conjugative plasmid. The remaining 17% (6/36) of species had no extracellular proteins on any of their mobilizable plasmids, and so the % for both was 0.

We also analysed this data using a MCMCglmm analysis to control for phylogeny, and found that there was no significant difference between the proportion of genes coding for extracellular proteins for mobilizable plasmids that co-occurred with conjugative plasmid(s) compared to those that did not co-occur with conjugative plasmids (Table S2, Rows 38 & 39). This suggests that co-occurence with a conjugative plasmid has little impact on whether mobilizable plasmids carry genes for extracellular proteins.

Number of environments.

We used recently published data which assigned bacterial species to living in one or more of five broad environments: host, soil, sediment, wastewater and water[7–9]. Of species in our analysis, 36 had been assigned to at least one of these environments. We found no significant correlation between the number of environments a species was found in and how likely genes coding for extracellular proteins were to be on plasmids (Figure S9) (MCMCglmm; Table S2, row 34). We also found no significant correlation when we supplemented the published environmental data with information from the literature, so that all species in our dataset were included in the analysis (Extended Data Figure 6a; Supp X) (MCMCglmm; Table S2, row 35).

Garcia-Garcera and Rocha (2020) found that the proportion of a species' genome which coded for extracellular proteins increased with the number of environments a species was found in[8]. This is a slightly different, but related question. When we asked the same question with our data, we found a non-significant pattern, but in the same direction: the number of five broad environments in which each species was found was positively correlated with the proportion of genes coding for extracellular proteins across the genome increased (Fig S10)

(MCMCglmm; Table S2, row 36). Garcia-Garcera and Rocha analysed data for over 1000 bacterial species, and so had greater statistical power to obtain a significant result. They also used MCMCglmm to control for phylogeny. In addition, this relationship could be relatively weak because the five environments are very broad and there is likely to be significant variability within these environments.

Core vs accessory genes.

Bacterial genes are often split up into 'core' genes, found in all genomes of a species, and 'accessory' genes, found in only a subset of a species' genomes[10]. Species which encounter more variable environments are expected to have relatively more accessory genes compared to core genes in their genomes[11]. Consequently, the proportion of each species' genomes composed of 'core' genes could be used as a proxy of environmental variability, by assuming that species which encounter more variable environments will have a smaller proportion of core genes. We used data from PanX[12] to calculate the proportion of each species' genomes which were core. We found no significant correlation between the proportion of each species' genomes which are core genes and the likelihood that genes coding for extracellular proteins are on plasmids (Extneded Data Figure 6b) (MCMCglmm; Table S2, row 37).

Effect sizes, variance explained and significance.

The percentage of variance explained that is considered biologically significant is subjective and can depend upon the kind of data you are examining, and the field of research. In many areas of evolution and ecology, 5-10% can be a reasonable baseline, but in some areas 1% could be argued for[1,13]. For example, when including all analyses both significant and non-significant in the field of behavioural ecology, the average variance explained is approximately 4%, and so a value greater than 4% would be above background noise[14]. In particularly successful areas, such as the field of sex allocation, where a relatively good fit between theory and data can be expected, the percentage variance explained can average 28% across studies within species, and be as high as 93%[15,16].

References

1. Crawley, M. J. *Statistics: An Introduction Using R*. (John Wiley & Sons, 2014).
2. 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).
3. 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).
4. Ramsay, J. P. & Firth, N. Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Curr. Opin. Microbiol.* **38**, 1–9 (2017).
5. Rodríguez-Rubio, L. *et al.* Extensive antimicrobial resistance mobilization via multicopy plasmid encapsidation mediated by temperate phages. *J. Antimicrob. Chemother.* **75**, 3173–3180 (2020).
6. 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).
7. 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).
8. 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).
9. 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).
10. Domingo-Sananes, M. R. & McInerney, J. O. Mechanisms That Shape Microbial Pangenomes. *Trends Microbiol.* **0**, (2021).
11. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040 (2017).
12. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* **46**, e5 (2018).
13. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Routledge, 1988).
14. 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).
15. West, S. A., Shuker, D. M. & Sheldon, B. C. Sex-ratio adjustment when relatives interact: a test of constraints on adaptation. *Evol. Int. J. Org. Evol.* **59**, 1211–1228 (2005).
16. West, S. *Sex Allocation*. (Princeton University Press, 2009).

S2 – Supplementary Tables

	Chromosome(s)		Plasmid(s)	
Protein Location	Number	% of Total	Number	% of Total
Cytoplasmic	1614	59.9%	134	59.8%
Cytoplasmic Membrane	893	33.1%	71	31.6%
Extracellular	52	1.9%	5	2.4%
	Gram-Negative			
Periplasmic	109	4.0%	15	5.3%
Outer Membrane	76	2.8%	5	1.9%
	Gram-Positive			
Cell Wall	39	1.5%	2	1.5%
Unknown	957	26.3%	224	38.3%

Table S1. Summary of location of genes encoding each subcellular localisation across species.

For schematic of these localisations see Figure S1. Cytoplasmic, cytoplasmic membrane and extracellular protein values are the mean number per genome calculated across all genomes of a species, and then the means across all species. Periplasmic and outer membrane values are the mean calculated across only Gram-negative species, while cell wall values are the mean calculated across only Gram-positive species. Percentages are out of all genes with a known localisation, except for unknown protein percentages which are of all proteins.

S2

Table S2. MCMCglmm analyses

We ran all MCMCglmm models with uninformative priors (V=1, nu=0.002).

Note: Unless otherwise stated, we arcsine square root transformed all proportion data.

	Model description	Sample size	Posterior mean	95% Credible Interval	pMCMC	R^2 value (if calculated)
Location of extracellular proteins within bacterial genomes						
1a	Difference in plasmid and chromosome extracellular proportions ~ 1. Random effects: phylogeny + number of genomes per species.	1632 genomes	0.004	-0.063 to 0.057	0.87 (NS)	Phylogeny = 0.17. Number of genomes per species = 0.47
1b	Difference in plasmid and chromosome extracellular proportions ~ 1. Random effects: number of genomes per species.	1632 genomes	0.007	-0.021 to 0.034	0.644 (NS)	
2	Ratio of plasmid and chromosome extracellular proportions ~ 1. Random effects: phylogeny + number of genomes per species.	1632 genomes	1.017	0.695 to 1.348	N/A (1 is within 95% CI, so ratio is not significantly different to 1).	
3	Each genome assigned 1 if plasmid > chromosome proportion, and 0 if plasmid < chromosome proportion.	1632 genomes	17.82	-69.90 to 128.97	0.558 (NS)	

S2

	Model uses categorical family response variable. Assigned value ~ 1. (This asks whether more 0s or 1s in the data). Random effects: phylogeny + number of genomes per species.					
4	Difference in plasmid and chromosome extracellular proportions ~ 1. Proportion data un-transformed before calculating difference. Random effects: phylogeny + number of genomes per species.	1632 genomes	0.017	-0.021 to 0.057	0.332	Phylogeny = 0.34. Number of genomes per species = 0.46.
Location of other protein classes within bacterial genomes						
5	Difference in plasmid and chromosome cytoplasmic proportions ~ 1. Random effects: phylogeny + number of genomes per species.	1632 genomes	0.090	-0.008 to 0.209	0.074 (NS)	
6	Difference in plasmid and chromosome cytoplasmic membrane proportions ~ 1. Random effects: phylogeny + number of genomes per species.	1632 genomes	-0.129	-0.295 to 0.012	0.088 (NS)	
7	Difference in plasmid and chromosome periplasmic proportions ~ 1. Random effects: phylogeny + number of genomes per species.	1027 genomes (only Gram-negative species)	-0.048	-0.183 to 0.127	0.482 (NS)	

S2

8	Difference in plasmid and chromosome outer membrane proportions ~ 1. Random effects: phylogeny + number of genomes per species.	1027 genomes (only Gram-negative species)	-0.075	-0.192 to 0.040	0.158 (NS)	
9	Difference in plasmid and chromosome cell wall proportions ~ 1. Random effects: phylogeny + number of genomes per species.	605 genomes (only Gram-positive species)	-0.028	-0.120 to 0.052	0.418 (NS)	
10	Difference in plasmid and chromosome unknown localisation proportions ~ 1. Random effects: phylogeny + number of genomes per species.	1632 genomes	0.156	0.089 to 0.224	0.002 (**)	
Plasmid mobility and extracellular proteins						
11	Slope value of mean plasmid extracellular proportion vs mobility ~ 1. Random effect: phylogeny.	40 slopes (one for each species with all three plasmid mobilities)	0.006	-0.040 to 0.052	0.73 (NS)	Phylogeny = 0.33.
12	Mean plasmid extracellular proportion ~ plasmid mobility. (Mobility as a factor with three levels) Random effect = phylogeny.	138 (mean for each plasmid mobility, so most species (40) have three data points)	Conjugative compared to non-mobilizable = 0.013. Mobilizable compared to non-mobilizable = -0.019.	Conjugative compared to non-mobilizable = -0.023 to 0.055. Mobilizable compared to non-mobilizable = -0.060 to 0.016.	Conjugative compared to non-mobilizable = 0.514 (NS). Mobilizable compared to non-mobilizable = 0.354 (NS).	
13	Mean plasmid extracellular proportion ~ plasmid mobility. (Here, non-mobiilzable = 1,	138 (mean for each plasmid mobility, so	Intercept = 0.098. Slope = 0.006.	Intercept = 0.001 to 0.183. Slope = -0.012 to 0.028.	Intercept = 0.042 (*) Slope = 0.546 (NS)	

	mobilizable = 2, conjugative = 3, so mobility is numeric and model is a regression).	most species (40) have three data points)				
14	Plasmid extracellular proportion ~ plasmid mobility. (Mobility as a factor with three levels) Random effects = phylogeny + number of plasmids per species.	3522 (one for each plasmid with a mobility prediction)	Conjugative compared to non-mobilizable = 0.015. Mobilizable compared to non-mobilizable = -0.033.	Conjugative compared to non-mobilizable = 0.004 to 0.026. Mobilizable compared to non-mobilizable = -0.044 to -0.023.	Conjugative compared to non-mobilizable = 0.008 (**). Mobilizable compared to non-mobilizable = <0.001 (***).	Plasmid mobility = 0.015. Phylogeny = 0.13. Number of plasmids per species = 0.29.
15	Plasmid extracellular proportion ~ plasmid mobility. (Here, non-mobiilzable = 1, mobilizable = 2, conjugative = 3, so mobility is numeric and model is a regression).	3522 (one for each plasmid with a mobility prediction)	Intercept = 0.102. Slope = 0.006.	Intercept = 0.046 to 0.170. Slope = -0.0002 to 0.011.	Intercept = 0.008 (**) Slope = 0.056 (NS).	
16	Mean difference in plasmid and chromosome extracellular proportions ~ mean proportion of plasmids which are conjugative. Random effect = phylogeny.	51 (mean difference and conjugative proportion for each species)	Intercept = -0.0003. Slope = -0.001.	Intercept = -0.075 to 0.076. Slope = -0.084 to 0.064.	Intercept = 0.996 (NS). Slope = 0.988 (NS).	
17	Mean difference in plasmid and chromosome extracellular proportions ~ mean proportion of plasmids which are conjugative or mobilizable. Random effect = phylogeny.	51 (mean difference and conjugative/ mobilizable proportion for each species)	Intercept = -0.016. Slope = 0.017.	Intercept = -0.125 to 0.079. Slope = -0.076 to 0.101.	Intercept = 0.78 (NS). Slope = 0.668 (NS).	
18	Mean plasmid extracellular proportion ~ mean proportion of plasmids which are conjugative. Random effect = phylogeny.	51 (mean extracellular proportion and	Intercept = 0.133. Slope = -0.006.	Intercept = 0.061 to 0.205. Slope = -0.087 to 0.065.	Intercept = 0.008 (**). Slope = 0.91 (NS).	

S2

		conjugative proportion for each species)				
19	Mean plasmid extracellular proportion ~ mean proportion of plasmids which are conjugative or mobilizable. Random effect = phylogeny.	51 (mean extracellular proportion and conjugative/ mobilizable proportion for each species)	Intercept = 0.109. Slope = 0.024.	Intercept = 0.004 to 0.221. Slope = -0.069 to 0.109.	Intercept = 0.05 (*). Slope = 0.578 (NS).	
20	Mean difference in non-mobilizable plasmid and chromosome extracellular proportions ~ 1. Random effect = phylogeny.	48 (mean difference for each species, 3 species had no genomes with a non-mobilizable plasmid(s))	0.016	-0.085 to 0.054	0.638 (NS)	
21	Mean difference in conjugative/mobilizable plasmid and chromosome extracellular proportions ~ 1. Random effect = phylogeny.	48 (mean difference for each species, 3 species had no genomes with a mobilizable/ conjugative plasmid(s))	-0.041	-0.117 to 0.051	0.292 (NS)	
22	Mean difference in conjugative plasmid and chromosome extracellular proportions ~ 1.	44 (mean difference for each species, 7	0.004	-0.078 to 0.102	0.924 (NS)	

S2

	Random effect = phylogeny.	species had no genomes with a conjugative plasmid(s))				
Host-range of pathogens						
23	Difference in plasmid and chromosome extracellular proportions ~ pathogenicity/host range (factor with three levels: non-pathogen, narrow host-range pathogen, and broad host-range pathogen). Random effects: phylogeny + number of genomes per species.	701 genomes (all genomes from 25 species)	Non-pathogen compared to broad host-range pathogen = -0.161. Narrow host-range pathogen compared to broad host-range pathogen = -0.222.	Non-pathogen compared to broad host-range pathogen = -0.252 to -0.067. Narrow host-range pathogen compared to broad host-range pathogen = -0.322 to -0.123.	Non-pathogen compared to broad host-range pathogen = <0.001 (***). Narrow host-range pathogen compared to broad host-range pathogen = <0.001 (***)	Pathogenicity/ host-range = 0.35. Phylogeny = 0.11. Number of genomes per species = 0.28.
24	Difference in plasmid and chromosome extracellular proportions ~ pathogenicity (factor with two levels: non-pathogen and pathogen). Random effects: phylogeny + number of genomes per species.	701 genomes (all genomes from 25 species)	Pathogen compared to non-pathogen = 0.106.	Pathogen compared to non-pathogen = -0.22 to 0.218.	Pathogen compared to non-pathogen = 0.092 (NS)	
25	Difference in plasmid and chromosome extracellular proportions ~ pathogenicity/host-range (factor with two levels: non-pathogen and narrow host-range pathogen).	389 genomes (all genomes from 15 species)	Non-pathogen compared to narrow host-range pathogen = 0.031.	Non-pathogen compared to narrow host-range pathogen = -0.065 to 0.127.	Non-pathogen compared to narrow host-range pathogen = 0.482 (NS).	
Pathogenicity of extracellular proteins						

26	Difference in plasmid and chromosome pathogenic extracellular proportions ~ host range. Only in broad and narrow host-range pathogens. Random effects: phylogeny + number of genomes per species.	474 genomes (genomes from 15 species)	Narrow host-range compared to broad host-range = -0.209.	Narrow host-range compared to broad host-range = -0.350 to -0.086.	Narrow host-range compared to broad host-range = 0.012 (*).	
27	Difference in plasmid and chromosome non-pathogenic extracellular proportions ~ host-range. Only in broad and narrow host-range pathogens. Random effects: phylogeny + number of genomes per species.	474 genomes (genomes from 15 species)	Narrow host-range compared to broad host-range = -0.034.	Narrow host-range compared to broad host-range = -0.108 to 0.035.	Narrow host-range compared to broad host-range = 0.296 (NS).	
28	Difference in plasmid and chromosome pathogenic extracellular proportions ~ human pathogenicity (factor with two levels: human or non-human). Only in broad and narrow host-range pathogens.	474 genomes (genomes from 15 species)	Non-human compared to human = 0.012.	Non-human compared to human = -0.156 to 0.187.	Non-human compared to human = 0.838 (NS).	
29	Difference in plasmid and chromosome non-pathogenic extracellular proportions ~ human pathogenicity. Only in broad and narrow host-range pathogens.	474 genomes (genomes from 15 species)	Non-human compared to human = -0.008.	Non-human compared to human = -0.074 to 0.059.	Non-human compared to human = 0.812 (NS).	
30	Difference in plasmid and chromosome pathogenic extracellular proportions ~ host-range + human pathogenicity. Only in broad and narrow host-range pathogens.	474 genomes (genomes from 15 species)	Host-range = -0.212. Human pathogenicity = -0.021.	Host-range = -0.366 to -0.77. Human pathogenicity = -0.157 to 0.105.	Host-range = 0.012 (*). Human pathogenicity = 0.740 (NS).	

S2

Pathogenic extracellular proteins and plasmid mobility						
31	Slope value of mean plasmid pathogenic extracellular proportion vs mobility ~ 1. Only broad host-range pathogens with plasmids of all three moiblities). Random effect: phylogeny.	7 (a slope for each broad host-range pathogen species with plasmids of all three mobilities)	-0.020	-0.224 to 0.185	0.774 (NS)	
32	Mean plasmid pathogenic extracellular proportion ~ plasmid mobility. (Mobility as a factor with three levels) All broad host-range pathogen species. Random effect: phylogeny.	26 (mean for each plasmid mobility; seven have 3 data points, three have 1 or 2).	Mobilizable compared to non-mobilizable = 0.0001. Conjugative compared to non-mobilizable = -0.049.	Mobilizable compared to non-mobilizable = -0.179 to 0.139. Conjugative compared to non-mobilizable = -0.212 to 0.099.	Mobilizable compared to non-mobilizable = 0.974. (NS) Conjugative compared to non-mobilizable = 0.528 (NS).	
33	Mean plasmid pathogenic extracellular proportion ~ plasmid mobility. (Mobility as a factor with three levels) All narrow host-range pathogen species. Random effect: phylogeny.	11 (mean for each plasmid mobility; two have 3 data points, three have 1 or 2).	Mobilizable compared to non-mobilizable = 0.003. Conjugative compared to non-mobilizable = 0.121.	Mobilizable compared to non-mobilizable = -0.128 to 0.118. Conjugative compared to non-mobilizable = -0.020 to 0.260.	Mobilizable compared to non-mobilizable = 0.972 (NS). Conjugative compared to non-mobilizable = 0.076 (NS).	
Number of five broad environments						
34	Difference in plasmid and chromosome extracellular proportions ~ number of environments.	1360 genomes (all genomes from 36 species with data on	Intercept = -0.026. Slope = 0.013.	Intercept = -0.098 to 0.057. Slope = -0.015 to 0.042.	Intercept = 0.498 (NS). Slope = 0.350 (NS).	

	Random effects: phylogeny + number of genomes per species.	number of environments)				
35	Difference in plasmid and chromosome extracellular proportions ~ number of environments (supplemented with literature). Random effects: phylogeny + number of genomes per species.	1632 genomes	Intercept = 0.017. Slope = -0.006.	Intercept = -0.055 to 0.115. Slope = -0.036 to 0.016.	Intercept = 0.562 (NS). Slope = 0.492 (NS).	
36	Genome extracellular proportion ~ number of environments (supplemented with literature). Random effects: phylogeny + number of genomes per species.	1632 genomes	Intercept = 0.138. Slope = 0.001.	Intercept = 0.102 to 0.181. Slope = -0.004 to 0.007.	Intercept = <0.001 (***). Slope = 0.596 (NS).	
Core vs accessory genome						
37	Difference in plasmid and chromosome extracellular proportion ~ core gene proportion. Random effects: phylogeny + number of genomes per species.	1632 genomes	Intercept = -0.075. Slope = -0.084.	Intercept = -0.041 to 0.205. Slope = -0.218 to 0.034.	Intercept = 0.228 (NS). Slope = 0.170 (NS).	
Gene content of mobilizable plasmids present with and without conjugative plasmids						
38	Proportion of genes coding extracellular proteins for mobilizable plasmid(s) in genome ~ whether conjugative plasmid also present in genome. Random effects: phylogeny.	46 species (those which had >= 1 genome with a mobilizable plasmid)	Without conjugative compared to conjugative = 0.002.	Without conjugative compared to conjugative = -0.032 to 0.038.	Without conjugative compared to conjugative = 0.912 (NS).	
39	Mean difference in extracellular proportion of mobilizable plasmids for genomes with vs without conjugative plasmids ~ 1.	35 species (those which had >=1 genome with a	Intercept = 0.003.	Intercept = -0.066 to 0.061.	Intercept = 0.922 (NS).	

	Random effects: phylogeny.	mobilizable plasmid both with and without a conjugative plasmid.				

Table S3. Measures of Bacterial Lifestyle and Environmental Variability

Below is a table of literature references used to categorise species': (i) pathogenicity; (ii) host-range (if pathogenic and not opportunistic/other); (iii) presence in five broad environments.

Species	**Gram-stain**	**Pathogenicity**	**Host-range**	**Environments (original Garcia-Garcera & Rocha[1] data)**	**Environments (supplemented with literature)**	**Literature references**
Acinetobacter baumannii	Negative	Opportunistic/ other		Water, wastewater, soil, host	Water, wastewater, sediment, soil, host	[2–5]
Acinetobacter pittii	Negative	Opportunistic/ other			Water, wastewater, sediment, soil, host	[5,6]
Bacillus anthracis	Positive	Pathogen	Broad	Water, soil	Water, soil, host	[7,8]
Bacillus cereus	Positive	Opportunistic/ other		Water, wastewater, soil	Water, wastewater, soil, host	[9,10]
Bacillus subtilis	Positive	Non-pathogen		Soil, host	Soil, host	[11,12]
Bacillus thuringiensis	Positive	Pathogen	Broad	Water, soil	Water, soil, host	[13,14]
Bacillus velezensis	Positive	Non-pathogen			Water, soil, host	[15,16]
Buchnera aphidicola	Negative	Non-pathogen			Host	[17]
Campylobacter coli	Negative	Opportunistic/ other		Host	Host	[18]
Campylobacter jejuni	Negative	Opportunistic/ other		Host	Host	[18]
Chlamydia psittaci	Negative	Pathogen	Broad	Host, sediment	Host	[19,20]
Chlamydia trachomatis	Negative	Pathogen	Narrow	Host, sediment	Host	[21,22]

S2

Citrobacter freundii	Negative	Opportunistic/ other			Water, wastewater, sediment, soil, host	[23]
Clostridium botulinum	Positive	Opportunistic/ other		Water, wastewater, sediment, soil, host	Water, wastewater, sediment, soil, host	[24,25]
Enterobacter cloacae	Negative	Opportunistic/ other		Host	Water, wastewater, sediment, soil, host	[26,27]
Enterobacter hormaechei	Negative	Opportunistic/ other			Water, wastewater, sediment, soil, host	[27,28]
Enterococcus faecalis	Positive	Opportunistic/ other		Host	Host	[29]
Enterococcus faecium	Positive	Opportunistic/ other		Host	Host	[29]
Escherichia coli	Negative	Opportunistic/ other		Water, wastewater, soil, host	Water, wastewater, soil, host	[30,31]
Helicobacter pylori	Negative	Pathogen	Narrow		Host	[32,33]
Klebsiella aerogenes	Negative	Opportunistic/ other		Soil, host	Soil, host	[27,34]
Klebsiella oxytoca	Negative	Opportunistic/ other			Water, wastewater, soil, host	[35]
Klebsiella pneumoniae	Negative	Opportunistic/ other		Soil, host	Water, wastewater, soil, host	[35]
Lactobacillus brevis	Positive	Non-pathogen		Host	Host, wastewater	[36,37]
Lactobacillus paracasei	Positive	Non-pathogen		Host	Host, wastewater	[37]
Lactobacillus plantarum	Positive	Non-pathogen		Soil, Host	Soil, host, wastewater	[37]
Lactobacillus sakei	Positive	Non-pathogen		Host	Host, wastewater	[37]
Lactococcus lactis	Positive	Opportunistic/ other		Host	Host	[38,39]
Legionella pneumophila	Negative	Opportunistic/ other		Water, sediment, soil	Water, sediment, soil, host	[40,41]
Leuconostoc mesenteroides	Positive	Opportunistic/ other		Host	Host	[42]
Listeria monocytogenes	Positive	Opportunistic/ other		Wastewater, soil	Wastewater, soil, host	[43]
Neisseria gonorrhoeae	Negative	Pathogen	Narrow	Host	Host	[44]
Phaeobacter inhibens	Negative	Opportunistic/ other			Host, water	[45]
Piscirickettsia salmonis	Negative	Pathogen	Narrow		Host	[46]
Proteus mirabilis	Negative	Opportunistic/ other		Host	Water, wastewater, soil, host	[47]

S2

Pseudomonas aeruginosa	Negative	Opportunistic/ other		Water, wastewater, soil	Water, wastewater, sediment, soil, host	[48,49]
Pseudomonas syringae	Negative	Pathogen	Broad	Water, soil, host	Water, soil, host	[50–52]
Ralstonia solanacearum	Negative	Pathogen	Broad	Water, soil	Water, wastewater, soil, host	[53,54]
Rhizobium leguminosarum	Negative	Non-pathogen		Soil	Soil, host	[55]
Rhizobium phaseoli	Negative	Non-pathogen			Soil, host	[56]
Salmonella enterica	Negative	Pathogen	Broad	Host	Host, wastewater	[57]
Serratia marcescens	Negative	Opportunistic/ other			Water, wastewater, sediment, soil, host	[58,59]
Sinorhizobium meliloti	Negative	Non-pathogen		Soil, host	Soil, host	[60]
Staphylococcus aureus	Positive	Opportunistic/ other		Sediment, host	Host	[61,62]
Staphylococcus epidermidis	Positive	Opportunistic/ other		Soil, host	Host	[63]
Vibrio parahaemolyticus	Negative	Opportunistic/ other			Water, host	[64]
Xanthomonas citri	Negative	Pathogen	Narrow	Soil, host	Soil, host	[65–67]
Xylella fastidiosa	Negative	Pathogen	Broad	Water, sediment, soil	Water, sediment, soil, host	[68]
Yersinia enterocolitica	Negative	Pathogen	Broad		Water, wastewater, soil, host	[69,70]
Yersinia pestis	Negative	Pathogen	Broad		Host, soil	[71]
Yersinia pseudotuberculosis	Negative	Pathogen	Broad		Host, soil	[72,73]

Table S3 References

1. 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).
2. Fournier, P. E., Richet, H. & Weinstein, R. A. The Epidemiology and Control of Acinetobacter baumannii in Health Care Facilities. *Clin. Infect. Dis.* **42**, 692–699 (2006).
3. Howard, A., O'Donoghue, M., Feeney, A. & Sleator, R. D. Acinetobacter baumannii. *Virulence* **3**, 243–250 (2012).
4. Yang, H., Liang, L., Lin, S. & Jia, S. Isolation and Characterization of a Virulent Bacteriophage AB1 of Acinetobacter baumannii. *BMC Microbiol.* **10**, 131 (2010).
5. Anane A, Y., Apalata, T., Vasaikar, S., Okuthe, G. E. & Songca, S. Prevalence and molecular analysis of multidrug-resistant Acinetobacter baumannii in the extra-hospital environment in Mthatha, South Africa. *Braz. J. Infect. Dis.* **23**, 371–380 (2019).
6. Al Atrouni, A., Joly-Guillou, M.-L., Hamze, M. & Kempf, M. Reservoirs of Non-baumannii Acinetobacter Species. *Front. Microbiol.* **7**, (2016).
7. Spencer, R. C. Bacillus anthracis. *J. Clin. Pathol.* **56**, 182–187 (2003).
8. Koehler, T. M. Bacillus anthracis physiology and genetics. *Mol. Aspects Med.* **30**, 386–396 (2009).
9. Bottone, E. J. Bacillus cereus, a Volatile Human Pathogen. *Clin. Microbiol. Rev.* **23**, 382–398 (2010).
10. Messelhäußer, U. & Ehling-Schulz, M. Bacillus cereus—a Multifaceted Opportunistic Pathogen. *Curr. Clin. Microbiol. Rep.* **5**, 120–125 (2018).
11. Harwood, C. R. Bacillus subtilis and its relatives: molecular biological and industrial workhorses. *Trends Biotechnol.* **10**, 247–256 (1992).
12. Earl, A. M., Losick, R. & Kolter, R. Ecology and genomics of Bacillus subtilis. *Trends Microbiol.* **16**, 269 (2008).
13. Argôlo-Filho, R. C. & Loguercio, L. L. Bacillus thuringiensis Is an Environmental Pathogen and Host-Specificity Has Developed as an Adaptation to Human-Generated Ecological Niches. *Insects* **5**, 62–91 (2013).
14. Garbutt, J., Bonsall, M. B., Wright, D. J. & Raymond, B. Antagonistic competition moderates virulence in Bacillus thuringiensis. *Ecol. Lett.* **14**, 765–772 (2011).
15. Rabbee, M. F. *et al.* Bacillus velezensis: A Valuable Member of Bioactive Molecules within Plant Microbiomes. *Molecules* **24**, (2019).
16. Reva, O. N. *et al.* Genetic, Epigenetic and Phenotypic Diversity of Four Bacillus velezensis Strains Used for Plant Protection or as Probiotics. *Front. Microbiol.* **10**, (2019).
17. Moran, N. A. & Mira, A. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. *Genome Biol.* **2**, research0054.1 (2001).
18. Sheppard, S. K. & Maiden, M. C. J. The Evolution of Campylobacter jejuni and Campylobacter coli. *Cold Spring Harb. Perspect. Biol.* **7**, (2015).
19. Andersen, A. A. Serotyping of US Isolates of Chlamydophila Psittaci from Domestic and Wild Birds. *J. Vet. Diagn. Invest.* **17**, 479–482 (2005).
20. Harkinezhad, T., Geens, T. & Vanrompay, D. Chlamydophila psittaci infections in birds: A review with emphasis on zoonotic consequences. *Vet. Microbiol.* **135**, 68–77 (2009).

21. Elwell, C., Mirrashidi, K. & Engel, J. Chlamydia cell biology and pathogenesis. *Nat. Rev. Microbiol.* **14**, 385–400 (2016).
22. Witkin, S. S., Minis, E., Athanasiou, A., Leizer, J. & Linhares, I. M. Chlamydia trachomatis: the Persistent Pathogen. *Clin. Vaccine Immunol. CVI* **24**, (2017).
23. Ranjan, K. P. & Ranjan, N. Citrobacter: An emerging health care associated urinary pathogen. *Urol. Ann.* **5**, 313–314 (2013).
24. Peck, M. W. Biology and Genomic Analysis of Clostridium botulinum. in *Advances in Microbial Physiology* (ed. Poole, R. K.) vol. 55 183–320 (Academic Press, 2009).
25. Shukla, H. D. & Sharma, S. K. Clostridium botulinum: A Bug with Beauty and Weapon. *Crit. Rev. Microbiol.* **31**, 11–18 (2005).
26. Keller, R., Pedroso, M. Z., Ritchmann, R. & Silva, R. M. Occurrence of Virulence-Associated Properties inEnterobacter cloacae. *Infect. Immun.* **66**, 645–649 (1998).
27. Sanders, W. E. & Sanders, C. C. Enterobacter spp.: pathogens poised to flourish at the turn of the century. *Clin. Microbiol. Rev.* **10**, 220–241 (1997).
28. Wang, Z. *et al.* First report of Enterobacter hormaechei with respiratory disease in calves. *BMC Vet. Res.* **16**, 1 (2020).
29. Byappanahalli, M. N., Nevers, M. B., Korajkic, A., Staley, Z. R. & Harwood, V. J. Enterococci in the Environment. *Microbiol. Mol. Biol. Rev. MMBR* **76**, 685–706 (2012).
30. van Elsas, J. D., Semenov, A. V., Costa, R. & Trevors, J. T. Survival of Escherichia coli in the environment: fundamental and public health aspects. *ISME J.* **5**, 173–183 (2011).
31. Scholz, R. L. & Greenberg, E. P. Sociality in Escherichia coli: Enterochelin Is a Private Good at Low Cell Density and Can Be Shared at High Cell Density. *J. Bacteriol.* **197**, 2122–2128 (2015).
32. Brown, L. M. Helicobacter Pylori : Epidemiology and Routes of Transmission. *Epidemiol. Rev.* **22**, 283–297 (2000).
33. Hooi, J. K. Y. *et al.* Global Prevalence of Helicobacter pylori Infection: Systematic Review and Meta-Analysis. *Gastroenterology* **153**, 420–429 (2017).
34. Wesevich, A. *et al.* Newly Named Klebsiella aerogenes (formerly Enterobacter aerogenes) Is Associated with Poor Clinical Outcomes Relative to Other Enterobacter Species in Patients with Bloodstream Infection. *J. Clin. Microbiol.* **58**, (2020).
35. Bagley, S. T. Habitat association of Klebsiella species. *Infect. Control IC* **6**, 52–58 (1985).
36. Feyereisen, M. *et al.* Comparative genome analysis of the Lactobacillus brevis species. *BMC Genomics* **20**, (2019).
37. Duar, R. M. *et al.* Lifestyles in transition: evolution and natural history of the genus Lactobacillus. *FEMS Microbiol. Rev.* **41**, S27–S48 (2017).
38. Song, A. A.-L., In, L. L. A., Lim, S. H. E. & Rahim, R. A. A review on Lactococcus lactis: from food to factory. *Microb. Cell Factories* **16**, 55 (2017).
39. Aguirre, M. & Collins, M. D. Lactic acid bacteria and human clinical infection. *J. Appl. Bacteriol.* **75**, 95–107 (1993).
40. Newton, H. J., Ang, D. K. Y., van Driel, I. R. & Hartland, E. L. Molecular pathogenesis of infections caused by Legionella pneumophila. *Clin. Microbiol. Rev.* **23**, 274–298 (2010).
41. Steinert, M., Hentschel, U. & Hacker, J. Legionella pneumophila: an aquatic microbe goes astray. *FEMS Microbiol. Rev.* **26**, 149–162 (2002).
42. Bou, G. *et al.* Nosocomial Outbreaks Caused by Leuconostoc mesenteroides subsp. mesenteroides. *Emerg. Infect. Dis.* **14**, 968–971 (2008).

43. Ivanek, R., Gröhn, Y. T. & Wiedmann, M. Listeria monocytogenes in multiple habitats and host populations: review of available data for mathematical modeling. *Foodborne Pathog. Dis.* **3**, 319–336 (2006).
44. Hill, S. A., Masters, T. L. & Wachter, J. Gonorrhea - an evolving disease of the new millennium. *Microb. Cell* **3**, 371–389.
45. Bramucci, A. R. *et al.* The Bacterial Symbiont Phaeobacter inhibens Shapes the Life History of Its Algal Host Emiliania huxleyi. *Front. Mar. Sci.* **5**, (2018).
46. Fryer, J. L. & Hedrick, R. P. Piscirickettsia salmonis: a Gram-negative intracellular bacterial pathogen of fish. *J. Fish Dis.* **26**, 251–262 (2003).
47. Drzewiecka, D. Significance and Roles of Proteus spp. Bacteria in Natural Environments. *Microb. Ecol.* **72**, 741–758 (2016).
48. Pellett, S., Bigley, D. V. & Grimes, D. J. Distribution of Pseudomonas aeruginosa in a riverine ecosystem. *Appl. Environ. Microbiol.* **45**, 328–332 (1983).
49. Sandoz, K. M., Mitzimberg, S. M. & Schuster, M. Social cheating in Pseudomonas aeruginosa quorum sensing. *Proc. Natl. Acad. Sci.* **104**, 15876–15881 (2007).
50. Morris, C. E. *et al.* The life history of the plant pathogen Pseudomonas syringae is linked to the water cycle. *ISME J.* **2**, 321–334 (2008).
51. Rohmer, L., Kjemtrup, S., Marchesini, P. & Dangl, J. L. Nucleotide sequence, functional characterization and evolution of pFKN, a virulence plasmid in Pseudomonas syringae pathovar maculicola. *Mol. Microbiol.* **47**, 1545–1562 (2003).
52. 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).
53. Álvarez, B., López, M. M. & Biosca, E. G. Biocontrol of the Major Plant Pathogen Ralstonia solanacearum in Irrigation Water and Host Plants by Novel Waterborne Lytic Bacteriophages. *Front. Microbiol.* **10**, (2019).
54. Gutarra, L., Herrera, J., Fernandez, E., Kreuze, J. & Lindqvist-Kreuze, H. Diversity, Pathogenicity, and Current Occurrence of Bacterial Wilt Bacterium Ralstonia solanacearum in Peru. *Front. Plant Sci.* **8**, (2017).
55. Labes, G., Ulrich, A. & Lentzsch, P. Influence of Bovine Slurry Deposition on the Structure of Nodulating Rhizobium leguminosarum bv. viciae Soil Populations in a Natural Habitat. *Appl. Environ. Microbiol.* **62**, 1717–1722 (1996).
56. Lowendorf, H. S. & Alexander, M. Identification of Rhizobium phaseoli Strains That Are Tolerant or Sensitive to Soil Acidity. *Appl. Environ. Microbiol.* **45**, 737–742 (1983).
57. Andino, A. & Hanning, I. Salmonella enterica: Survival, Colonization, and Virulence Differences among Serovars. *Sci. World J.* **2015**, (2015).
58. Hejazi, A. & Falkiner, F. R. Serratia marcescens. *J. Med. Microbiol.* **46**, 903–912 (1997).
59. Huang, G. *et al.* Isolation of a Novel Heterotrophic Nitrification–Aerobic Denitrification Bacterium Serratia marcescens CL1502 from Deep-Sea Sediment. *Environ. Eng. Sci.* **34**, 453–459 (2017).
60. Roumiantseva, M. L. *et al.* Diversity of Sinorhizobium meliloti from the Central Asian Alfalfa Gene Center. *Appl. Environ. Microbiol.* **68**, 4694–4697 (2002).
61. Tong, S. Y. C., Davis, J. S., Eichenberger, E., Holland, T. L. & Fowler, V. G. Staphylococcus aureus infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clin. Microbiol. Rev.* **28**, 603–661 (2015).

62. Pollitt, E. J. G., West, S. A., Crusz, S. A., Burton-Chellew, M. N. & Diggle, S. P. Cooperation, Quorum Sensing, and Evolution of Virulence in Staphylococcus aureus. *Infect. Immun.* **82**, 1045–1051 (2014).
63. Otto, M. Staphylococcus epidermidis – the "accidental" pathogen. *Nat. Rev. Microbiol.* **7**, 555–567 (2009).
64. de Souza Santos, M., Salomon, D., Li, P., Krachler, A.-M. & Orth, K. 8 - Vibrio parahaemolyticus virulence determinants. in *The Comprehensive Sourcebook of Bacterial Protein Toxins (Fourth Edition)* (eds. Alouf, J., Ladant, D. & Popoff, M. R.) 230–260 (Academic Press, 2015). doi:10.1016/B978-0-12-800188-2.00008-2.
65. Vieira, G. *et al.* Terrestrial and marine Antarctic fungi extracts active against Xanthomonas citri subsp. citri. *Lett. Appl. Microbiol.* **67**, 64–71 (2018).
66. Patané, J. S. L. *et al.* Origin and diversification of Xanthomonas citri subsp. citri pathotypes revealed by inclusive phylogenomic, dating, and biogeographic analyses. *BMC Genomics* **20**, 700 (2019).
67. 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).
68. Baldi, P. & La Porta, N. Xylella fastidiosa: Host Range and Advance in Molecular Identification Techniques. *Front. Plant Sci.* **8**, (2017).
69. Fredriksson-Ahomaa, M., Stolle, A. & Korkeala, H. Molecular epidemiology of Yersinia enterocolitica infections. *FEMS Immunol. Med. Microbiol.* **47**, 315–329 (2006).
70. Harvey, S., Greenwood, J. R., Pickett, M. J. & Mah, R. A. Recovery of Yersinia enterocolitica from streams and lakes of California. *Appl. Environ. Microbiol.* **32**, 352–354 (1976).
71. Eisen, R. J. *et al.* Persistence of Yersinia pestis in Soil Under Natural Conditions. *Emerg. Infect. Dis.* **14**, 941–943 (2008).
72. Santos-Montañez, J., Benavides-Montaño, J. A., Hinz, A. K. & Vadyvaloo, V. Yersinia pseudotuberculosis IP32953 survives and replicates in trophozoites and persists in cysts of Acanthamoeba castellanii. *FEMS Microbiol. Lett.* **362**, (2015).
73. Gemski, P., Lazere, J. R., Casey, T. & Wohlhieter, J. A. Presence of a virulence-associated plasmid in Yersinia pseudotuberculosis. *Infect. Immun.* **28**, 1044–1047 (1980).

S3 – Supplementary Figures for Genomic Analyses

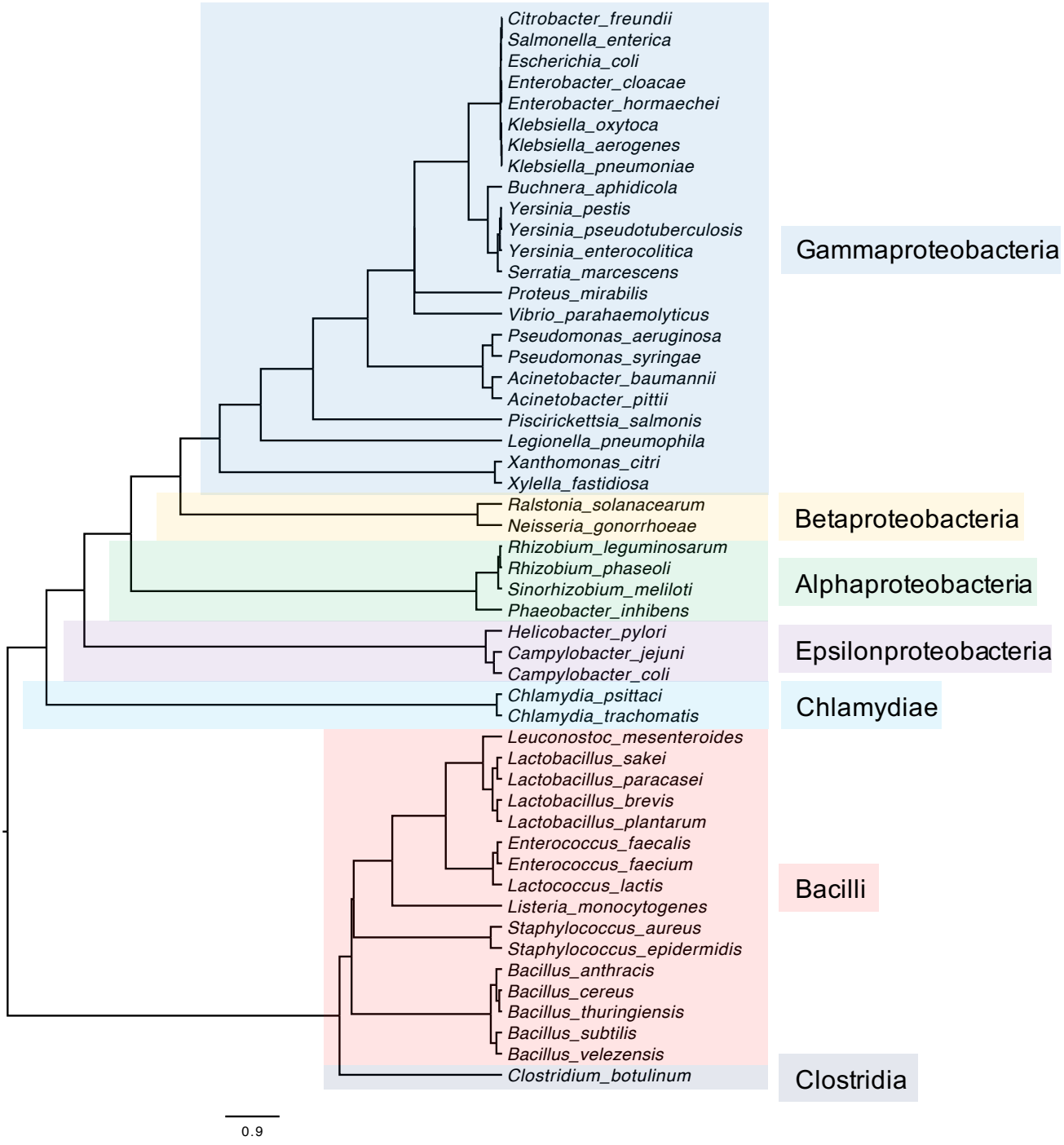


Figure S1. Phylogeny of all 51 species in our dataset.

Based on published 16S RNA maximum likelihood tree[67] and supplemented with additional published trees from the literature. Class is indicated by colour and corresponding labels.

S3

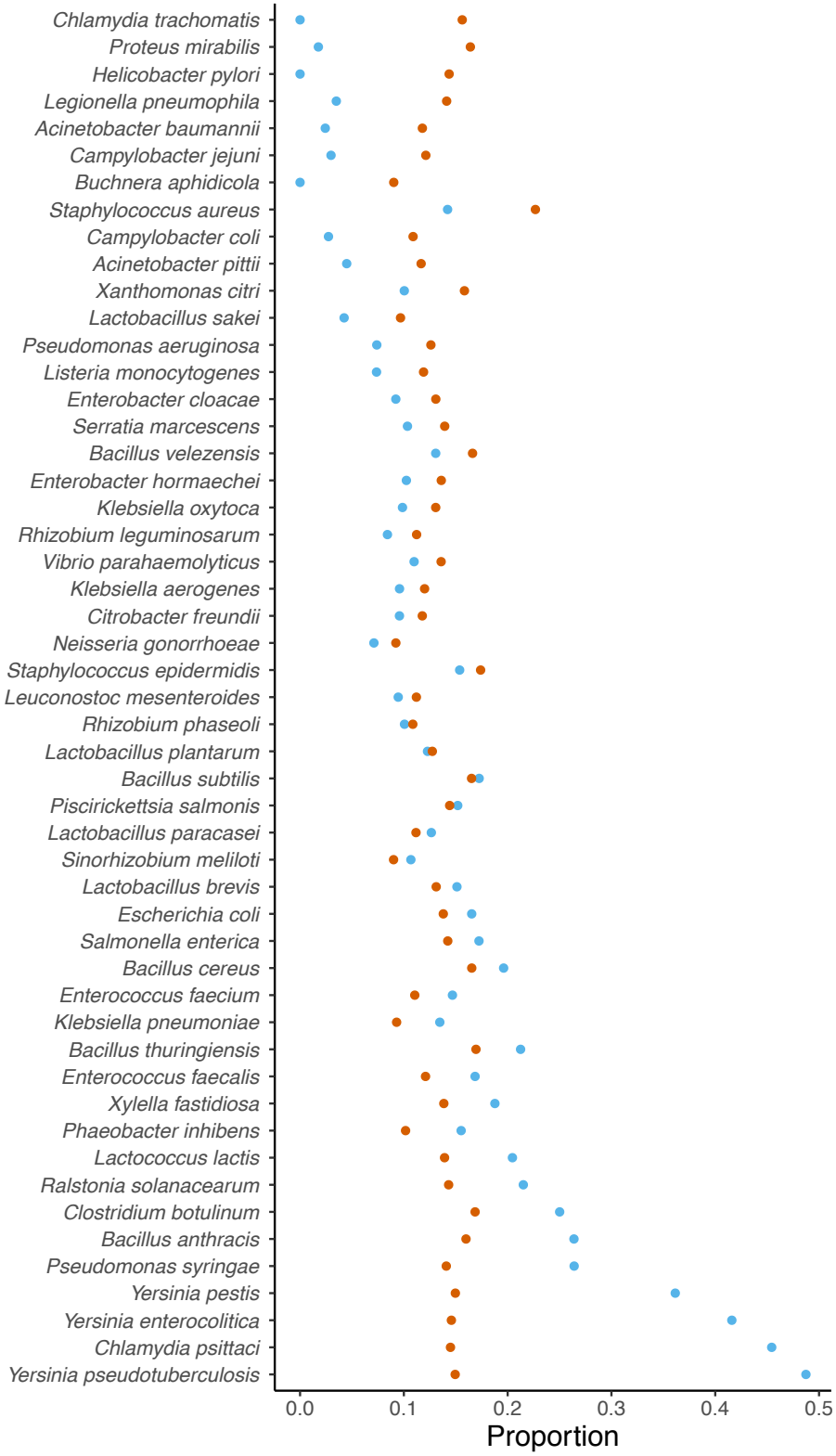


Fig S2. Proportion of proteins predicted as extracellular for plasmids and chromosomes. Each species has two proportions: the blue dot is the mean proportion of plasmid proteins predicted by PSORTb to be extracellular across all plasmids in that species, while the red dot is the mean proportion of plasmid proteins predicted to be extracellular across all chromosomes

in that species. It is clear that these proportions vary substantially across species, and this is particularly true for plasmids. Proportion data is arcsine square root transformed.

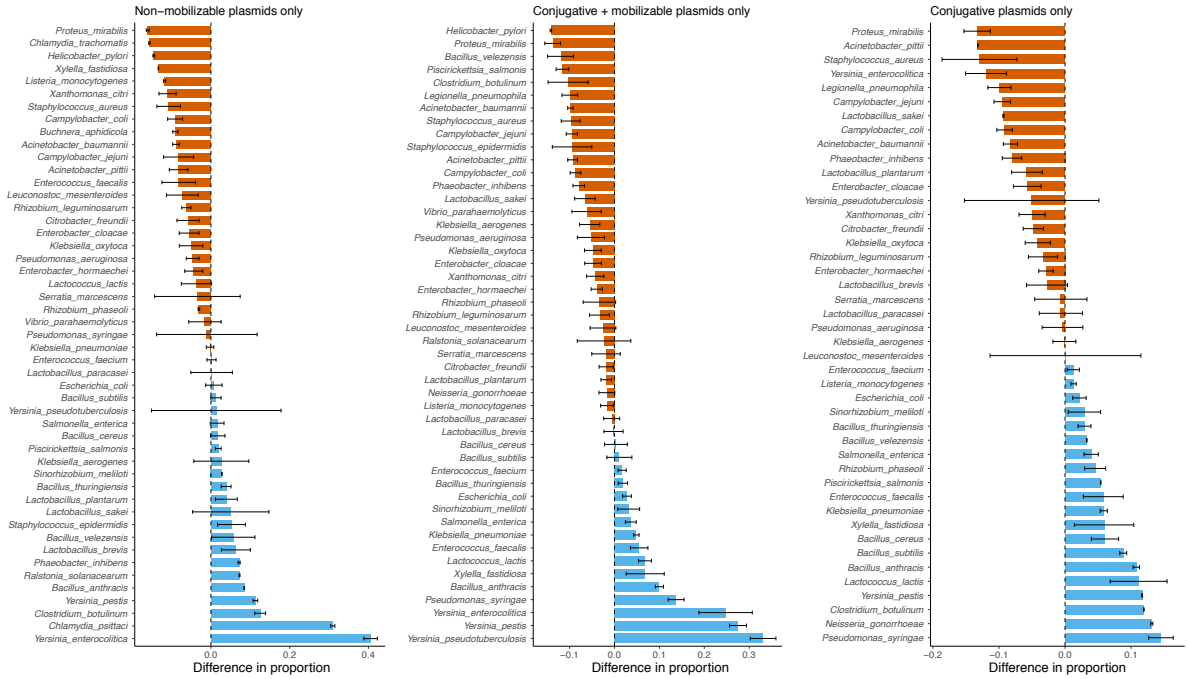


Fig S3. Extracellular proteins are not consistently overrepresented on plasmids of all three mobilities (non-mobilizable, mobilizable, conjugative)

The graphs are identical to Figure 3, but with only certain plasmids included in each. The left-hand graph shows the difference between chromosome and non-mobilizable plasmid proportion of genes encoding extracellular proteins. The middle graph shows the same difference but for conjugative and mobilizable plasmids together. The right-hand graph shows the difference with only conjugative plasmids.

S3

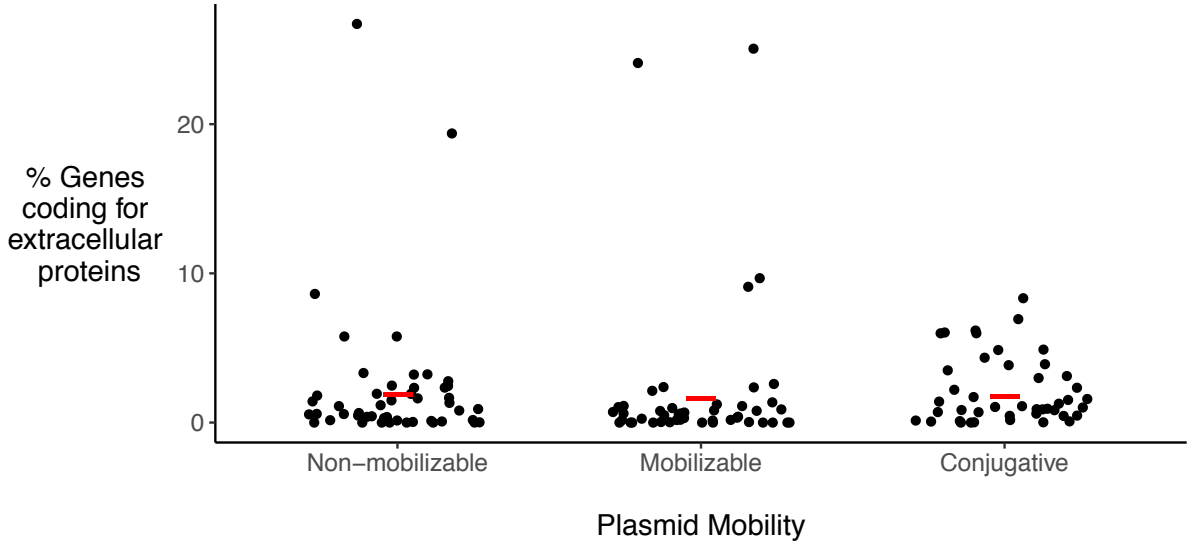


Figure S4. No difference in the mean % of genes coding for extracellular proteins across the three mobility types.

Dots indicate the mean % of genes coding for extracellular proteins of all plasmids of each mobility level for each species. All species data points are shown, including those which do not carry plasmids of all three mobility levels. Red bars indicate the mean across species for each mobility level.

S3

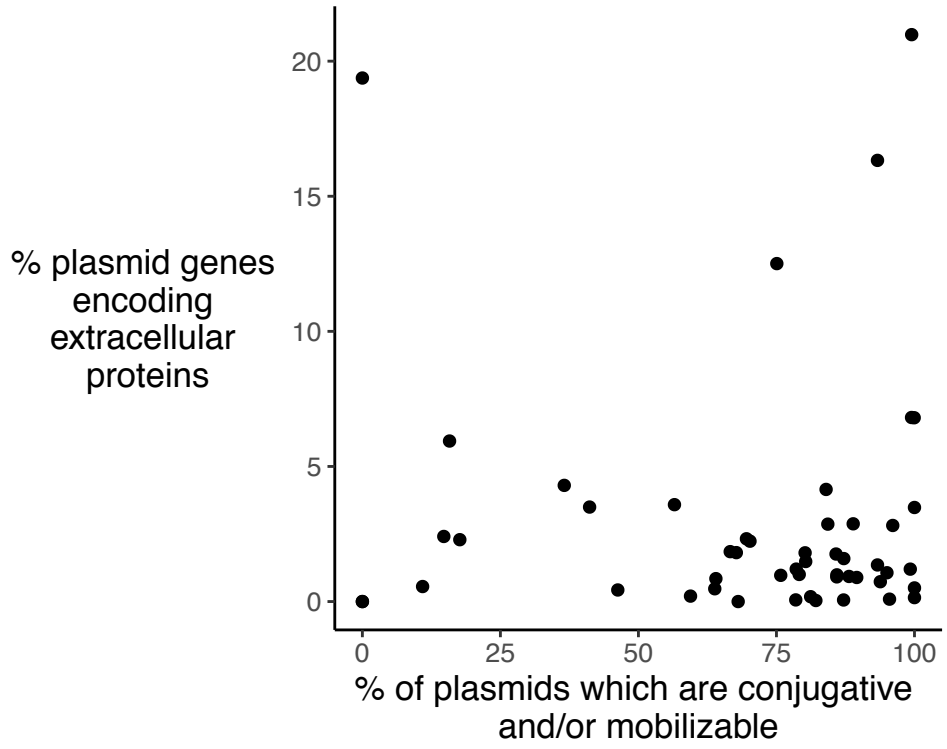


Figure S5. No effect of a species' plasmid mobility and % plasmid genes coding for extracellular proteins.

Dots indicate the mean for each species. The x-axis is the % of a species' plasmids which are conjugative/ mobilizable, and the y-axis indicates the % of a species' plasmid genes which code for extracellular proteins. There is no significant correlation (S3; Table S2, row 19).

S3

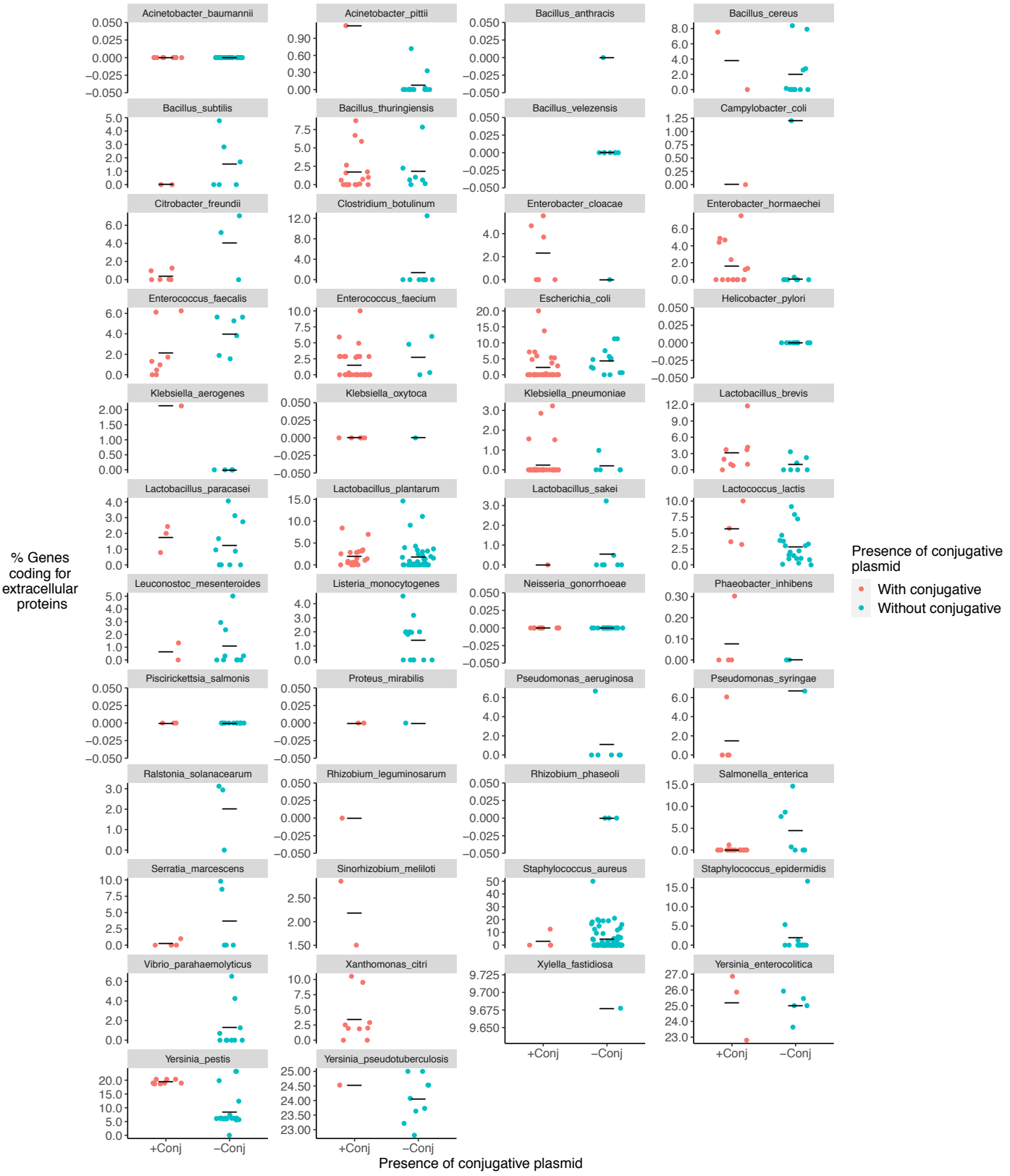


Figure S6. Co-occurrence of mobilizable plasmids with conjugative plasmids.

Each panel shows data for one of the 46 species which had at least one genome with at least one mobilizable plasmid. Each dot corresponds to a genome which had at least one mobilizable

plasmid. The y-axis shows the % of genes coding for extracellular proteins for each genomes' mobilizable plasmid(s). In the cases where two or more mobilizable plasmids were in the same genome, we calculated their mean % and plotted this, so that each genome is only plotted once. Genomes which also carry a conjugative plasmid are plotted on the left of each panel, and coloured red. Genomes which do not carry a conjugative plasmid are on the right of each panel, and coloured green. The black bars indicate the mean of each of these two categories. Overall, species are highly variable in both the number of genomes with mobilizable plasmids that co-occur with conjugative plasmids, and the % of genes that code for extracellular proteins of their mobilizable plasmids. It is clear that, across species, the means of red dots are not consistently greater than the means of blue dots with respect to the y-axis.

S3

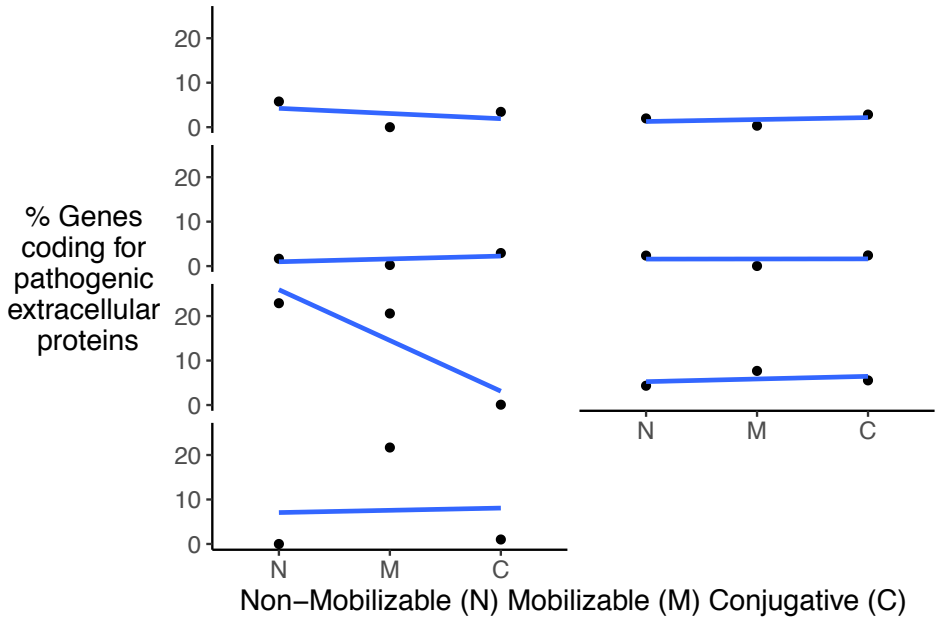


Fig S7. Pathogenic extracellular proteins are not more likely to be carried by higher mobility plasmids in broad host-range pathogen species.

Each panel shows data for one of the 7 broad host-range pathogen species which carried plasmids of all three mobilities. Dots in each panel indicate the mean % of genes coding for pathogenic extracellular proteins of all plasmids of each mobility level. The blue lines are the linear regression of these three points. Overall, there is no consistent trend for genes that code for extracellular proteins to be on more mobile plasmids.

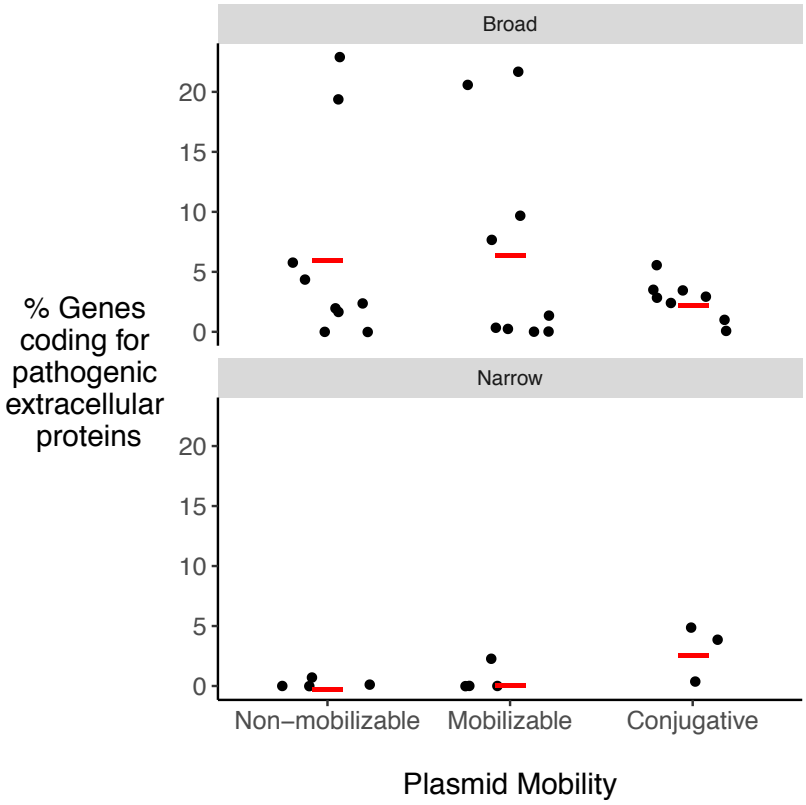


Figure S8. Pathogenic extracellular proteins are not more likely to be carried by more mobile plasmids in both broad and narrow host-range pathogen species

Dots indicate the mean % of genes coding for extracellular proteins of all plasmids of each mobility level for each species. All pathogen species data points are shown, including those which do not carry plasmids of all three mobility levels. Red bars indicate the mean across species for each mobility level.

S3

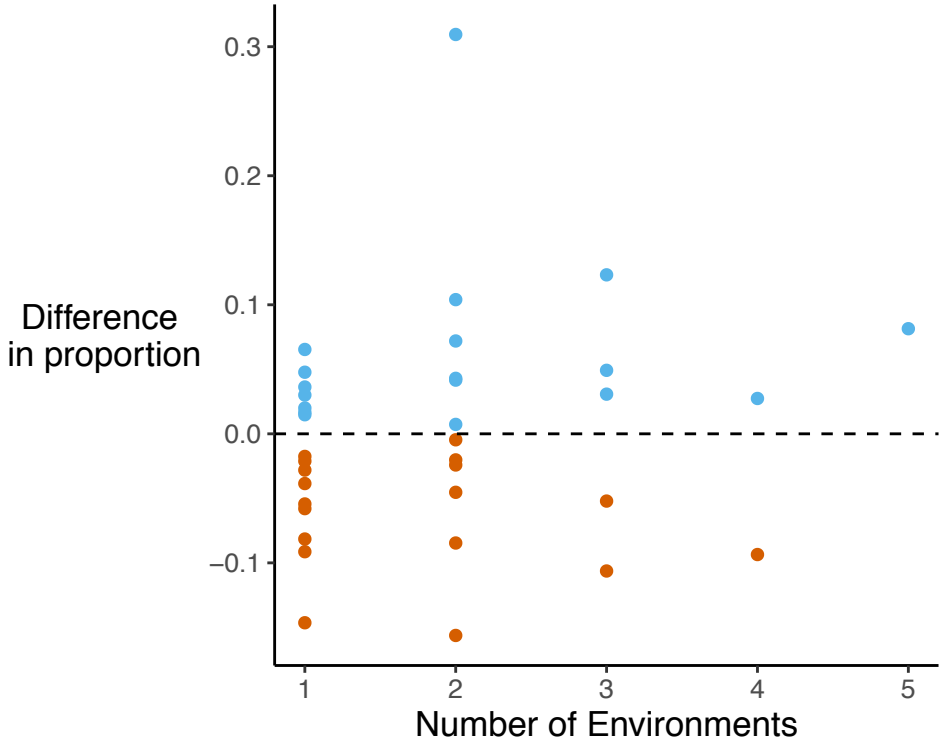


Figure S9. No significant correlation between the number of five broad environments a species is found in and how overrepresented or underrepresented extracellular proteins are on plasmids.

The x-axis shows the original published data of the number of five broad environments a species is found in, with 36 of the species in our dataset represented in the dataset. The y-axis 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 overrepresented on plasmids, while species in red are those with extracellular proteins overrepresented on chromosomes.

S3

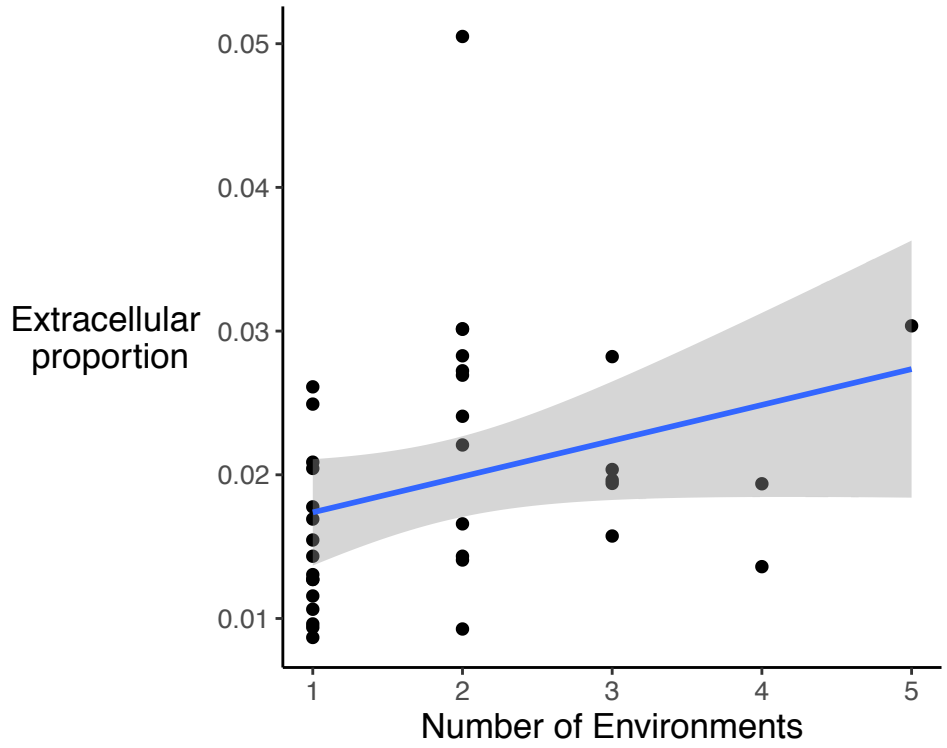


Figure S10. Positive but non-significant correlation between the number of five broad environments a species is found in and the proportion of the genome which encodes extracellular proteins.

The x-axis shows the original published data of the number of five broad environments a species is found in, with 36 of the species in our dataset represented in the dataset. The y-axis shows the proportion of all genes in the genome which code for extracellular proteins. The blue line is the linear regression.

Appendix A

Kin selection for cooperation in natural bacterial populations

Laurence J. Belcher[1*], Anna E. Dewar[1], Melanie Ghoul[1a], Stuart A. West[1a]

[1] Department of Zoology, University of Oxford, Oxford, OX1 3SZ, United Kingdom

a Joint last author

*Corresponding author

Email: laurence.belcher@zoo.ox.ac.uk

Classification: Biological Sciences – Evolution

Keywords: Public goods, inclusive fitness, relatedness, population genetics

Abstract

Bacteria produce a range of molecules that are secreted from the cell, and can provide a benefit to the local population of cells. Laboratory experiments have suggested that these 'public goods' molecules represent a form of cooperation, favoured because they benefit closely related cells (kin selection). However, there is a relative lack of data demonstrating kin selection for cooperation in natural populations of bacteria. We used molecular population genetics to test for signatures of kin selection at the genomic level, in natural populations of the opportunistic pathogen *Pseudomonas aeruginosa*. We found consistent evidence from multiple traits that genes controlling putatively cooperative traits have higher polymorphism, greater divergence, and are more likely to harbour deleterious mutations relative to genes controlling putatively private traits which are expressed at similar rates. We estimate that the relatedness for social interactions in *P. aeruginosa* is $r = 0.84$. Our results suggest that cooperation has been favoured by kin selection, demonstrating how molecular population genetics can be used to study the evolution of cooperation in natural populations.

Significance statement

Bacteria secrete many molecules outside the cell, where they provide benefits to other cells. One potential reason for producing these 'public goods' is that they benefit closely related cells who share the gene for cooperation (kin selection). While many laboratory studies have supported this hypothesis, there is a lack of evidence that kin selection favours cooperation in natural populations. We examined bacterial genomes from the environment and used population genetics theory to analyse the DNA sequences. Our analyses suggest that public goods cooperation has indeed been favoured by kin selection in natural populations.

Introduction

The growth and success of many bacteria appears to depend upon a stunning array of cooperative behaviours (1–3). Cells produce and secrete a range of factors that benefit the local group of cells, and so act as cooperative 'public goods'. Examples include molecules to scavenge iron (siderophores) (4), enzymes that break down proteins (proteases) (5) and molecules to aid cell movement (rhamnolipids) (6).

The potential problem with such cooperation is that it can be exploited by non-cooperators ('cheats') who do not produce public goods, but can still benefit from those produced by others (7). A likely solution to this problem in bacteria is that clonal growth keeps close relatives together, and limited diffusion keeps public goods close to producers (8). Consequently, the benefits of cooperation tend to be shared with related cells that share the gene for cooperation, and so cooperation is favoured by kin selection (9).

However, most evidence for cooperation and kin selection in bacteria has come from laboratory experiments (10–18). To what extent are test tube cultures, often utilising extreme gene knockouts, representative of natural populations? (1, 12). A problem here is that while bacteria and other microorganisms offer many advantages for laboratory experiments, they can be very difficult to study in their natural environment.

Population genetics offers a way to study natural populations, because kin selection can leave signatures ('footprints') of selection at the genomic level (10–12, 15, 19–28). In a clonal population, where the relatedness (r) between interacting cells is r=1, the benefits of cooperating will always be passed onto other individuals who carry the gene for cooperation. In contrast, as relatedness decreases (r<1), the benefits of cooperation will increasingly be passed onto individuals who do not carry the gene for cooperation (Figure 1a). This reduces (dilutes) the kin selected benefit of cooperation, making beneficial mutations less likely to fix, and deleterious mutations more likely to fix (Figure 1b) (9, 25).

Population genetic theory therefore predicts that, in non-clonal populations (r<1), cooperative traits favoured by kin selection will show increased polymorphism and divergence relative to traits that provide private benefits (Figure 1c & d). Non-clonal populations appear to be very common in bacteria. At the scale of the social interaction, groups often contain multiple

species, let alone multiple lineages of the same species (17, 29, 30). In addition, molecular and genomic studies have demonstrated selection for non-cooperative cheats, that exploit the cooperation of others, as well as a diversity of mechanisms for attacking nonrelatives (14, 16, 31). Clonal interactions seem to be limited to extreme cases such as cyanobacteria filaments (30).

We tested for genomic signatures of kin selection for cooperation in the opportunistic pathogen *Pseudomonas aeruginosa*. Laboratory experiments have suggested that *P. aeruginosa* produces a range of cooperative public goods, that facilitate both growth and virulence (4, 32, 33). A potential problem with genomic analyses is that they can be confounded by conditional gene expression. If a gene is only occasionally expressed, in certain conditions, this can also lead to relaxed selection, making beneficial mutations less likely to fix and deleterious mutations more likely to fix (10, 22). We controlled for this influence of conditional gene expression by making targeted comparisons between cooperative and private traits that are likely to be expressed at similar rates.

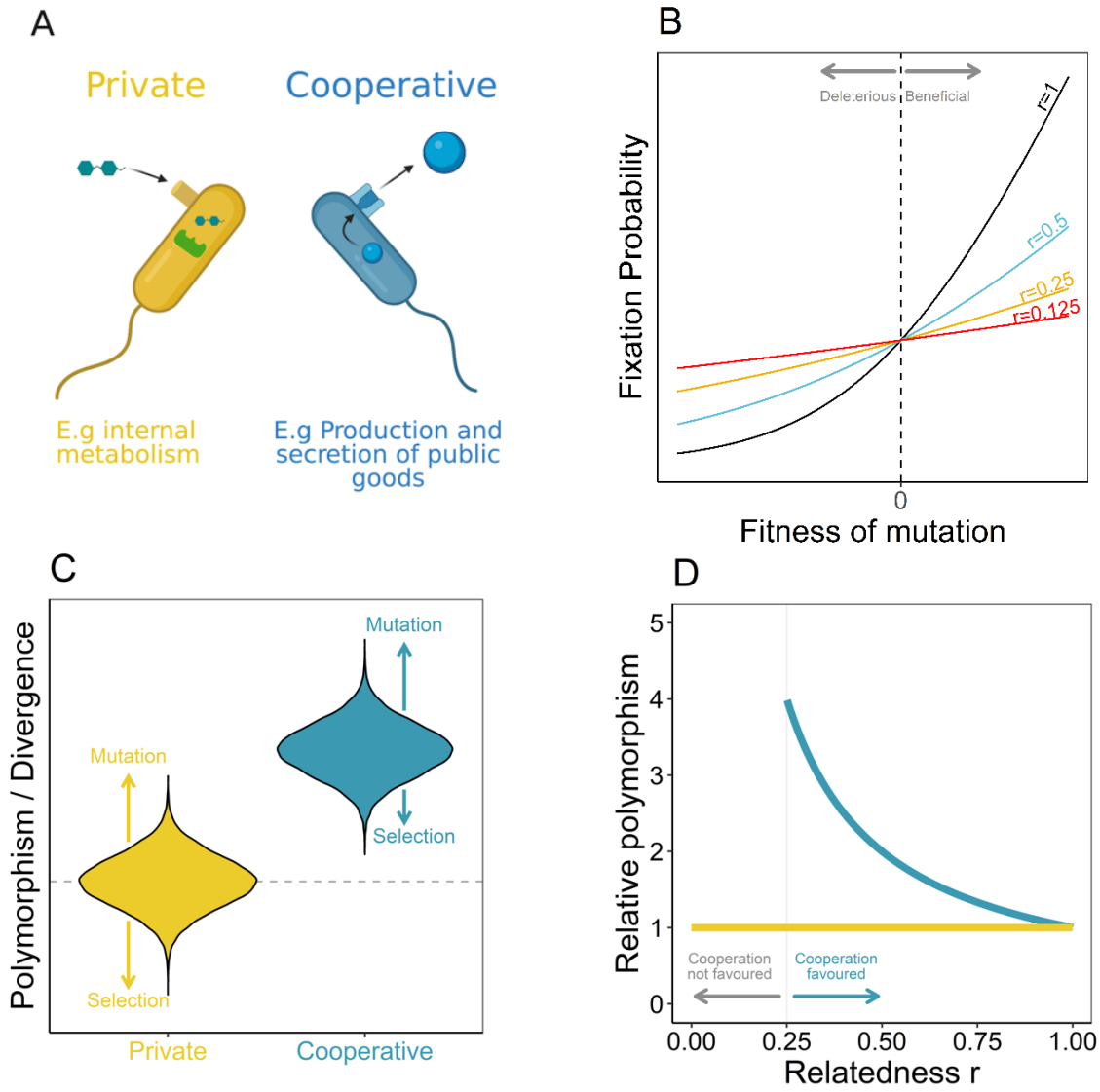


Figure 1: Population genetic theory for cooperative traits. (**A**) Representation of how traits are categorized as private or cooperative. Cooperative traits are those involving the production and secretion of molecules where the fitness benefits can potentially be shared with other cells in the local group. Private traits are those where the fitness benefits are only felt by the individual expressing the gene (e.g. internal metabolism). **B**) Probability of fixation for deleterious or beneficial mutations of varying effect (x-axis) for mutations influencing private (black line) and cooperative (social) (all lines) traits. In clonal populations, where the relatedness (*r*) between interacting individuals *r*=1, the prediction is the same for mutations influencing private and cooperative traits (black line). As relatedness decreases, the prediction changes for mutations influencing cooperation, with beneficial mutations become less likely to fix, and deleterious mutations more likely to fix. Consequently, in non-clonal populations, there is relaxed selection on genes controlling cooperative traits relative to those controlling private traits. Adapted from van Dyken & Wade 2010. (**C**) Prediction for relative polymorphism and divergence for cooperative (blue) relative to private (yellow) genes assuming a fixed r<1. Due to the increased fixation likelihood of deleterious mutations, and decreased fixation likelihood of beneficial mutation, genes for cooperative traits should have relatively greater levels of polymorphism and divergence. (**D**) Predicted polymorphism of private (yellow) and cooperative (blue) genes as relatedness varies for a trait where cooperation is favoured when $r > 0.25$. For private traits, polymorphism is independent of relatedness. For cooperative traits, expected polymorphism relative to a private trait is inversely proportional to r when cooperation is favoured. When r=1, there is no difference in polymorphism between cooperative and private traits. When r<0.25 cooperation is not favoured, so relatedness no longer predicts the level of polymorphism observed.

Results and Discussion

We compared genetic variation in traits which are hypothesised to be cooperative with traits that are hypothesised to be private (Figure 1). The predicted results from the population genetic analysis for kin selection and other competing hypotheses are shown in Table 1. As no single measure can separate the different possible forms of selection, it is important to consider all of these measures together. We examined 41 genomes of *P. aeruginosa* environmental isolates, focusing our analyses on six groups of traits where the cooperative and private traits were likely to be expressed at relatively similar rates (Supplementary Table 6).

Table 1: Predicted results from population genetics analysis for four different forms of selection (positive/directional selection, kin-selection, balancing selection, and purifying selection). Levels of divergence, polymorphism, frequency of deleterious mutations are shown as values for cooperative genes relative to private genes. Tajima's D uses information about the frequency of polymorphism, and predictions are shown as the absolute value, with extreme values indicative of positive or balancing selection. Mcdonald-Kreitman compares levels of polymorphism and divergence, and a significant result is indicative of either positive or balancing selection.

Selection type	Divergence	Polymorphism	Deleterious mutations	Tajima's D	Mcdonald-Krietman
Positive	High	Low	-	$<<0$	$p<0.05$
Kin-selection	High	High	High	≈ 0	n.s
Balancing	Low	High	-	$>>0$	$p<0.05$
Purifying	Low	Low	-	≈ 0	n.s

Quorum Sensing

We started by examining genes induced by the quorum sensing (QS) signalling system (34, 35). This system regulates gene expression in response to the density of a diffusible signal molecule produced by cells. As cell density increases, the concentration of the signal molecule also increases, leading to the upregulation of many genes. In *P. aeruginosa*, the quorum sensing network regulates several hundred genes, which comprise approximately 6% of the genome (36).

There are four advantages to examining the quorum sensing system. First, it regulates a number of traits which are hypothesised to be cooperative, as well as a number of traits that have only

private benefits (Figure 1a) (37, 38). For example, the secretion of enzymes to digest proteins outside the cell (cooperative), versus the production of enzymes to metabolise molecules within the cell (private). Second, control by the shared quorum sensing network means that the genes coding for these different traits are likely to be expressed at relatively similar rates on average (34, 35). This allows us to control for the potentially confounding influence that expression rates may have on patterns of genetic variation (22). Third, co-regulation of genes acts as a control for mutations in non-coding regulatory and promoter regions that could affect the production of public goods. Fourth, the large size of the network means that there are sufficient genes for a meaningful comparison (See Methods).

We used a combination of gene annotations and experimental data to assign genes as controlling either cooperative or private traits (See Methods). For example, we categorised the extracellular elastase Las*B* as cooperative, because it has been shown to be an exploitable public good in laboratory experiments (39). We also included several other extracellular proteases controlled by QS signalling, such as PIV and PepB*,* which can provide benefits to the local group of cells and are known virulence factors (40, 41). Private traits include genes encoding proteins such as Nuh, an intracellular enzyme that allows cells to metabolise adenosine within the cell (5). The set of cooperative genes and their function are given in Supplementary Table 1. Our set of genes contains some that respond specifically to only one of the two major QS signals, so we checked the robustness of our results by restricting the analysis to only genes that respond to both QS signals in Supplement S9.

Quorum Sensing: Polymorphism

We found that genes regulating cooperative traits had significantly higher levels of polymorphism than genes regulating private traits (Figure 2; ANOVA $F_{1,2351}$=12.0, p<0.01. Tukey's HSD p=0.009). This difference was also significant when examining synonymous and non-synonymous sites separately (Synonymous: ANOVA $F_{1,2350}$=30.0, p<10^{-7}; Tukey's HSD p=0.004. Non-synonymous: Kruskal-Wallis $X^2(2) = 22.7, p < 10^{-4}$; Dunn Test p=0.04. Supplementary Figure 1). In all cases, the average pairwise nucleotide diversity per site (π) was significantly higher in cooperative genes compared with private genes. We discuss possible reasons for increased polymorphism being manifest at synonymous sites as well as non-synonymous sites in the following section.

We also found the same pattern of elevated polymorphism in cooperative genes when comparing to a background set of 2459 private genes not involved in the QS system (Supplement S5). This background set was made up of genes whose proteins localise to the cytoplasm, since these are the class of gene least likely to have a cooperative function. However, some cytoplasmic genes will be critical to the process of producing and secreting public goods, particularly in complex public goods such as pyoverdine that require several biosynthesis steps (42). Examining quorum sensing controlled genes, the ratio of non-synonymous to synonymous polymorphism did not differ significantly between genes controlling cooperative versus private traits (ANOVA $F_{1,2338}$=32.4, p<10^{-7}. Tukey's HSD p=0.963). However, QS-regulated private genes had a significantly higher ratio than the background set of private genes (Tukey HSD p<0.03) (Supplementary Figure 2). This result reflects the finding that polymorphism is increased at both non-synonymous and synonymous sites in cooperative compared to private genes, and that QS-regulated genes may be under overall stronger selection than the background set of private genes. This could be because QS-regulated genes include many virulence factors and genes with large fitness effects such as those involved in biofilms, social motility, and obtaining nutrients (38).

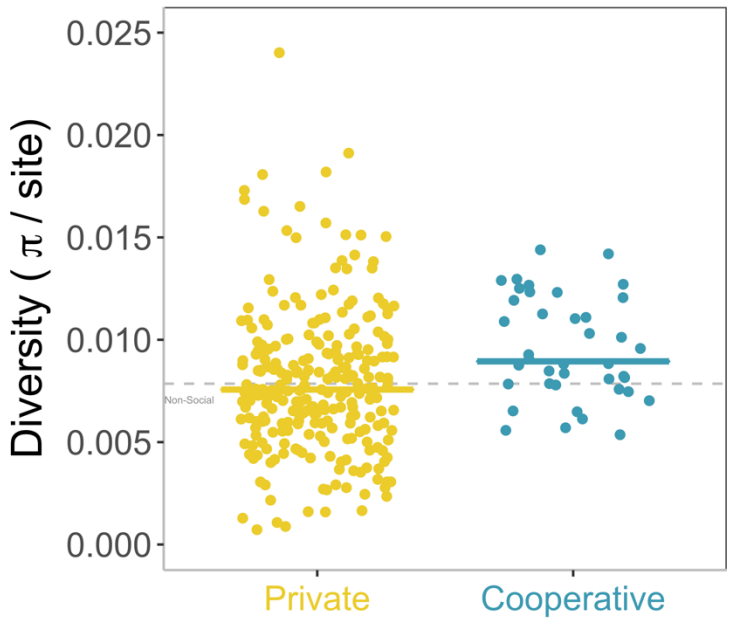


Figure 2: Nucleotide diversity per site for private QS (yellow) and cooperative QS (blue) genes. Each dot represents a gene, and the horizontal line represents the median for each group. The grey dotted line represents the median for private genes across the genome. Genes for cooperative traits showed significantly higher polymorphism than genes for private traits.

Quorum Sensing: Divergence

We found that genes regulating cooperative traits had significantly higher divergence than genes regulating private traits (Figure 3). We measured divergence as the rate of protein evolution, quantified as the number of substitutions per site when comparing the reference genome PAO1 to the known taxonomic outlier PA7 (43). The difference was significant when examining both non-synonymous (Figure 3A; Kruskal-Wallis $X^2(2) = 25.5, p < 10^{-5}$. Dunn Test p=0.045) and synonymous sites (Figure 3B; ANOVA $F_{1,2118}$=0.08, p=0.771. Tukey's HSD p=0.03).

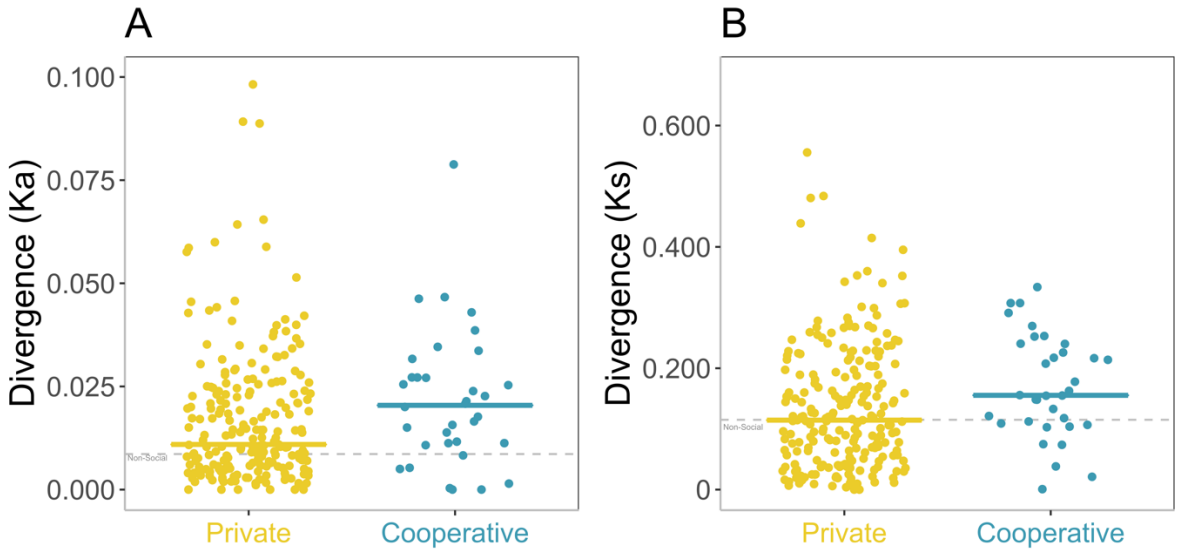


Figure 3: Divergence at non-synonymous (A) and synonymous (B) sites, measured as rates of protein evolution (e.g. non-synonymous substitutions per non-synonymous site) for Private QS (yellow) and Cooperative QS (blue) genes. Each point represents a gene, and the horizontal line represents the median for each group. The grey dotted line represents the median for private genes across the genome. Genes for cooperative traits showed significantly higher divergence than genes for private traits.

Divergence was significantly elevated at both non-synonymous and synonymous sites in cooperative genes, and the ratio of non-synonymous to synonymous divergence does not differ between the two classes of gene (Kruskal-Wallis $X^2(2) = 37.8, p < 10^{-8}$. Dunn Test p=0.40). However, both cooperative and private quorum sensing genes have a significantly higher ratio than the background private genes (Tukey HSD cooperative $p<10^{-2}$, private $p<10^{-6}$), and cooperative genes have slightly higher median ratio than private genes (Supplementary Figure 3). Overall levels of both polymorphism and divergence are consistent with earlier work (see Supplement S8).

Our finding that cooperative genes have significantly elevated polymorphism at both synonymous and non-synonymous sites suggests that mutations at synonymous sites are under

selection, and not evolving neutrally. In microbes, there is substantial evidence that synonymous mutations have fitness effects (44), such as increasing antibiotic resistance (45) and generating public goods cheats in viruses (46). Synonymous mutations in pyoverdine biosynthesis genes repeatedly occur in experimental evolution of *P. aeruginosa* biofilms (47), and synonymous mutations in QS genes of *V. campbellii* are associated with intermediate QS phenotypes (48). Similar patterns of elevated polymorphism at both non-synonymous and synonymous sites was also found in the social microbe *D. discoideum* (10). We did not find evidence for systematic differences in codon usage that could explain the synonymous variation that we see (Supplement S1).

Quorum Sensing: deleterious mutations

Population genetic theory also predicts that deleterious mutations are more likely be observed in genes controlling cooperative traits which are maintained by kin selection (10, 25). This prediction is a result of relaxed selection making deleterious mutations less likely to be removed by selection. We tested this prediction by looking overrepresentation of a subset of loss-of-function mutations that are easily identifiable. Specifically, we looked for (1) mutations that generate stop codons; (2) frameshift mutations. Our previous designation of cooperative genes was based on searching the literature for QS-regulated genes that have been demonstrated to be cooperative in the lab. Because of this, we don't know how many other 'cooperative' genes there are that were not included in our previous dataset. Therefore, to test whether genes with deleterious mutations were more likely to be cooperative, we needed to use a proxy of cooperative genes that examined all genes in the genome. We used the production of extracellular proteins as a proxy for cooperation, as has been done previously (49, 50), since this can be systematically calculated for the whole genome using the protein subcellular localization prediction tool PSORTb (51).

We found that deleterious mutations were more common in genes controlling the production of extracellular proteins, and which where therefore more likely to be cooperative. Of the 359 genes which have known protein localization and at least one deleterious mutation, 12 code for extracellular proteins (3.3%). Genes coding for extracellular proteins make up 1.6% of all genes with known protein localization, but 3.3% of genes with deleterious mutations, which represents a significant overrepresentation of genes coding for extracellular proteins in genes containing deleterious mutations (binomial test, $p<0.05$). Additionally, this increased to 4.4%

(19/431 mutations) when we counted the total number of deleterious mutations in genes coding for extracellular proteins (rather than number of genes with at least one mutation), suggesting that extracellular proteins are also likely to contain multiple mutations per gene. Interestingly, we observed a particularly high rate of deleterious mutations in LasR, the master regulator of the QS system. Whilst LasR isn't an extracellular protein, LasR mutants are common in generating 'cheaters' in clinical isolates (5, 52), and we show here that they also appear to be common in environmental isolates.

Quorum Sensing: robustness and competing hypotheses.

Our conclusion that kin selection favors cooperation was further supported by five further analyses which eliminated alternative explanations for the patterns that we observed. First, genes for cooperative traits could alternatively have significantly greater polymorphism than genes for private traits if they were more likely to be under balancing selection. For example, due to frequency dependent selection between cooperators and cheats (11, 12, 25, 53, 54). However, we found no evidence that genes for cooperative traits are overrepresented in genes evolving under balancing selection, and no evidence that balancing selection explained the elevated polymorphism we observed (Supplement S3).

Second, genes for cooperative traits could have significantly greater divergence than genes for private traits because they are more likely to be under positive selection and therefore have fixed adaptive differences (24, 25). However, we found no evidence that genes for cooperative traits are overrepresented in genes evolving under positive selection, and no evidence that positive selection explains the elevated divergence we observe (Supplement S4). The population genetic parameters that we analysed are designed to test deviation from neutral expectations, and therefore have various underlying assumptions. Neutral theory (55) is based on the idea that polymorphisms are added by mutation, and their fate is largely determined by drift (56). This means that populations are at mutation-drift equilibrium, and we can make predictions about the level or polymorphism we expect in a population. We can then use tests like Tajima's D or the Mcdonald-Krietman test to test for deviations from the predictions of the standard neutral model. Whilst we cannot completely rule out problems in interpreting these tests due to issues such as selection acting at different sites in subpopulations (see Supplement S14), no alternative hypotheses can explain the patterns we see across multiple sets of isolates, and across multiple traits.

Third, our findings could reflect some other shared aspect of cooperative genes, rather than being cooperative *per se*. We performed a functional annotation of all the QS controlled genes using the eggNOG database (57), which splits genes into functional categories such as 'metabolism', 'cellular processes and signaling', and 'information storage and processing'. We found that whilst genes for cooperative traits are overrepresented in genes annotated as 'metabolism' and underrepresented in genes annotated as 'Information storage and processing', there was no difference in polymorphism between these two functional categories (Supplementary Figure 4). Whilst we did find a difference for divergence (Supplementary Figure 5), it is 'information storage and processing' genes that have higher divergence. Overall, it appears that there is no other shared function of genes for cooperation that explains greater divergence and polymorphism.

Fourth, cooperative genes could appear more polymorphic and divergent than private genes because of differences in gene length. In human genomes at least, shorter genes tend to have higher expression (58) and greater divergence (59) than longer genes. If cooperative genes tend to be much shorter than private genes this could bias our results, even though we control for gene length by using polymorphism measures calculated per site, and control for variation in expression by analysing QS controlled genes which should have similar average expression. However, cooperative genes did not differ in length compared to private genes (t-test $t = 0.448, p = 0.657$). Further, when considering all genes, there is no significant correlation between gene length and polymorphism (Pearson's correlation $t = -0.650, p = 0.516$). We checked the robustness of this analysis by removing the bottom quartile of genes (<188 amino acids) from our analysis, and found that this makes no difference to the qualitative results (Supplement S7).

Fifth, if cooperative and private genes differed in their likelihood of being transferred horizontally over their evolutionary history, that could effect comparisons due to the inherent problems that horizontal gene transfer raises in population genetics (60). We conducted an analysis using pangenome data (Supplement S11), showing that the cooperative genes we used are either part of the core genome, or present in most strains with rare duplications. More generally, recent work has shown that across bacteria cooperative genes are not more likely to be on plasmids (and therefore transferred) than chromosomes, including in *P. aeruginosa* (61).

Appendix A

Other Forms of Cooperation

Our analyses on quorum sensing provided support for cooperation being favoured by kin selection. We then tested the robustness of this conclusion by examining five other cases where we could compare genes for cooperative and private traits that were likely to be expressed at similar rates: (1) iron-scavenging siderophore pyoverdine; (2) iron-scavenging siderophore pyochelin; (3) antimicrobial resistance; (4) toxins; (5) adhesion and movement (Figure 4, Table 2). As each comparison considers traits with the same or similar fitness components, the strength of selection is expected to be similar between the 'private' and 'cooperative' genes, aiding comparisons with theory (25). We have focused on cooperation, because we are examining genes for cooperative traits, but if r<1 then we could also expect selection for conflict and exploitation, as has been examined in the slime mould *Dictyostelium discoideum* (62, 63).

Table 2: Additional comparisons of cooperative vs. private genes. We examined five scenarios. In the first two of these we compared genes for the same trait with either private (uptake) or cooperative (production and export) fitness consequence: pyoverdine and pyochelin. For the other three, we compared genes for traits with similar functions, but where traits varied in the extent to which they were relatively private or relatively cooperative: antimicrobial resistance; toxins and adhesion / movement.

Comparison	Relatively private genes	Relatively cooperative genes
Pyoverdine	Genes involved in the uptake and use of iron-bound pyoverdine in the cell.	Genes involved in the biosynthesis and export of pyoverdine into the extracellular space.
Pyochelin	Genes involved in the uptake and use of iron-bound pyochelin in the cell.	Genes involved in biosynthesis and export of pyochelin into the extracellular milieu.
Antimicrobial resistance	Genes that control efflux pumps, which expel unaltered antibiotics back into the environment, and outer porins, which alter resistance through traits such as membrane stability.	Genes where the antibiotic is modified and all cells in the local population benefit. This includes the production of beta-lactamases and enzymes that de-activate aminoglycoside antibiotic.
Toxins	Genes which control mechanisms to eliminate competitors via direct contact and the injection of toxins, such as the type IV secretion system (T6SS). This may still provide an indirect benefit to other cells, but relatively less than diffusible toxins.	Genes involved in the production of bacteriocins to eliminate competitors such as R and F pyocins, which diffuse through the environment.
Adhesion and movement	Genes which allow cells to stick-to and move across surfaces, such as flagella and pili.	Genes producing extracellular polysaccharides and rhamnolipids that allow cells to stick and move together.

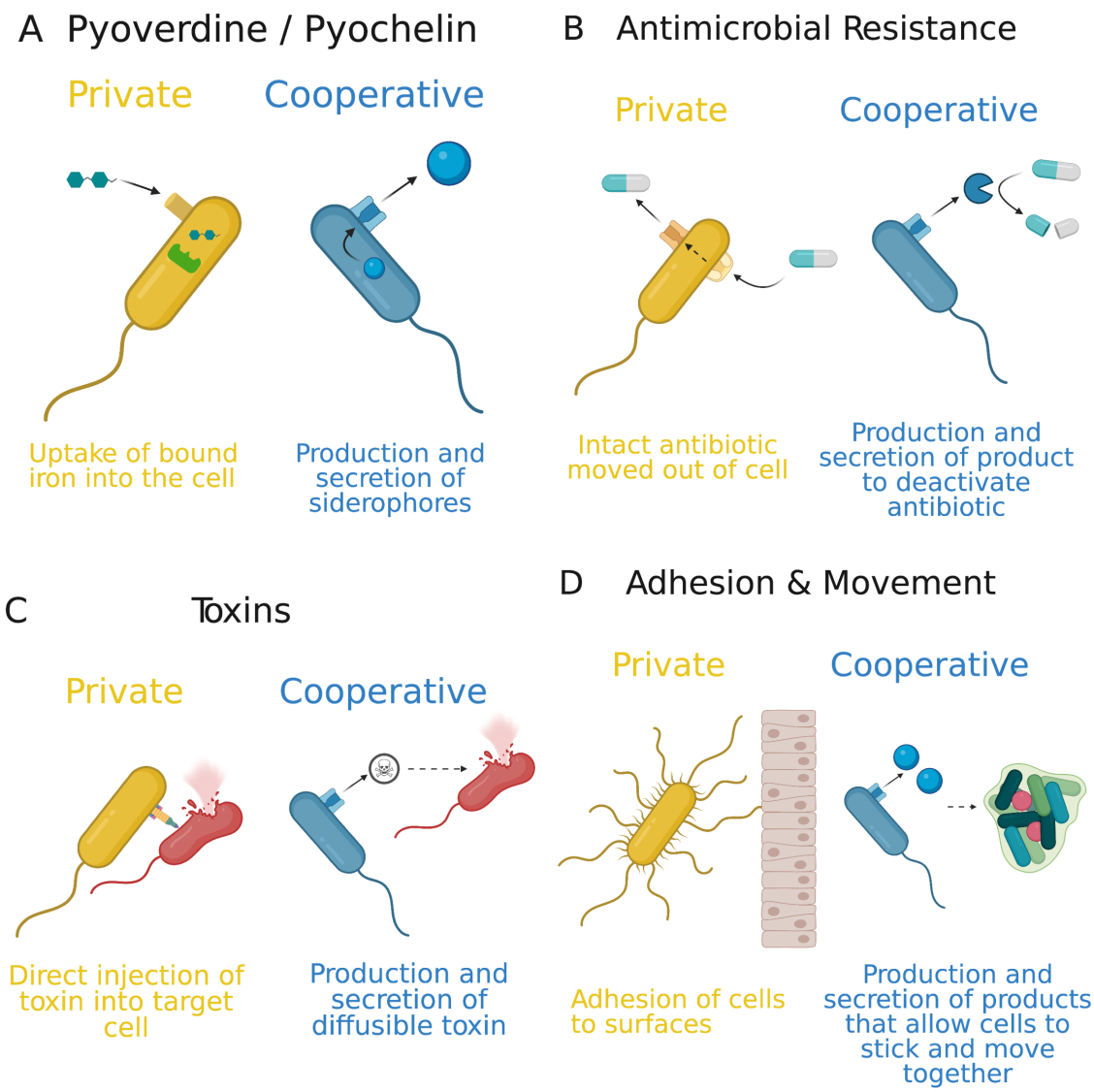


Figure 4: Secondary comparisons of cooperative vs. private traits. (**A**) pyoverdine and pyochelin siderophores. (**B**) Antimicrobial resistance. (**C**) Toxins. (**D**) Cell adhesion and movement. For more details see Table 2 and Methods. Gene lists for each comparison are in Supplementary Tables 2-4

Examining across these different cases, we consistently found that genes coding for relatively cooperative traits were more polymorphic and showed greater divergence than genes coding for relatively private traits. Comparing across all six cases, including quorum sensing, the average level of polymorphism was consistently greater (6/6 cases) in genes coding for cooperative traits (Figure 5; Wilcoxon signed rank exact test, V=21, p=0.03). We found analogous patterns when analyzing synonymous and non-synonymous sites separately (Synonymous: 6/6 cases, Wilcoxon signed rank exact test, V=21, p=0.03 Supplementary

Figure 6; Non-synonymous: 5/6 cases, Wilcoxon signed rank exact test, V=20, p=0.06 Supplementary Figure 7).

Comparing across all six cases, the average level of non-synonymous divergence was consistently greater (6/6 cases) in genes coding for cooperative traits (Figure 6; Wilcoxon signed rank exact test, V=21, p=0.03), with divergence also higher when analyzing synonymous divergence separately (Supplementary Figure 8: 6/6 cases, Wilcoxon signed rank exact test, V=21, p=0.03).

In the above analysis, we examined whether there was a consistent pattern across different types of trait, taking each trait type as a single data point (n=6). One reason that we have taken this relatively conservative approach is that the six traits differ in their power to test between cooperative and private traits. For example, with toxins, adhesion and movement, we are comparing relatively private traits that are likely to still have some cooperative benefit, compared with relatively more cooperative traits (Table 2). With antibiotic resistance, private and cooperative traits can also involve resistance to difference antibiotics (Table 2). Nonetheless, while some of these other five comparisons could have had less power than our analysis of quorum sensing, we found the same consistent pattern across all cases (Figures 5 & 6).

As an alternative analysis, we also combined all genes from all traits into a single data set (n=92 cooperative genes, n=405 private genes). In this case, we also found the same pattern, that genes for cooperative traits showed significantly greater polymorphism and divergence (nucleotide diversity: $t_{175}=3.920$, $p<0.001$; non-synonymous divergence $t_{147}=4.353$, $p<0.0001$) (Supplement S6). We did not analyse the patterns within each type of trait separately, because the sample size in some groups was too low. For example, we were only able to analyse three private pyochelin genes, and four cooperative conflict genes (R and F pyocin bacteriocins).

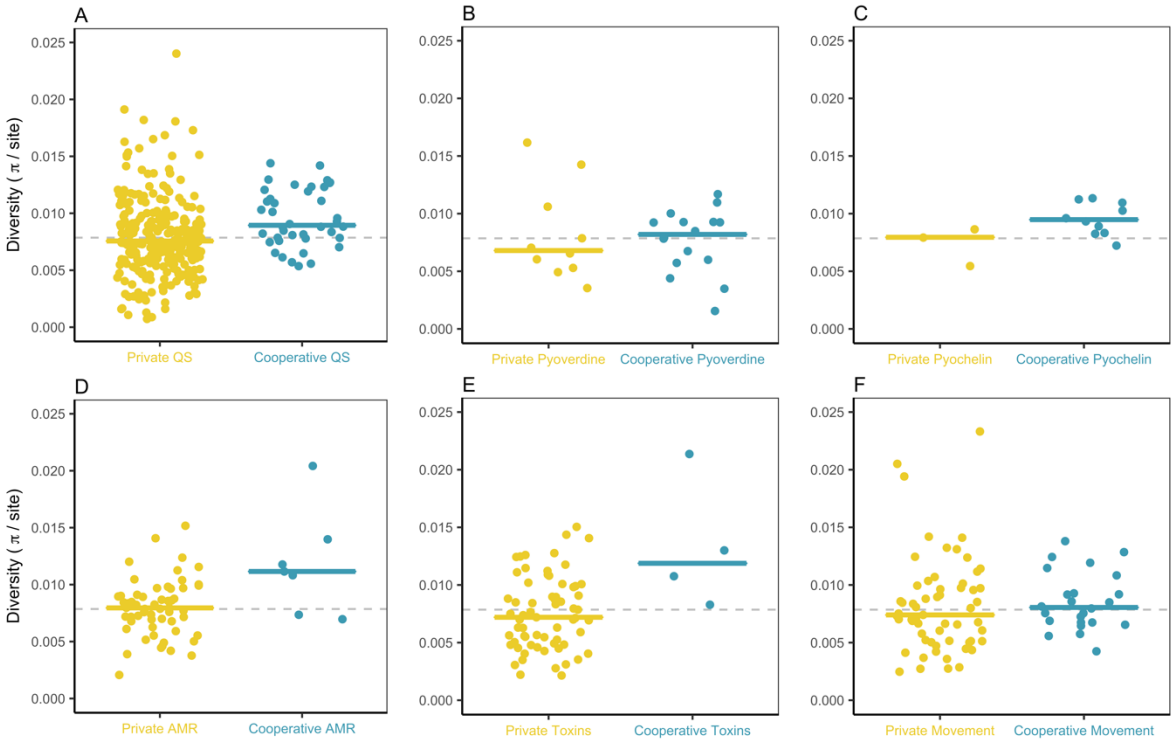


Figure 5: Private versus cooperative comparisons for six trait types for polymorphism (nucleotide diversity). Panel A is the private versus cooperative comparison for quorum sensing genes, from the main analysis (Fig. 2B), shown for comparison. Across different traits, genes for cooperative traits showed a consistent trend towards higher polymorphism than genes for private traits.

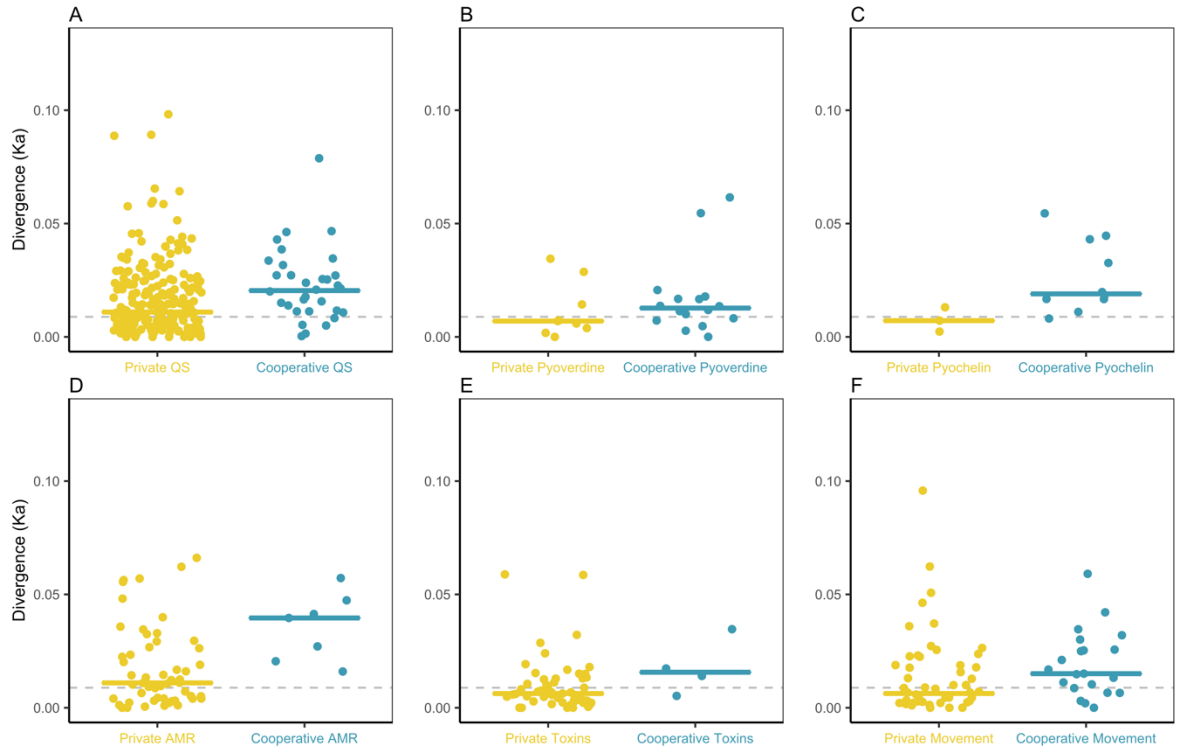


Figure 6: Private versus cooperative comparisons for six trait types for divergence (non-synonymous). Panel A is the private versus cooperative comparison for quorum sensing genes, from the main analysis (Fig. 3), shown for comparison. Across different traits, genes for cooperative traits showed a consistent trend towards higher divergence than genes for private traits.

Clinical Isolates

The robustness of our results was also supported when we analysed whole genomes from 41 clinical isolates. While most clinical strains are often acquired from the environment (64), it is generally thought that they aren't transmitted back to the environment (65). We therefore focused on environmental isolates because they are more likely to represent natural populations. Furthermore, certain environmental conditions such as treatment with antibiotics may affect diversity at some genes (e.g. those involved in immune escape) but not others, so we were decided not to analyze clinical and environmental isolates together. Nonetheless, when analyzing clinical isolates, we found the same qualitative patterns, with genes for cooperative traits showing increased polymorphism consistent with relaxed selection (Supplement S2). Further, our results for polymorphism and divergence are in line with previous studies in this species (Supplement S8).

Appendix A

Relatedness

Given that the predicted degree to which selection is relaxed on cooperative traits is inversely proportional to relatedness between producers and recipients of cooperative traits, we can use our data to estimate relatedness. We do this by comparing the relative level of polymorphism between cooperative and private QS-regulated genes, as we can make direct predictions of relative polymorphism from a simple population genetic model with some assumptions (Supplement S12). In particular, because the theory is about comparing one cooperative gene with one private gene under equal strength of selection, we have to assume that the magnitude and distribution of selection coefficients on cooperative and private traits is on average the same.

We estimate that relatedness is $r = 0.84$ for the natural isolates and $r = 0.85$ for the clinical isolates (Supplement S12). This method allows us to estimate relatedness in natural populations, when it would otherwise be problematic to estimate directly. In order to estimate relatedness directly, it would be necessary to both genotype cells, and identify the spatial scale over which social interactions take place. This is possible in cases where interactions take place in a defined social group such as a fruiting body (18). In contrast, things get much more difficult with public goods, especially as cells live and grow in a range of different environments, and produce a variety of public goods (66). Indeed, laboratory data could even lead to very misleading estimates. In contrast, by using an indirect population genetics approach, we are effectively letting natural selection work out the spatial scale of interaction for us (67). Natural selection will respond to the average relatedness, which will depend upon all the factors that would be hard or impossible for us to directly estimate.

Other Species

Our results build upon previous studies to show how cooperative social behaviours can be favored by kin selection in an analogous way across the natural world. Van Dyken and Wade's (15) groundbreaking analysis on quorum sensing genes across seven bacterial species found similarly increased polymorphism and divergence, but did not have sufficient information at the time to distinguish between private and cooperative quorum sensing controlled traits, or control for expression rates (22). Population genetic analyses on the slime mould *Dictyostelium discoideum* have examined both social conflict and relaxed selection (10–12, 68). In the social

insects, genes for cooperative traits diverge and evolve faster (28, 69), yet experience lower rates of adaptive evolution (19). Furthermore, selection on cooperative worker traits appears to be relaxed with increased mating frequency, when relatedness is lower (20, 21, 26). Social insects have the advantage that genes can be readily separated by gene expression data into worker traits which are presumably cooperative (because workers are largely sterile) and queen traits that are likely to evolve under direct fitness effects (25, 70–72).

Conclusions

Molecular population genetics offers a powerful tool to study how selection acts in natural populations (56). In combination with theory, this type of analysis can determine the extent to which microbial traits are cooperative, and how important this cooperation is in microbes (10–12, 15, 22–25). These results add to the growing evidence that cooperation plays an important role in natural populations of bacteria and other microorganisms. Experiments carried out in hosts have shown that natural populations engage in cooperation (73–76), and can be exploited by non-cooperative cheats (13, 14, 16, 77, 78). Looking across species, comparative studies have found higher levels of cooperation in species where the relatedness between interacting individuals is higher (17, 30). We have shown here that molecular population genetics can also provide evidence for the role of cooperation in natural populations.

Methods

Controlling for levels of expression

The central predictions of elevated divergence and polymorphism are characteristic of relaxed selection, but there are factors alongside kin selection (indirect fitness) that can lead to relaxed selection. Notably, conditional expression can also produce the same effect, via the same mechanism of weakening the association between possessing a genotype and producing a phenotype that can be seen by selection. Specifically, if a gene is expressed by only a fraction of individuals, or by all individuals but in only a fraction of generations, selection is relaxed (22).

In order to control for the effect of conditional expression we restricted our primary analysis to the subset of genes co-induced by the quorum sensing (QS) signalling system. Quorum sensing is a mechanism for coordinating gene expression whereby diffusible signals accumulate as cell density increases, eventually reaching a threshold where the receptor is activated and expression of a set of genes is triggered. In *P. aeruginosa* there are several hundred genes whose expression is controlled by QS signalling, of which there is an overrepresentation of proposed cooperative traits, as well as many private traits (37, 38). We therefore compare cooperative genes to private genes within this set of QS controlled genes, allowing us to control for the effect of conditional expression. In a separate analysis, we assess whether conditional expression itself predicts levels of polymorphism and divergence (Supplement S10).

Categorisation of genes

For the main analysis we focus on genes induced by QS signalling in the *Pseudomonas aeruginosa* reference strain PAO1, for which we use the set of genes described in (36). Within these 315 genes, we selected a set of genes that are putatively 'cooperative' by manually assessing gene function from annotation in the Pseudomonas Genome Database (79), as well as a literature search for any experiments demonstrating a cooperative fitness effect. This was determined by looking for studies that show the basic prediction for a public good (producers outperform non-producers clonally, but non-producers outperform producers in groups). The set of cooperative genes and their function is shown in Supplementary Table 1. This is not

intended to be a fully comprehensive list of genes with any cooperative effect; indeed there are several QS induced genes of unknown/predicted function which are plausibly cooperative, and several categories of genes that may have at least some cooperative component. We compared cooperative QS to private QS genes for the main comparison, and made further comparisons to a background set of private genes in the rest of the genome. For this set of background genes, we used proteins localise to the cytoplasm, as these are the class of gene least likely to have a cooperative function. Such cytoplasmic genes are known to be over-represented with essential genes (80), which suggest an overrepresentation of genes with functions such as central metabolism and replication.

For some analyses where we needed a set of cooperative genes across the whole genome, we followed the approach of previous studies which have used extracellular localization as a proxy for sociality (49, 81). Extracellular localization can be reliably and systematically calculated using PSORTb (51). Whilst it is evident that not all cooperative genes are extracellular and not all extracellular proteins are cooperative, any strong effect of sociality is very likely to be captured by this proxy. For further investigation into properties that may differ between cooperative and private genes, we used eggNOG functional annotations (57).

Secondary comparisons

In our secondary analysis we examined five comparisons of cooperative vs. private genes (Table 2). Firstly, we used pyoverdine, an iron-scavenging siderophore which is extremely well studied for sociality (4, 32, 82). We separated the genes involved in the pyoverdine pathway into cooperative and private components, which is possible thanks to good knowledge of the function and localization of the genes involved (42). We classified genes involved in the biosynthesis and export of pyoverdine into the extracellular milieu as cooperative, and genes involved in the uptake and disassociation of iron-bound pyoverdine in the cell as private (Supplementary Table 2). Pyochelin, the secondary siderophore of *P. aeruginosa*, was separated into cooperative and private components using the same principles, forming our next secondary comparison.

For the two iron-scavenging comparisons we separated a single trait into cooperative and private function, whereas for the other comparisons here we use separate traits for the private vs. cooperative comparison, whilst making effort to ensure that the traits are directly comparable.

Antimicrobial resistance (AMR) is a broad feature which has been well-studied in *P. aeruginosa* for its social fitness effects. There are many ways in which cells can express resistance. One such mechanism is through the production of beta-lactamases which detoxify the environment, and therefore can provide cooperative benefits to the local population (83). Aminoglycoside resistance can also be a cooperative trait, as the antibiotic is modified and therefore the environment is detoxified (84). This is in contrast to efflux pumps, which expel unaltered antibiotics back into the environment (85), and therefore have private fitness effects. Outer porins are another private mechanism (86), which alter resistance through traits such as membrane stability (87). The genes used in this analysis are shown in Supplementary Table 3.

Toxin production is another aspect of bacterial life that can be separated into relatively cooperative and private components. In *P. aeruginosa* there are various mechanisms by which strains compete with and kill each other, which can again be separated into cooperative and private components. Type VI secretion systems (T6SS) involve direct contact with competitors and the use of a needle to inject toxins (88), therefore having a private fitness effect. By contrast, bacteriocins such as R and F pyocins don't require direct contact, and diffuse through the environment (33), which allows cooperative fitness effects on other cells. Elimination of competitors via direct contact can still have a cooperative social benefit, and so our comparison here is between relatively cooperative and relatively private. The gene list for R and F pyocins comes from Ghoul *et al.* 2015 (33). Note that we only use the R and F pyocins and not the S pyocins. R and F pyocins are made up of many genes, which form a structure that resembles a bacteriophage tail (89). S pyocins however consist only of killing and immunity genes (33), and so are less comparable with T6SS. The T6SS gene list comes from the set of genes in the known three distinct T6SS loci in *P. aeruginosa* (88), alongside the *vgr* genes (90) (Supplementary Table 4).

The final cooperative vs. private comparison we used was a broad distinction between extracellular polysaccharides and rhamnolipids that allow cells to stick and move together and are presumed cooperative, and flagella and pili that allow cells to stick-to and move across surfaces. For extracellular polysaccharides (EPS) we used the genes for two of the major *P. aeruginosa* polysaccharides PSL and PEL (91), but not the third polysaccharide EPS alginate which is only a major component of EPS production in clinical settings (92). For rhamnolipids, we used the three biosynthesis genes (6), which are known to be a cooperative trait. For flagella,

we used the gene list in Dasgupta *et al.* 2003 (93) . For pili, we used the gene list in the review by Burrows 2012 (94). This category lumps together some different functions, and represents our most tentative grouping (Supplementary Table 5).

We used Paired samples Wilcoxon tests to test if cooperative genes differ significantly from private genes for each population genetic parameter, with the cooperative and private comparison for each trait type forming a pair. We chose the non-parametric form of a paired t-test because the sample size is quite low for the cooperative genes in some comparisons, so differences were rarely normally distributed and means were strongly effected by extreme values. We calculated two-sided p-values using the wilcox.test function in R.

Sequences

P. aeruginosa is an opportunistic pathogen, with most clinical strains also widely spread environmentally (64). To avoid complications from the selection faced in clinical settings, we focused our primary analysis on environmental isolates. It is generally thought that whilst clinical infections are acquired from the environment, clinical isolates generally aren't transmitted back to the environment (65). We chose strains from a list of *Pseudomonas aeruginosa* strains on the Pseudomonas Genome Database accessed at pseudomonas.com (79). We gathered all available meta data on isolation sources and locations, and first filtered for strains for which the raw sequence read data was publicly available (in the form of an SRA archive), and then further filtered for strains which were unambiguously environmental (by first removing any strains for which the meta-data mentioned 'human', 'clinical', or the name of a disease, and then further removing any records for which it wasn't possible to ascertain their source). This gave a list of 96 possible strains at the time of analysis. This strain list had heavy representation of multiple strain collections from the same locality or environment type, so we took a smaller sample of 41 strains by sampling randomly whilst ensuring that no country featured more than five times. We also screened the isolates to make sure no strain had a known mutator element such as *mutS* that could increase diversity and affect comparisons. Whilst one strain had an in-frame deletion mutation in the mismatch repair gene *mutL*, removing this strain makes no difference to our conclusions (Supplement S13). The 41 strains used are shown in Supplementary Table 6.

SNP Calling

We downloaded raw sequencing reads for each of the 41 strains plus the outgroup PA7 (SRA: SRR9418201) from the European Bioinformatics Institute's European Nucleotide archive (www.ebi.ac.uk/ena) – see Supplementary Table 6 for the relevant ID of each sequencing run. We trimmed reads for each strain to remove adapters and low quality reads using Trimmomatic (95). We removed leading and trailing reads with a quality score <3, and also removed reads if average quality in a four base sliding window was below 20. The resulting reads were quality-control checked using FastQC (96).

Next, we mapped reads for each strain, and aligned to the reference strain PAO1 (Accession: SAMN02603714) using BWA (97). We sorted and converted the resulting SAM files to BAM files using SAMtools (98). We then removed PCR duplicates using Picard tools (99).

We called variants on all strains using BCFtools (100), and converted to a VCF file for analysis. Next, we filtered variants to removed INDELs, and further quality filtering conducted using the default settings of the vcfutils python script in SAMtools (98) to filter for minimum mapping quality (=10), minimum read depth (=2), and minimum p-value for strand bias (=0.001).

We used the featureCounts tool in Subread (101) to assess coverage of each gene in each strain, removing any strains with <2 reads in >50% of genes (which in this case was no strains). We used the coverageBed tool in BEDtools (102) to analyse what proportion of each gene's length had been mapped – so that we could adjust per-site population genetic measures to the mapped length of a gene, rather than the length of the gene in the reference genome.

We removed any site in the genome which hadn't been called in >80% of strains. This meant that each site had a call in at least 33 strains. We conducted a brief power analysis by removing 8 strains from the VCF file to ensure that downstream population genetic measures would not substantially altered by this lowering of sample size. After filtering, we had a VCF file with a total of 391,770 SNPs among the 41 environmental strains (not including the outgroup).

Population Genetic Analysis

Appendix A

We conducted the majority of the molecular population genetics analysis using the PopGenome package (version 2.7.5) in R (103). Specifically, we calculated the parameters nucleotide diversity π, Tajima's D, Fu and Li's F*, Fu and Li's D*, Mcdonald Kreitman test, Neutrality index, alpha, and the Direction of Selection statistic using the PopGenome package. For statistic where an outgroup is necessary, we used PA7, a known taxonomic outlier commonly used (43). Where necessary, to obtain per-site measures, we calculated parameters separately for synonymous and non-synonymous sites and scaled to the relevant mapped length. Genes with mapped length <50% were removed from the analysis at this stage – leaving a final set of 5234 genes

We calculated rates of protein evolution (k_a/k_s) by comparison of the reference strain PAO1 to a known taxonomic outlier, PA7 (43). Next, we extracted SNPs for PA7 from the VCF file, and inserted them into the sequence of PAO1 using the 'FastaAlternateReferenceMaker' tool in the GATK suite (104). We compared this pseudo-genome sequence to the sequence of PAO1 using the seqinR package in R (105) to determine k_a, k_s, and k_a/k_s for each gene. Genes which weren't aligned between the two strains return were removed from this analysis.

For some tests, we conducted further analysis by analyzing whether cooperative genes were overrepresented in the subset of genes which had a statistically significant result for a given parameter. Some tests such as Mcdonald-Kreitman are designed to test the null hypothesis for an individual gene. We used various measures that use the same information as the MKT to allow comparisons across genes (e.g. neutrality index, alpha, direction of selection statistic), and we also extracted the set of genes for which the test is significant (meaning an excess of either non-synonymous substitutions or nonsynonymous polymorphism). For statistics that use data on the site frequency distribution (Tajima's D, Fu & Li's D*, Fu & Li's F*), we also extracted the genes with a significant value. For Tajima's D this was conducted using the beta distribution test (106) conducted in the R package Pegas (107). For Fu & Li's D*/F* statistics we used the critical value's from the original paper (108) for n=100 genes to test significance at the α=0.025 level. Although we have many more genes than 100, the critical value for these tests will be proportional to ln(n) so this is a reasonable approximation. After extracting the subset of genes which are significant for a given test, we test for whether cooperative genes (see below) are over- or under-represented in this class using a binomial test.

Appendix A

One signature of relaxed selection on sociality genes is an increase in deleterious mutations, such as those which have large disruptive effects on the function of a gene. For this analysis, we annotated variants with SNPeff (109) and counted mutations that generate premature stop codons. We included INDELs at this stage so that we could also count frameshift mutations.

To test whether cooperative genes are over- or under-represented in a set of genes, it is necessary to use a proxy for cooperative genes, because our designation of cooperative genes is not a systematic genome-wide assignment and so we cannot confidently say if any number is an overrepresentation since we don't know how many 'cooperative' genes there are in the genome. We used extracellular proteins as a proxy for cooperative genes, which has been used several times before (49, 50) and can be systematically calculated for a whole genome using PSORTb (51). Although it is evident that not all cooperative genes are extracellular and not all extracellular proteins are cooperative, if there was a strong signature of sociality captured by measures such as Tajima's D, we would expect to see an effect with this proxy.

Statistical Analysis

We used R (110) for all statistical analyses and graph plotting. For the main analysis comparing cooperative QS genes to private QS genes, we used a background set of genes for comparison, which comprised all genes in the genome localized as cytoplasmic by PSORTb. This created a large set of genes, of which some may of course be cooperative, but are arguably the group least likely to be cooperative.

Where possible, we used an ANOVA to analyze whether there were any significant differences between our three classes of genes (cooperative, private and background). Data were transformed using the Box-Cox transformation (111), which finds a value of λ such that the transformation $\frac{y^{\lambda}-1}{\lambda}$ gives the best approximation of a normal distribution. Transformed variables were checked for normality with the Kolmogorov-Smirnov test. For some variables, the Box-Cox transformation was not appropriate (as the formulation used does not allow zeros) so a transformation of the form $log(y+c)$ was used, where c is a constant. After transformation and checking the assumptions of ANOVA tests, we conducted the omnibus ANOVA in R, and used Tukey's HSD for post-hoc comparisons. Where data transformation was not sufficient to meet the assumptions of an ANOVA, we used the non-parametric

Kruskal-Wallis test, which compared medians in a ranked-order approach. The Dunn test was used for post-hoc comparisons of Kruskal-Wallis tests, and was only performed where the omnibus test was significant.

Figures

Results figures were all produced using the 'ggplot2' package in R (112). Conceptual figures were created with BioRender.com.

Acknowledgements

We thank Jason Wolf, Thomas Scott, Guy Cooper, Louis Bell-Roberts, Chunhui Hao, and three anonymous reviewers for helpful comments on the manuscript. This work was supported by a BBSRC studentship (AED) and the European Research Council (834164: LJB & SAW; SESE: MG).

References

1. J. E. Strassmann, O. M. Gilbert, D. C. Queller, Kin discrimination and cooperation in microbes. Annu. Rev. Microbiol. 65, 349–367 (2011).
2. S. A. West, A. S. Griffin, A. Gardner, S. P. Diggle, Social evolution theory for microorganisms. Nat Rev Microbiol 4, 597–607 (2006).
3. L. McNally, M. Viana, S. P. Brown, Cooperative secretions facilitate host range expansion in bacteria. Nat. Commun. 5 (2014).
4. A. S. Griffin, S. A. West, A. Buckling, Cooperation and competition in pathogenic bacteria. Nature 430, 1024–1027 (2004).
5. K. M. Sandoz, S. M. Mitzimberg, M. Schuster, Social cheating in Pseudomonas aeruginosa quorum sensing. Proc. Natl. Acad. Sci. U. S. A. 104, 15876–15881 (2007).
6. J. B. Xavier, W. Kim, K. R. Foster, A molecular mechanism that stabilizes cooperative secretions in Pseudomonas aeruginosa. Mol. Microbiol. 79, 166–179 (2011).
7. M. Ghoul, A. S. Griffin, S. A. West, Toward an evolutionary definition of cheating. Evolution (N. Y). 68, 318–331 (2014).
8. R. Kümmerli, A. S. Griffin, S. A. West, A. Buckling, F. Harrison, Viscous medium promotes cooperation in the pathogenic bacterium Pseudomonas aeruginosa. Proc. R. Soc. B Biol. Sci. 276, 3531–3538 (2009).
9. W. D. Hamilton, The genetical evolution of social behaviour. I. J. Theor. Biol. 7, 1–16 (1964).
10. J. L. de Oliveira, et al., Conditional expression explains molecular evolution of social genes in a microbe. Nat. Commun. 10, 3284 (2019).
11. S. Noh, K. S. Geist, X. Tian, J. E. Strassmann, D. C. Queller, Genetic signatures of microbial altruism and cheating in social amoebas in the wild. Proc. Natl. Acad. Sci. U. S. A. 115, 3096–3101 (2018).
12. E. A. Ostrowski, et al., Genomic Signatures of Cooperation and Conflict in the Social Amoeba. Curr. Biol. 25, 1661–1665 (2015).
13. J. B. Bruce, G. A. Cooper, H. Chabas, S. A. West, A. S. Griffin, Cheating and resistance to cheating in natural populations of the bacterium Pseudomonas fluorescens. Evolution (N. Y). 71, 2484–2495 (2017).
14. E. Butaite, M. Baumgartner, S. Wyder, R. Kümmerli, Siderophore cheating and cheating resistance shape competition for iron in soil and freshwater Pseudomonas communities. Nat. Commun. 8, 414 (2017).
15. J. D. van Dyken, M. J. Wade, Detecting the molecular signature of social conflict: Theory and a test with bacterial quorum sensing genes. Am. Nat. 179, 436–450 (2012).
16. O. X. Cordero, L. -a. Ventouras, E. F. DeLong, M. F. Polz, Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. Proc. Natl. Acad. Sci. 109, 20059–20064 (2012).
17. C. Simonet, L. McNally, Kin selection explains the evolution of cooperation in the gut microbiota. Proc. Natl. Acad. Sci. (2021) https:/doi.org/10.1073/pnas.2016046118.
18. O. M. Gilbert, K. R. Foster, N. J. Mehdiabadi, J. E. Strassmann, D. C. Queller, High relatedness maintains multicellular cooperation in a social amoeba by controlling cheater mutants. Proc. Natl. Acad. Sci. 104, 8913–8917 (2007).
19. M. R. Warner, A. S. Mikheyev, T. A. Linksvayer, Genomic Signature of Kin Selection in an Ant with Obligately Sterile Workers. Mol. Biol. Evol. 34, 1780–1787 (2017).
20. B. G. Hunt, et al., Relaxed selection is a precursor to the evolution of phenotypic plasticity. Proc. Natl. Acad. Sci. U. S. A. 108, 15936–15941 (2011).
21. B. G. Hunt, et al., Sociality Is Linked to Rates of Protein Evolution in a Highly Social Insect. Mol. Biol. Evol. 27, 497–500 (2010).
22. J. D. Van Dyken, M. J. Wade, The genetic signature of conditional expression. Genetics 184, 557–570 (2010).
23. J. D. Van Dyken, T. A. Linksvayer, M. J. Wade, Kin Selection–Mutation Balance: A Model for the Origin, Maintenance, and Consequences of Social Cheating. Am. Nat. 177, 288–300 (2011).
24. T. A. Linksvayer, M. J. Wade, Theoretical predictions for sociogenomic data: The effects of kin selection and sex-limited expression on the evolution of social insect genomes. Front. Ecol. Evol. 4, 1–10 (2016).

25. T. A. Linksvayer, M. J. Wade, Genes with social effects are expected to harbor more sequence variation within and between species. Evolution (N. Y). 63, 1685–1696 (2009).
26. D. W. Hall, M. A. D. Goodisman, The effects of kin selection on rates of molecular evolution in social insects. Evolution (N. Y). 66, 2080–2093 (2012).
27. D. W. Hall, S. V. Yi, M. A. D. Goodisman, Kin selection, genomics and caste-antagonistic pleiotropy. Biol. Lett. 9, 1–4 (2013).
28. S. Vojvodic, et al., The transcriptomic and evolutionary signature of social interactions regulating honey bee caste development. Ecol. Evol. 5, 4795–4807 (2015).
29. R. Kümmerli, et al., Co-evolutionary dynamics between public good producers and cheats in the bacterium Pseudomonas aeruginosa. J. Evol. Biol. 28, 2264–2274 (2015).
30. M. Ghoul, et al., Bacteriocin-mediated competition in cystic fibrosis lung infections. Proc. R. Soc. B Biol. Sci. 282 (2015).
31. S. Azimi, A. D. Klementiev, M. Whiteley, S. P. Diggle, Bacterial Quorum Sensing during Infection. Annu. Rev. Microbiol. (2020) https:/doi.org/10.1146/annurev-micro-032020-093845.
32. S. T. Rutherford, B. L. Bassler, Bacterial quorum sensing: Its role in virulence and possibilities for its control. Cold Spring Harb. Perspect. Med. 2, 1–25 (2012).
33. M. Schuster, C. P. Lostroh, T. Ogi, E. P. Greenberg, Identification, timing, and signal specificity of Pseudomonas aeruginosa quorum-controlled genes: A transcriptome analysis. J. Bacteriol. 185, 2066–2079 (2003).
34. K. B. Gilbert, T. H. Kim, R. Gupta, E. P. Greenberg, M. Schuster, Global position analysis of the Pseudomonas aeruginosa quorum-sensing transcription factor LasR. Mol. Microbiol. 73, 1072–1085 (2009).
35. M. Schuster, D. J. Sexton1, B. A. Hense, Why quorum sensing controls private goods. Front. Microbiol. 8, 1–16 (2017).
36. R. Chen, E. Déziel, M. C. Groleau, A. L. Schaefer, E. P. Greenberg, Social cheating in a Pseudomonas aeruginosa quorum-sensing variant. Proc. Natl. Acad. Sci. U. S. A. (2019) https:/doi.org/10.1073/pnas.1819801116.
37. J. L. Bradshaw, et al., Pseudomonas aeruginosa Protease IV Exacerbates Pneumococcal Pneumonia and Systemic Disease . mSphere 3, 1–10 (2018).
38. T. Robinson, P. Smith, E. R. Alberts, M. Colussi-pelaez, M. Schuster, crossm Aminopeptidase in the Pseudomonas aeruginosa RpoS Response. 11, 1–20 (2020).
39. M. T. Ringel, T. Brüser, The biosynthesis of pyoverdines. Microb. Cell 5, 424–437 (2018).
40. P. H. Roy, et al., Complete genome sequence of the multiresistant taxonomic outlier Pseudomonas aeruginosa PA7. PLoS One 5, 1–10 (2010).
41. E. Lebeuf-Taylor, N. McCloskey, S. F. Bailey, A. Hinz, R. Kassen, The distribution of fitness effects among synonymous mutations in a gene under selection. bioRxiv, 1–16 (2019).
42. M. P. Zwart, et al., Unraveling the causes of adaptive benefits of synonymous mutations in TEM-1 β-lactamase. Heredity (Edinb). 121, 406–421 (2018).
43. M. Meir, et al., Competition between social cheater viruses is driven by mechanistically different cheating strategies. Sci. Adv. 6 (2020).
44. S. Azimi, et al., Allelic polymorphism shapes community function in evolving Pseudomonas aeruginosa populations. ISME J. (2020) https:/doi.org/10.1038/s41396-020-0652-0.
45. E. L. Bruger, D. J. Synder, V. S. Cooper, C. M. Waters, Quorum sensing provides a molecular mechanism for evolution to tune and maintain investment in cooperation. bioRxiv (2020) https:/doi.org/10.1101/2020.06.29.178467.
46. M. Garcia-Garcera, E. P. C. Rocha, Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. Nat. Commun. 11, 1–11 (2020).
47. T. Nogueira, M. Touchon, E. P. C. Rocha, Rapid Evolution of the Sequences and Gene Repertoires of Secreted Proteins in Bacteria. PLoS One 7, 1–10 (2012).
48. N. Y. Yu, 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).
49. L. R. Hoffman, et al., Pseudomonas aeruginosa lasR mutants are associated with cystic fibrosis lung disease progression. J. Cyst. Fibros. 8, 66–70 (2009).
50. J. Gore, H. Youk, A. van Oudenaarden, Snowdrift game dynamics and facultative cheating in yeast. Nature 459, 253–256 (2009).
51. A. Ross-Gillespie, et al., Frequency Dependence and Cooperation: Theory and a Test with

Bacteria. Am. Nat. 170, 331–342 (2007).
52. M. Kimura, Evolutionary Rate at the Molecular Level. Nature 217, 624–626 (1968).
53. M. W. Hahn, Molecular Population Genetics (OUP USA, 2018).
54. J. Huerta-Cepas, et al., EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47, D309–D314 (2019).
55. A. O. Urrutia, L. D. Hurst, The signature of selection mediated by expression on human genes. Genome Res. 13, 2260–2264 (2003).
56. I. Lopes, G. Altab, P. Raina, J. P. de Magalhães, Gene Size Matters: An Analysis of Gene Length in the Human Genome. Front. Genet. 12 (2021).
57. E. P. C. Rocha, Neutral theory, microbial practice: Challenges in bacterial population genetics. Mol. Biol. Evol. 35, 1338–1347 (2018).
58. A. Dewar, et al., Plasmids facilitate pathogenicity, not cooperation, in bacteria. Nat. Ecol. Evol.
59. J. P. Pirnay, et al., Pseudomonas aeruginosa population structure revisited. PLoS One 4 (2009).
60. E. A. Ozer, et al., The Population Structure of Pseudomonas aeruginosa Is Characterized by Genetic Isolation of exoU+ and exoS+ Lineages. Genome Biol. Evol. 11, 1780–1796 (2019).
61. E. A. Ostrowski, Enforcing Cooperation in the Social Amoebae. Curr. Biol. 29, R474–R484 (2019).
62. C. Tong, et al., Comparative Genomics Identifies Putative Signatures of Sociality in Spiders. Genome Biol. Evol. 12, 122–133 (2020).
63. S. Patalano, et al., Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. Proc. Natl. Acad. Sci. U. S. A. (2015) https:/doi.org/10.1073/pnas.1515937112.
64. P. G. Ferreira, et al., Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes. Genome Biol. 14 (2013).
65. B. A. Taylor, A. Cini, C. D. R. Wyatt, M. Reuter, S. Sumner, The molecular basis of socially mediated phenotypic plasticity in a eusocial paper wasp. Nat. Commun. 12, 1–10 (2021).
66. K. P. Rumbaugh, et al., Quorum Sensing and the Social Evolution of Bacterial Virulence. Curr. Biol. (2009) https:/doi.org/10.1016/j.cub.2009.01.050.
67. E. J. G. Pollitt, S. A. West, S. A. Crusz, M. N. Burton-Chellew, S. P. Diggle, Cooperation, quorum sensing, and evolution of virulence in Staphylococcus aureus. Infect. Immun. (2014) https:/doi.org/10.1128/IAI.01216-13.
68. L. Zhou, L. Slamti, C. Nielsen-LeRoux, D. Lereclus, B. Raymond, The social biology of quorum sensing in a naturalistic host pathogen system. Curr. Biol. (2014) https:/doi.org/10.1016/j.cub.2014.08.049.
69. S. Gu, et al., Competition for iron drives phytopathogen control by natural rhizosphere microbiomes. Nat. Microbiol. (2020) https:/doi.org/10.1038/s41564-020-0719-8.
70. S. B. Andersen, R. L. Marvig, S. Molin, H. K. Johansen, A. S. Griffin, Long-term social dynamics drive loss of function in pathogenic bacteria. Proc. Natl. Acad. Sci. U. S. A. 112, 10756–10761 (2015).
71. J. L. Sachs, M. O. Ehinger, E. L. Simms, Origins of cheating and loss of symbiosis in wild Bradyrhizobium. J. Evol. Biol. (2010) https:/doi.org/10.1111/j.1420-9101.2010.01980.x.
72. R. M. Fisher, C. K. Cornwallis, S. A. West, Group formation, relatedness, and the evolution of multicellularity. Curr. Biol. 23, 1120–1125 (2013).
73. G. L. Winsor, et al., Enhanced annotations and features for comparing thousands of Pseudomonasgenomes in the Pseudomonas genome database. Nucleic Acids Res. (2016) https:/doi.org/10.1093/nar/gkv1227.
74. C. Peng, F. Gao, Protein localization analysis of essential genes in prokaryotes. Sci. Rep. (2014) https:/doi.org/10.1038/srep06001.
75. T. Nogueira, et al., Horizontal Gene Transfer of the Secretome Drives the Evolution of Bacterial Cooperation and Virulence. Curr. Biol. 19, 1683–1691 (2009).
76. S. O'Brien, A. M. Luján, S. Paterson, M. A. Cant, A. Buckling, Adaptation to public goods cheats in Pseudomonas aeruginosa. Proc. R. Soc. B Biol. Sci. 284 (2017).
77. E. Amanatidou, et al., Biofilms facilitate cheating and social exploitation of β-lactam resistance in Escherichia coli. npj Biofilms Microbiomes 5, 1–10 (2019).
78. K. Poole, Aminoglycoside resistance in Pseudomonas aeruginosa. Antimicrob. Agents Chemother. 49, 479–487 (2005).
79. K. Poole, Pseudomonas aeruginosa: Resistance to the max. Front. Microbiol. 2, 1–13 (2011).

80. S. Chevalier, et al., Structure, function and regulation of Pseudomonas aeruginosa porins. FEMS Microbiol. Rev. 41, 698–722 (2017).
81. R. E. W. Hancock, F. S. L. Brinkman, Function of Pseudomonas porins in uptake and efflux. Annu. Rev. Microbiol. 56, 17–38 (2002).
82. J. D. Mougous, et al., A virulence locus of Pseudomonas aeruginosa encodes a protein secretion apparatus. Science (80-.). 312, 1526–1530 (2006).
83. Y. Michel-Briand, C. Baysse, The pyocins of Pseudomonas aeruginosa. Biochimie 84, 499–510 (2002).
84. R. D. Hood, et al., a Toxin to Bacteria. Cell 7, 25–37 (2011).
85. Y. Irie, et al., The Pseudomonas aeruginosa PSL polysaccharide is a social but noncheatable trait in biofilms. MBio (2017) https:/doi.org/10.1128/mBio.00374-17.
86. D. J. Wozniak, et al., Alginate is not a significant component of the extracellular polysaccharide matrix of PA14 and PAO1 Pseudomonas aeruginosa biofilms. Proc. Natl. Acad. Sci. U. S. A. 100, 7907–7912 (2003).
87. N. Dasgupta, et al., A four-tiered transcriptional regulatory circuit controls flagellar biogenesis in Pseudomonas aeruginosa. Mol. Microbiol. (2003) https:/doi.org/10.1046/j.1365-2958.2003.03740.x.
88. L. L. Burrows, Pseudomonas aeruginosa twitching motility: Type IV pili in action. Annu. Rev. Microbiol. 66, 493–520 (2012).
89. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics (2014) https:/doi.org/10.1093/bioinformatics/btu170.
90. S. Andrews, FastQC. Babraham Bioinforma. (2010).
91. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (2009) https:/doi.org/10.1093/bioinformatics/btp324.
92. H. Li, et al., The Sequence Alignment/Map format and SAMtools. Bioinformatics (2009) https:/doi.org/10.1093/bioinformatics/btp352.
93. Broad Institute, Picard toolkit. Broad Institute, GitHub Repos. (2019).
94. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (2011) https:/doi.org/10.1093/bioinformatics/btr509.
95. Y. Liao, G. K. Smyth, W. Shi, FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (2014) https:/doi.org/10.1093/bioinformatics/btt656.
96. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics (2010) https:/doi.org/10.1093/bioinformatics/btq033.
97. B. Pfeifer, U. Wittelsbürger, S. E. Ramos-Onsins, M. J. Lercher, PopGenome: An efficient swiss army knife for population genomic analyses in R. Mol. Biol. Evol. 31, 1929–1936 (2014).
98. A. McKenna, et al., The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. (2010) https:/doi.org/10.1101/gr.107524.110.
99. D. Charif, J. R. Lobry, "SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis" in (2007) https:/doi.org/10.1007/978-3-540-35306-5_10.
100. F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics (1989).
101. E. Paradis, Pegas: An R package for population genetics with an integrated-modular approach. Bioinformatics 26, 419–420 (2010).
102. Y. X. Fu, W. H. Li, Statistical tests of neutrality of mutations. Genetics (1993) https:/doi.org/10.1093/genetics/133.3.693.
103. P. Cingolani, et al., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). (2012) https:/doi.org/10.4161/fly.19695.
104. R, R: A language and environment for statistical computing. R Found. Stat. Comput. Vienna, Austria. (2020).
105. G. E. P. Box, D. R. Cox, An Analysis of Transformations. J. R. Stat. Soc. Ser. B (1964) https:/doi.org/10.1111/j.2517-6161.1964.tb00553.x.
106. H. Wickham, ggplot2 Elegant Graphics for Data Analysis (2016).

Appendix B

Are plasmid-carried genes for cooperation less complex?

Abstract

Many mechanisms are proposed to determine which traits are more likely to be transferred horizontally via plasmids, two of which have received a lot of attention: (i) genes with lower levels of connectivity and (ii) genes coding for cooperative traits, tend to be transferred horizontally and are therefore more likely to be found on plasmids. However, the second mechanism was not supported by a newly empirical study[1], which reminded us that there might be an interplay between the two mechanisms in determining what traits are carried on plasmids. With a comparative analysis across 161 diverse prokaryotes species including both bacteria and archaea, we found that genes on plasmids were less connected than genes on chromosomes, and this finding could also be applied to genes coding for extracellular proteins, which are likely to be cooperative genes. Based on these results, our study then suggested that gene complexity represented by gene connectivity could be a factor restricting the horizontal transfer of cooperative genes to plasmids.

Keywords: Horizontal gene transfer, cooperative trait, the complexity hypothesis, plasmid

Introduction

Plasmids are extra-chromosomal genetic structures that can autonomously replicate and be transferred between cells. Along with phages and integrative conjugative elements (ICEs), plasmids are also the key vectors of horizontal gene transfer (HGT) and indispensable elements for shaping the genome of prokaryotes[2–7]. Due to their ability to move genetic materials between organisms that are not in a parent-offspring relationship, plasmids are recognized to be one of the most important genetic elements in the evolution of prokaryotes[8–12]. This makes plasmids an important research target not only in the fields of bacterial genetics but also in the fields of evolutionary biology[13,14].

The genes carried on plasmids are known to differ from those found on chromosomes. From an evolutionary perspective, plasmids are fundamentally self-interested entities, thus bear genes that promote their own replication and spread[15–17]. These genes are generally 'core' genes responsible for their vertical and horizontal transmission, such as those encoding proteins that direct plasmid replication, partitioning to daughter cells and conjugation[18]. Furthermore, plasmids encode genes that allow them to persist in the face of multiple constraints[19]. Examples include toxin-antitoxin systems, which ensure plasmids maintenance during segregation[20–22]; anti-restriction systems[23,24] and biofilm formation[25,26], which help plasmids bypass the mechanistic barriers to HGT[27]; and Type IV CRISPR-Cas systems, which mediate conflicts between plasmids[28]. On the other hand, plasmids encode 'accessory' genes that are beneficial to their hosts by increasing the range of environmental conditions to which their host can adapt[14,16]. For instance, the accessory functions conferred by plasmids genes include virulence factors[29,30], antimicrobial resistance[31–34], digestion of most classes of carbohydrates[35], undergoing anoxygenic photosynthesis[36], and bacteriocins[37].

One specific case of accessory traits carried on plasmids is the production of cooperative 'public goods', which are protein products released that benefit local cell populations. A number of theoretical models have suggested that cooperative traits are favoured to be on plasmids because it strengthens the cooperation between cells. By allowing cooperative genes to reinfect 'cheats' that evade the production of public goods, and turning them into 'cooperators', the HGT of cooperative traits can help stabilize the cooperation of neighbouring cells[38–40]. This explanation, however, was not supported by a recent across species comparative analysis, which found that cooperative genes were not overrepresented on plasmids[1]. Further analysis of their study showed that when cooperative genes helped promote the pathogen host-range, they were more likely to be on plasmids[1]. This finding suggested that there might be multiple factors that affected whether the cooperative traits are on plasmids or chromosomes. Consequently, we turned to consider the influence of gene complexity on the locations of cooperative genes, which is a relatively well-understood factor that generally determines the tendency of genes to be successfully transferred and retained in evolution.

The complexity hypothesis suggested that genes involved in the complex interconnected cellular process whose products interact with large numbers of other gene products were transferred at lower rates than less connected genes[41,42]. Further studies confirmed an inverse correlation between the rate of horizontal gene transfer and gene connectivity, which is defined as the number of protein-protein interactions (PPIs) of the gene product in the PPI networks[43–45]. This hypothesis can be naturally expanded by hypothesizing that the higher the connectivity of a gene, the more complex it is, and therefore the less likely the gene is found on the plasmid rather than the chromosome (Figure 1). This would explain the absence of core genes from plasmids in general senses, and could also affect the horizontal transfer and location of genes

encoding public goods. If the cooperative genes have higher connectivity than non-cooperative genes, even if they are more likely to exert their cooperative advantages on plasmids, they would still be constrained to be on chromosomes. When the connectivity of two types of genes is roughly the same, the cooperative genes could still be confined to be located on chromosomes. This could happen if the power that restricts the cooperative genes on chromosomes offsets the benefit of keeping them on plasmids. If the connectivity of cooperative genes is lower than that of non-cooperative genes, cooperative traits are more of accessory traits that are less likely to be regulated to remain on chromosomes. In this situation, factors other than gene connectivity might contribute more to the location of cooperative genes. Taken together, a new test that considers the role of gene connectivity in determining the horizontal transfer and location of genes encoding public goods should be conducted.

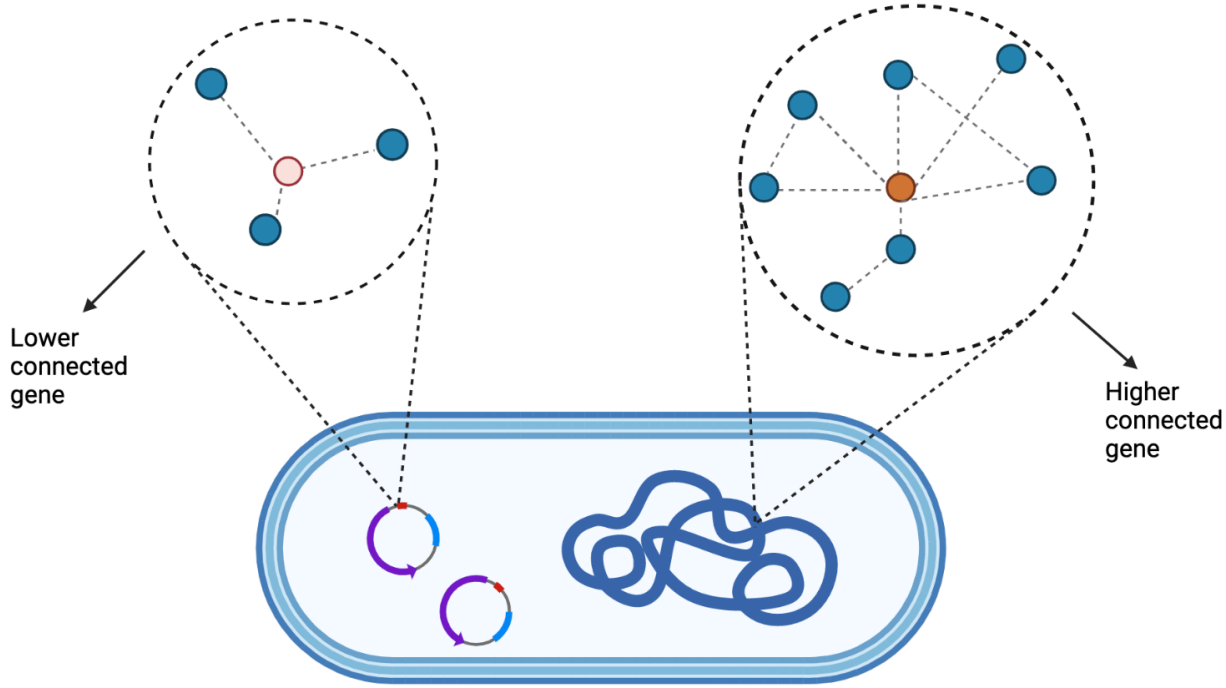


Figure 1. Highly connected genes are more likely to be found on chromosomes. According to the complexity hypothesis, genes with higher connectivity are more likely to be horizontally transferred than genes with lower connectivity. We extended this hypothesis by assuming that the higher the connectivity of a gene, the less likely the gene is found on the plasmid rather than chromosome. Therefore, we hypothesized that chromosomal genes had higher connectivity than plasmid genes.

In this study, we first tested the extension of the complexity hypothesis by comparing the connectivity of genes on chromosomes and genes on plasmids, using 167 genomes from 161 prokaryote species. Unlike previous study[43], we used phylogeny-based statistical methods that controlled the impact of phylogenetic relationships. There are two reasons for doing this. First, species share traits descended from the common ancestor, thus cannot be considered as independent data points[46]. Second, some genes only transferred at intra-genus or intra-species levels, which suggested that phylogenetic relationships might further limit the transfer of lower

connected genes[47]. Therefore, controlling for phylogeny could thus provide us with a fairer understanding of the complexity hypothesis. We then tested whether the effect of the complexity hypothesis constrained cooperative genes from transferring to be on plasmids. We followed the approach of previous studies which have treated genes coding for extracellular proteins as 'cooperative' genes[1,48–50]. To determine whether cooperative genes are different from private genes in the relative gene connectivity between chromosomes and plasmids, we simultaneously examined how sociality (cooperative or private) and location (chromosome or plasmid) affect the connectivity of a given gene.

Results

Chromosomal genes are more complex than plasmids genes.

We first compared the gene connectivity between chromosomal genes and plasmids genes for all genes in our dataset. We found that genes located on chromosomes had significantly higher levels of connectivity compared to genes on plasmids (Figure 2). The difference in connectivity between chromosomal genes and plasmid genes was significantly different from zero across all species, and the connectivity of chromosomal genes was higher than that of plasmid genes (MCMCglmm[51]; posterior mean = 15.135, 95% CI = 12.825 to 18.114, pMCMC < 0.001, n = 161 species, R^2 of phylogeny = 0.202; Figure 2, Table S1).

This result was also robust to alternative analysis when we looked at the ratio of connectivity between chromosomal genes and plasmid genes instead of the difference. The connectivity of chromosomal genes was on average 2.579 times higher than that of plasmid genes (MCMCglmm; posterior mean = 2.579, 95% CI = 2.026 to 3.176, pMCMC < 0.001, n = 161 species, R^2 of phylogeny = 0.341; Table S1). This pattern of elevated connectivity in genes on chromosomes was also significant when we used networks with different confidence thresholds of protein-protein interaction (Table S1). Increasing the threshold reduced the posterior mean of the differences in chromosome and plasmid connectivity, and also reduced the number of species included (Table S1). When we looked at the individual species level, chromosomal genes had higher connectivity than plasmids genes in 98.8% of species (159/161), the opposite pattern was found only in *Beijerinckia indica* and *Ralstonia pickettii* (Figure S1, Table S4).

Network size (the total number of proteins in a PPI network) could affect gene connectivity, which may bias our results. There is evidence that genes with the same connectivity have different impacts in networks of different sizes[52]. To examine whether network size influenced the connectivity of genes in our dataset, we tested whether they were correlated. We found no significant correlation between network size and gene connectivity (MCMCglmm; posterior mean = 0.000352, 95% CI = -0.000212 to 0.000994, pMCMC = 0.274, n = 161; Figure S4, Table S7). We also found no significant correlation between network size and difference of connectivity between the two replicons (Table S7), and also no significant correlation between network size and the ratio of connectivity (Table S7). These results suggested that network size did not affect our finding that the connectivity is increased in genes on chromosomes.

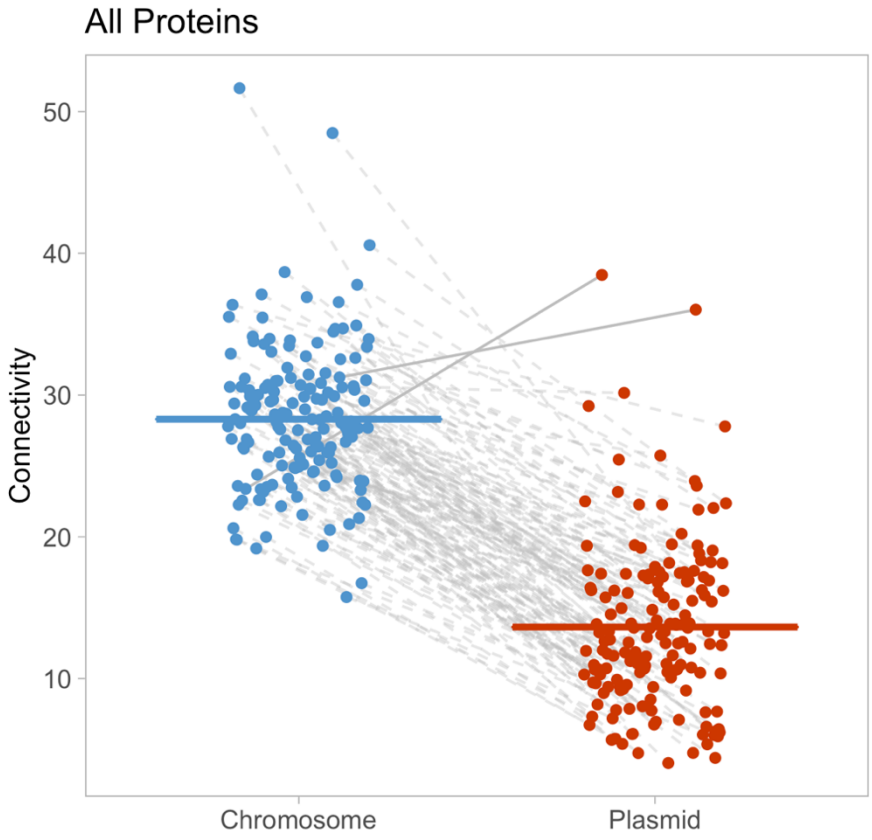


Figure 2. The relative connectivity between chromosomal genes and plasmid genes. Each dot represents the average connectivity of all genes in either the chromosome or plasmid(s) of one species. Chromosome and plasmid values of the same species are linked by a line. A solid line means the average connectivity of chromosomal genes is lower than that of plasmid genes, while a dashed line means the average connectivity of chromosomal genes is higher than that of plasmid genes. The two horizontal lines represent the mean for each group. For almost all species (159/161), chromosomal genes have a higher level of connectivity than plasmid genes.

The prediction of the complexity hypothesis holds for cooperative genes.

To determine whether genes are cooperative genes or private genes would affect their relative gene connectivity between chromosomes and plasmids, we examined the simultaneous influence of two explanatory variables in determining the connectivity of genes. The first explanatory variable is the location of a given gene (chromosome or plasmid), the second one is the sociality of a given gene (cooperative or private). We treated genes encoding extracellular proteins as cooperative genes, and genes encoding intracellular proteins as private genes.

For our main analysis using networks with a threshold of 400 (i.e., The confidence scores of all protein-protein interactions in these networks are higher than 400/1000), we found a significant interaction between two explanatory variables, which suggested that the difference in gene connectivity between chromosomal genes and plasmids genes was different in genes of various sociality (MCMCglmm; posterior mean = -4.470, 95% CI = -8.150 to -0.982,

pMCMC = 0.012, n = 161 species, Table S2). Although the relative gene connectivity between genes on two replicons was affected by the sociality, it was still obvious that genes coding for extracellular proteins (cooperative genes) on chromosomes are more connected than that on plasmids across all species (MCMCglmm; posterior mean = 10.307, 95% CI = 7.853 to 12.750, pMCMC < 0.001, n = 161 species, Figure 3, Figure S7, Table S2). When we looked at the individual species level, we found that genes coding for extracellular proteins on chromosomes had higher connectivity than those on plasmids in 81.4% of species (131/161). In contrast, 18.6% of species (30/161) displayed the opposite pattern, where genes coding for extracellular proteins on plasmids had higher connectivity than that on chromosomes (Figure S2, Table S5).

As a comparison, we also tested whether the prediction of the complexity hypothesis applied to genes that coded for intracellular proteins (private genes). We found that genes encoding intracellular proteins on chromosomes had significantly higher connectivity than that on plasmids across all species (MCMCglmm; posterior mean = 14.759, 95% CI = 13.695 to 15.654, pMCMC < 0.001, n = 161 species; Figure 3, Figure S7, Table S3). And the variations at the individual species level indicated that only 1.86% of species (3/161) displayed the opposite pattern, where genes coding for intracellular proteins on plasmids had higher connectivity than that on chromosomes (Figure S3, Table S6).

When we tested the robustness of this result by performing similar analyses at networks with other confidence thresholds, we found consistent patterns that genes coding for extracellular proteins on chromosomes are more connected than that on plasmids across all networks with different thresholds (Table S2). We also found that the interactions between different explanatory variables varied across networks with different thresholds. A similar significant interaction was found in networks with the lowest level of confidence (150/1000). However, in networks with higher confidence thresholds (700/1000 & 900/1000), the interactions between two explanatory variables were no longer significant, which suggested that the difference in gene connectivity between chromosomal genes and plasmids genes has not changed in genes of different sociality (Table S2).

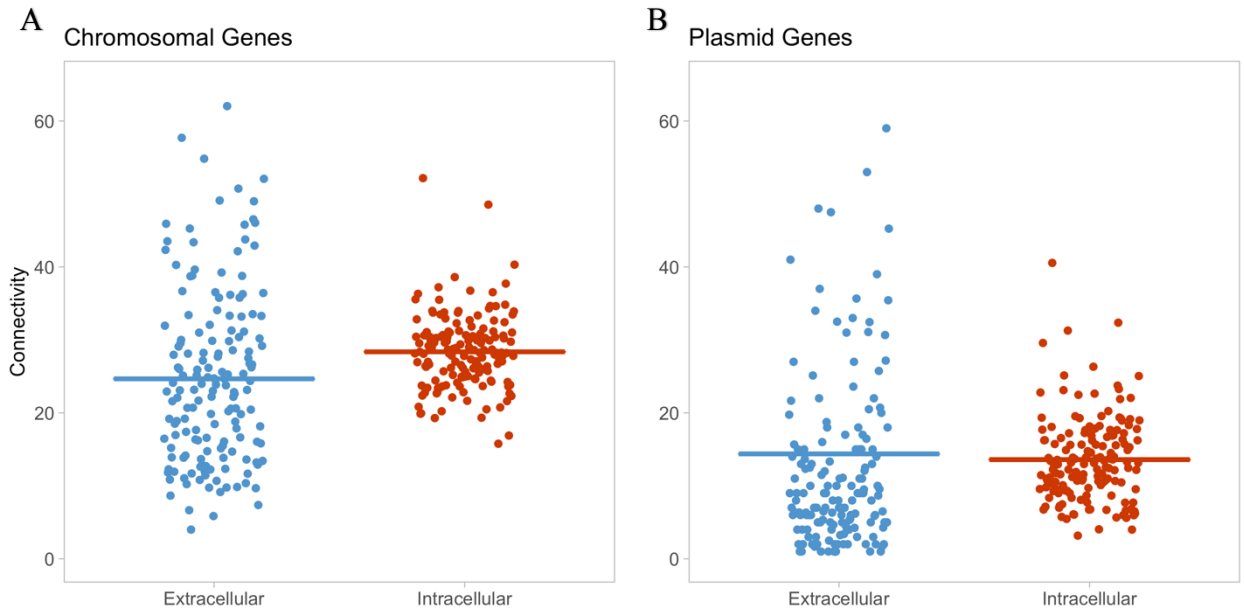


Figure 3. Connectivity of genes encoding extracellular proteins (cooperative) versus connectivity of genes encoding intracellular proteins (private). (A) cooperative versus private comparisons for chromosomal genes; (B) cooperative versus private comparisons for plasmid genes. Each dot represents the mean connectivity of all genes with certain types of protein products for one species. The horizontal line represents the mean for each group. Three species have been removed from figure 2B, because their genes on plasmids have much higher levels of connectivity. The complete version of Figure 2B can be found in the supplementary material (Figure S5). Although there were slight differences in gene connectivity between genes encoding extracellular proteins and genes encoding intracellular proteins, chromosomal genes were still more connected than plasmids genes regardless of their sociality.

Discussion

In this study, we examined the extension of the complexity hypothesis by comparing the connectivity of genes on chromosomes and genes on plasmids. We found strong evidence that genes with higher levels of connectivity were more likely to be on chromosomes rather than plasmids. We then tested whether the impact of gene connectivity constrained gene encoding extracellular proteins (cooperative genes) from being on plasmids. We found that the prediction of our previous analysis could be applied to cooperative genes, because our result suggested that chromosomal genes were more complex than plasmids genes for cooperative genes. These results suggested that gene connectivity could be a factor that restricts cooperative genes from locating on plasmids, even though these genes are more beneficial when on plasmids.

Our finding that genes on chromosomes had higher levels of connectivity was in principle consistent with previous studies[43–45] (Figure 2). The original complexity hypothesis was proposed to explain why informational genes (i.e., those involved in transcription, translation, and related processes) are less likely to be horizontally transferred[41]. Subsequent studies aimed at testing the validity of the complexity hypothesis also focused on genes that are hypothesized to undergo HGT. Based on previous studies, we extended the prediction of the complexity hypothesis by assuming that plasmids are less likely to harbour highly connected genes. The rationality of making such an extension is that plasmids are known to be essential drivers of horizontal transfer in prokaryotic evolution[2,3,12]. Horizontally transferred genes, which are detected by the presence and absence of genes across multiple phylogenetically related genomes, are likely to be transferred by the movement of plasmids. Therefore, it is reasonable to assume that HGT genes would share similar genetic characteristics with genes on plasmids. In addition, our results were robust throughout all networks with different thresholds, which suggested that the influence of gene connectivity in determining whether genes are on chromosomes or plasmids was powerful. Other mobile genetic elements (MGEs) such as integrative conjugative elements (ICEs), phages, and transposable elements are also important agents of HGT and could be the focus of research for testing the complexity hypothesis[53]. Our study, however, only focused on plasmids because they are the most representative and most widely studied MGEs that mediate HGT.

One of the advancements of our study is that we included 161 species in our analysis, including 7 archaeal species, which is significantly larger than previous studies. This allowed us as universally as possible to examine the applicability of the complexity hypothesis in determining the gene location of both bacteria and archaea. Another advancement is that we controlled for the misleading effects of phylogenetic relationships among all species in our analysis. Due to shared ancestry, species are not independent from each other, the patterns displayed across multiple species are likely to be the result of evolutionary history, not the result of evolutionary mechanisms[46,54,55]. Additionally, there was evidence that 16S rRNA genes, which are important informational genes, only transferred at intra-genus or intra-species levels[47]. 16S rRNA genes are not necessarily to be complex. If they are not compatible in distant organisms because of their conservation, they are also likely to be transferred less or only in closely related species[56]. Consequently, by accounting for phylogeny, we were closer to delineate the effect of gene connectivity in determining gene location.

We followed the previous studies that considered gene connectivity as a measurement of gene complexity. Although there was evidence that gene connectivity is a good prediction of gene essentiality[57–59], we did not blur these two related but different concepts in the context of our study. A gene is considered essential if it is required for the reproductive success of a cell or an organism[60]. However, a previous test of the complexity hypothesis has shown that gene

connectivity rather than functional essentiality acted as a more important barrier in restraining HGT[42]. Subsequent research further claimed that gene expression level was a major determinant of horizontal gene transferability for some species[61]. In addition, other network-based measurements such as betweenness centrality and hierarchy, have been suggested to be more significant indicators of gene essentiality in some cases[62–64]. Therefore, in our study of disentangling the factors regulating HGT, we focused more on the extent to which gene connectivity deciphered gene complexity. Our result thus cannot be interpreted as chromosomal genes are more essential than plasmid genes.

Our next analysis suggested that gene connectivity could explain why in some species, cooperative genes are more likely to be on chromosomes instead of plasmids. We found that cooperative genes were more connected on chromosomes than on plasmids across all species (Figure 3). Therefore, we claimed that highly connected cooperative genes could be confined to stay on chromosomes, even though they could provide more cooperative benefits when on plasmids. However, the influence of gene connectivity on determining the location of cooperative genes was not as strong as that for private genes. For private genes, genes on chromosomes are more connected than genes on plasmids in 98.14% of species (158/161). In contrast, for cooperative genes, 81.4% of species (131/161) displayed the same pattern, which was less than that for private genes. Although in networks with a confidence threshold of 400 or lower, the interaction between sociality and location in affecting gene connectivity was significant, the p-value (pMCMC = 0.012) was only slightly less than 0.05. This suggested that despite the relative gene connectivity between chromosomal genes and plasmids genes was different in genes with different sociality, such disparity was not meaningful, which was in line with our finding when looking at the individual level.

We found that the influence of sociality on gene connectivity between chromosomal genes and plasmids genes varied across networks with different thresholds. When we increased the thresholds of networks we used (700 & 900), the interactions between two variables were no longer significant. The threshold of a network displays the confidence cutoff of functional associations between every two proteins in the network (see **Methods**). Considering the evidence of protein-protein interactions in our dataset are from diverse sources, and it is difficult to evaluate the reliability of interactions derived from computational inferences, the threshold does not perfectly represent the real confidence of a given interaction[65]. Nonetheless, we were still able to conclude that as we scrutinized the reliability of protein associations more and more rigorously, the relative gene connectivity between chromosomal genes and plasmids genes were less likely to differ between cooperative genes and private genes.

Taken together, our study found strong evidence to claim that in prokaryotes, genes on chromosomes have higher levels of connectivity than genes on plasmids. This result supported and extended the original complexity hypothesis, which states that highly connected genes are transferred at lower rates than less connected genes. We then found a similar pattern in cooperative genes across different prokaryote species. Although this pattern was not as effective for cooperative genes as it was for private genes, the prediction of the complexity hypothesis could still be used to explain why cooperative genes were not overrepresented on plasmids as predicted by theories[38–40]. A natural progression of this work is to test the extension of the complexity hypothesis focusing on other MGEs. Further across species comparative studies are also required to validate whether cooperative genes are more likely to be found on MGEs rather than on chromosomes. By comparing the different selective pressures in determining the location of cooperative genes in terms of different MGEs, we would be able

to gain a clearer understanding of the role of HGT in affecting the evolution of prokaryotes, especially in affecting the evolution of cooperative traits in microbes.

Methods

Network Collection

We extracted protein-protein interaction (PPI) networks from STRING database version 11.0 [65](https://string-db.org/). We used PPI networks to calculate connectivity (see **Connectivity**) for genes in our dataset. Each strain in our dataset has a corresponding network (see **Database Matching and Genome Collection**). We chose STRING because it covers the largest number of organisms (5090), including microorganisms that allow cross-species comparative analysis. In addition, STRING is a more comprehensive PPI network database. Unlike other databases based on either experimental[66–69], or computational prediction interactions[70], STRING integrates both of them and includes direct (physical) and indirect (functional) associations. This allowed us to include as many species as possible in our analysis.

The evidence for each interaction in the STRING database is categorized into one of seven independent 'channels': neighbourhood, fusion, co-occurrence, co-expression, text-mining, experiments and databases. For each pair of interactions, a separate score is given per channel. A combined confidence score ranging from 0 – 1000 is denoted by combining and adjusting the scores from the different channels[71]. In our main analysis, we specified a threshold of 400 for the combined scores of the interactions, and any interaction below this threshold would not be considered. 400 is a medium confidence threshold according to the STRING database. To check the reproducibility of our results, we also repeated our analysis by setting different thresholds: 150 (low confidence), 700 (high confidence), and 900 (highest confidence). The results at different thresholds are presented in Supplement tables. To match with PSORTdb database (**see Database Matching and Genome Collection**), we retrieved all the available PPI networks by using the STRINGdb package (version 2.4.0) in R[65].

Categorization of Genes and Annotations of Replicons

To select genes that were putatively 'cooperative', we followed the methods of previous studies which have regarded genes with extracellular gene products as a proxy for 'cooperative' genes[48–50]. Although not all cooperative genes are extracellular and not all extracellular proteins are cooperative, any strong effect of sociality is likely to be captured by using this proxy[1]. We compiled the prediction results of the protein subcellular localization for each protein included in our analysis from PSORTdb 4.0 (https://db.psort.org/)[72]. PSORTdb was selected for its reliability and validity in systematically deducing both bacterial and archaeal SCLs.

PSORTdb gives a final prediction of the subcellular location for each protein. For Gram-positive bacteria, the program allocates proteins to one of four locations within the cell: cytoplasmic, cytoplasmic membrane, extracellular or cell wall. Many of the most well-studied Archaea contain the same basic components as classic Gram-positive bacteria[72]. For classic Gram-negative bacteria, proteins are assigned to one of five locations, where cell wall has been replaced by outer membrane or periplasmic. We excluded any proteins classified as "Unknown" by PSORTdb from our analysis, which accounted for 23.9% of all proteins we analyzed.

The PSORTdb outputs we used also included information about bacterial and archaeal replicons, from which we can infer whether the genes coding for proteins of interest are on plasmids or chromosomes. We initially collected precomputed PSORTdb results for all

available genomes, including 73136 replicons belonging to 8416 bacterial and archaeal strains, to keep all genomes that were also in the STRINGdb with a PPI network. All the PSORTdb results were retrieved and compiled using GNU Wget and R.

Database Matching and Genome Collection

To compare the connectivity (see below section) of genes encoding extracellular and intracellular proteins, we curated a list of bacterial and archaeal strains which were in both the PSORTdb and STRING databases. Each organism in STRING is assigned with a unique NCBI taxonomy ID as a specific identifier, whereas PSORTdb uses RefSeq genome/replicon accession to specify the genome. We therefore matched RefSeq accession numbers from PSORTdb with their corresponding NCBI taxonomy ID by adopting NCBI Entrez Direct Command Line Tools (https://www.ncbi.nlm.nih.gov/books/NBK179288/). By doing so, we were able to use the NCBI taxonomy ID to extract the PSORTdb results of all genomes which also had a PPI network(s). To allow us to compare chromosome and plasmid genes, we only considered genomes with PSORTdb results that included at least one plasmid sequence. Specifically, for our purpose of comparing the connectivity of genes coding for extracellular and intracellular proteins that are on plasmids, we omitted genomes with no extracellular protein-coding genes on their plasmids. This gave us a list of 161 species (167 genomes), which included 7 archaeal species (7 genomes) and 154 bacterial species (160 genomes).

For each gene in our dataset, we mapped the gene name to the STRING database identifier 'STRING_id' using the 'map' function of the R package STRINGdb version X[65]. This unique 'STRING_id' was used to calculate the connectivity and normalized connectivity (see **Connectivity**) for every individual gene. Genes that could not be mapped with 'STRING_id' were not included in our dataset.

Connectivity

In our analysis, we use the term gene connectivity to mean the same as the protein connectivity of its protein product in a PPI network. We followed previous studies to define protein connectivity as the number of protein-protein interactions (PPIs) in which the protein is embedded in the PPI network[43,61]. We used this definition because one fundamental assumption of the complexity hypothesis is that the more interactions a gene has with other genes, the more complex it is, and therefore the less likely the gene will be successfully transferred[41].

It is also noteworthy that network size (the total number of proteins in a PPI network) would affect gene connectivity[52]. Genes with the same connectivity have different impacts in networks of different sizes. To control the influence of network size on our across species analyses, we examine whether network size was correlated with the connectivity of genes in our dataset. We performed all calculations of connectivity using R package 'igragh'[73], and analogous functions written to check the package was working as expected.

Statistics

We carried out all statistical analysis and graph plotting in R (version 4.0.2). For all comparisons between groups that included all our species, we used the R package MCMCglmm[51]. To control for phylogenetic relationships between species in our dataset (see **Phylogenetic Reconstruction**), we used a phylogeny as random effects in our model. For each analysis, we used 1100000 model iterations with a starting burn-out phase of 100000, sampling every 1000 iterations. We then checked the reliability of all output models by looking at model convergence. After the model diagnoses, we reported the posterior mean, 95% Credible Intervals (functionally similar to 95% Confidence Intervals), and the pMCMC value (used here

as ‘p-value’) for each model. We also provided the R^2 for models in our main analysis using methods described in[74,75].

When we looked at the patterns within each species, we used the Kruskal-Wallis test which tests whether samples are from the same distribution to perform all comparisons between groups. The results are in Supplementary tables.

Phylogeny

To control for phylogenetic relationships between our species, we used a phylogenetic tree including all 161 species in our dataset (Fig S6). We put together this phylogeny using the methods of a recent study[1]. The tree was based on a recently published maximum likelihood tree of life using 16 ribosomal protein sequences data[76]. This tree typically has only one representative species of each genus. We first extracted all branches that matched species in our dataset by using the R package ‘ape’[77]. In cases where the representative species of a genus was not the same as our species from the same genus, we replaced the branch tip with our species, since all species from the same genus are equally related to species of sister genera. In cases where there were two species per genus in our dataset, we used the R package ‘phylotools’ to directly add the second species as an additional branch into their genera[78]. Where there were more than two species within a genus in our dataset, we consulted phylogenies from the literature (Supplementary table X) to add any within-genus clustering of species’ branches.

References

1. Dewar, A. E. *et al.* Plasmids do not consistently stabilize cooperation across bacteria, but may promote pathogen host-range. *Nature Ecology and Evolution* (2021).
2. Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & Cruz, F. de la. Mobility of Plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434–452 (2010).
3. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat Rev Genet* **16**, 472–482 (2015).
4. Brito, I. L. Examining horizontal gene transfer in microbial communities. *Nat Rev Microbiol* **19**, 442–453 (2021).
5. Wagner, A. *et al.* Mechanisms of gene flow in archaea. *Nat Rev Microbiol* **15**, 492–501 (2017).
6. Johnson, C. M. & Grossman, A. D. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annual Review of Genetics* **49**, 577–601 (2015).
7. Lang, A. S., Zhaxybayeva, O. & Beatty, J. T. Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol* **10**, 472–482 (2012).
8. Husnik, F. & McCutcheon, J. P. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Microbiol* **16**, 67–79 (2018).
9. Syvanen, M. Evolutionary Implications of Horizontal Gene Transfer. *Annual Review of Genetics* **46**, 341–358 (2012).
10. Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* **3**, 679–687 (2005).
11. Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. *Annual Review of Microbiology* **55**, 709–742 (2001).

12. Sørensen, S. J., Bailey, M., Hansen, L. H., Kroer, N. & Wuertz, S. Studying plasmid horizontal transfer in situ: a critical review. *Nat Rev Microbiol* **3**, 700–710 (2005).
13. Lehtinen, S., Huisman, J. S. & Bonhoeffer, S. Evolutionary mechanisms that determine which bacterial genes are carried on plasmids. *Evolution Letters* **5**, 290–301 (2021).
14. 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).
15. Werren, J. H. Selfish genetic elements, genetic conflict, and evolutionary innovation. *PNAS* **108**, 10863–10870 (2011).
16. 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).
17. Baltrus, D. A. Exploring the costs of horizontal gene transfer. *Trends in Ecology & Evolution* **28**, 489–495 (2013).
18. Pinto, U. M., Pappas, K. M. & Winans, S. C. The ABCs of plasmid replication and segregation. *Nat Rev Microbiol* **10**, 755–765 (2012).
19. Carroll, A. C. & Wong, A. Plasmid persistence: costs, benefits, and the plasmid paradox. *Can. J. Microbiol.* **64**, 293–304 (2018).
20. Melderen, L. V. & Bast, M. S. D. Bacterial Toxin–Antitoxin Systems: More Than Selfish Entities? *PLOS Genetics* **5**, e1000437 (2009).
21. Hernández-Arriaga, A. M., Chan, W. T., Espinosa, M. & Díaz-Orejas, R. Conditional Activation of Toxin-Antitoxin Systems: Postsegregational Killing and Beyond. *Microbiology Spectrum* **2**, 2.5.34 (2014).
22. Ni, S. *et al.* Conjugative plasmid-encoded toxin–antitoxin system PrpT/PrpA directly controls plasmid copy number. *PNAS* **118**, (2021).
23. Belogurov, A. A. *et al.* Antirestriction protein ard (type C) encoded by IncW plasmid psa has a high similarity to the "protein transport" domain of TraC1 primase of promiscuous plasmid RP411Edited by M. Gottesman. *Journal of Molecular Biology* **296**, 969–977 (2000).
24. Tock, M. R. & Dryden, D. T. The biology of restriction and anti-restriction. *Current Opinion in Microbiology* **8**, 466–472 (2005).
25. Ghigo, J.-M. Natural conjugative plasmids induce bacterial biofilm development. *Nature* **412**, 442–445 (2001).
26. Arnaouteli, S., Bamford, N. C., Stanley-Wall, N. R. & Kovács, Á. T. Bacillus subtilis biofilm formation and social interactions. *Nat Rev Microbiol* **19**, 600–614 (2021).
27. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nat Rev Microbiol* **3**, 711–721 (2005).
28. Pinilla-Redondo, R. *et al.* Type IV CRISPR–Cas systems are highly diverse and involved in competition between plasmids. *Nucleic Acids Research* **48**, 2000–2012 (2020).
29. Johnson, T. J. & Nolan, L. K. Pathogenomics of the Virulence Plasmids of Escherichia coli. *Microbiology and Molecular Biology Reviews* **73**, 750–774 (2009).
30. Pilla, G. & Tang, C. M. Going around in circles: virulence plasmids in enteric pathogens. *Nat Rev Microbiol* **16**, 484–495 (2018).
31. Che, Y. *et al.* Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. *PNAS* **118**, (2021).

32. Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clinical Microbiology Reviews* **31**, e00088-17.
33. Rozwandowicz, M. *et al.* Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *Journal of Antimicrobial Chemotherapy* **73**, 1121–1137 (2018).
34. San Millan, A. Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical Context. *Trends in Microbiology* **26**, 978–985 (2018).
35. Hehemann, J.-H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908–912 (2010).
36. Brinkmann, H., Göker, M., Koblížek, M., Wagner-Döbler, I. & Petersen, J. Horizontal operon transfer, plasmids, and the evolution of photosynthesis in Rhodobacteraceae. *ISME J* **12**, 1994–2010 (2018).
37. Riley, M. A. & Wertz, J. E. Bacteriocins: Evolution, Ecology, and Application. *Annual Review of Microbiology* **56**, 117–137 (2002).
38. Smith, J. The social evolution of bacterial pathogenesis. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268**, 61–69 (2001).
39. Ginty, S. E. M., Rankin, D. J. & Brown, S. P. Horizontal Gene Transfer and the Evolution of Bacterial Cooperation. *Evolution* **65**, 21–32 (2011).
40. 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. *Proceedings of the Royal Society B: Biological Sciences* **280**, 20130400 (2013).
41. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: The complexity hypothesis. *PNAS* **96**, 3801–3806 (1999).
42. Novick, A. & Doolittle, W. F. Horizontal persistence and the complexity hypothesis. *Biol Philos* **35**, 2 (2019).
43. 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).
44. Davids, W. & Zhang, Z. The impact of horizontal gene transfer in shaping operons and protein interaction networks – direct evidence of preferential attachment. *BMC Evol Biol* **8**, 23 (2008).
45. Lercher, M. J. & Pál, C. Integration of Horizontally Transferred Genes into Regulatory Interaction Networks Takes Many Million Years. *Molecular Biology and Evolution* **25**, 559–567 (2008).
46. Grafen, A. & Hamilton, W. D. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **326**, 119–157 (1989).
47. Tian, R.-M., Cai, L., Zhang, W.-P., Cao, H.-L. & Qian, P.-Y. Rare Events of Intragenus and Intraspecies Horizontal Transfer of the 16S rRNA Gene. *Genome Biology and Evolution* **7**, 2310–2320 (2015).
48. 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).
49. Nogueira, T. *et al.* Horizontal Gene Transfer of the Secretome Drives the Evolution of Bacterial Cooperation and Virulence. *Current Biology* **19**, 1683–1691 (2009).

50. 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).
51. Hadfield, J. D. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software* **33**, 1–22 (2010).
52. Li, X., Li, W., Zeng, M., Zheng, R. & Li, M. Network-based methods for predicting essential genes or proteins: a survey. *Briefings in Bioinformatics* **21**, 566–583 (2020).
53. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* **3**, 722–732 (2005).
54. Stone, G. N., Nee, S. & Felsenstein, J. Controlling for non-independence in comparative analysis of patterns across populations within species. *Philos Trans R Soc Lond B Biol Sci* **366**, 1410–1424 (2011).
55. Ives, A. R., Midford, P. E. & Garland, T., Jr. Within-Species Variation and Measurement Error in Phylogenetic Comparative Methods. *Systematic Biology* **56**, 252–270 (2007).
56. Janda, J. M. & Abbott, S. L. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. *Journal of Clinical Microbiology* **45**, 2761–2764 (2007).
57. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
58. He, X. & Zhang, J. Why Do Hubs Tend to Be Essential in Protein Networks? *PLOS Genetics* **2**, e88 (2006).
59. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101–113 (2004).
60. Rancati, G., Moffat, J., Typas, A. & Pavelka, N. Emerging and evolving concepts in gene essentiality. *Nat Rev Genet* **19**, 34–49 (2018).
61. Park, C. & Zhang, J. High Expression Hampers Horizontal Gene Transfer. *Genome Biology and Evolution* **4**, 523–532 (2012).
62. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLOS Computational Biology* **3**, e59 (2007).
63. Yu, H. & Gerstein, M. Genomic analysis of the hierarchical structure of regulatory networks. *PNAS* **103**, 14724–14731 (2006).
64. Bhardwaj, N., Kim, P. M. & Gerstein, M. B. Rewiring of Transcriptional Regulatory Networks: Hierarchy, Rather Than Connectivity, Better Reflects the Importance of Regulators. *Science Signaling* **3**, ra79–ra79 (2010).
65. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607–D613 (2019).
66. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research* **32**, D449–D451 (2004).
67. Orchard, S. *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* **42**, D358–D363 (2014).
68. Ammari, M. G., Gresham, C. R., McCarthy, F. M. & Nanduri, B. HPIDB 2.0: a curated database for host–pathogen interactions. *Database* **2016**, (2016).

69. Chatr-aryamontri, A. *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Research* **45**, D369–D379 (2017).
70. Zhang, Q. C., Petrey, D., Garzón, J. I., Deng, L. & Honig, B. PrePPI: a structure-informed database of protein–protein interactions. *Nucleic Acids Research* **41**, D828–D833 (2013).
71. von Mering, C. *et al.* STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Research* **33**, D433–D437 (2005).
72. Lau, W. Y. V. *et al.* PSORTdb 4.0: expanded and redesigned bacterial and archaeal protein subcellular localization database incorporating new secondary localizations. *Nucleic Acids Research* **49**, D803–D808 (2021).
73. Csardi, G. & Nepusz, T. The Igraph Software Package for Complex Network Research. *InterJournal* **Complex Systems**, 1695 (2005).
74. Nakagawa, S. & Schielzeth, H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* **4**, 133–142 (2013).
75. Nakagawa, S., Johnson, P. C. D. & Schielzeth, H. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface* **14**, 20170213 (2017).
76. Hug, L. A. *et al.* A new view of the tree of life. *Nat Microbiol* **1**, 1–6 (2016).
77. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
78. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**, 217–223 (2012).