The Role of Cooperation in Bacterial Ecology and Evolution

Chunhui Hao

Green Templeton College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2023

献给我的母亲

谁言寸草心

报得三春晖

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.

Chunhui Hao
Michaelmas Term 2023

Acknowledgments

I would like to express my gratitude to my supervisors, Stuart West and Melanie Ghoul. It was through your guidance that I grew to understand scientific research. You trained my skills and taught me to think, articulate, and pay attention to every detail as a scientist. Your support encouraged me to design and conduct scientific projects independently, enabling me to explore areas that intrigued me with confidence and determination.

Next, thank you to people whose insights, techniques, and wisdom enriched my thesis: Anna Dewar, Naoki Konno, Laurence Belcher, Zohar Katz, Craig MacLean, Josh Firth, Thomas Richards, Zheren Zhang, Thomas Scott, Ming Liu, Hanlun Liu, Xiaotong Zhang, Wenshuo Zhao, Miaoxiao Wang, Xinqiang Xi, Jordi Bascompte, and Wataru Iwasaki. A particular thank to Anna for her invaluable guidance in my initial phases and to Naoki for his meticulous attention to every facet of our work and for patiently addressing all my technical inquiries. Your professionalism has set a standard that I continually strive for.

Thank you to everyone in the West and Griffin Labs. Arriving in England as an international student, I couldn't have hoped for a more welcoming group. I'm especially grateful to the Thursday meeting group, Anna, Laurie, and Zohar, for their insightful and encouraging feedback.

My PhD journey began during the COVID pandemic. I couldn't have got through this without the support of my family and friends. To Zhijie, who always stood by me during my toughest moments, always comforting me and infusing me with strength and courage. To Jiahao, Zinan, and Jinyi – I'm fortunate to have the best housemates. Zihui and Yuan, our backyard BBQs will forever hold a special place in my heart.

Finally, to my parents, Jingbo and Zhen: your endless encouragement has always been my strength in this scientific endeavour. There is still a long road ahead, but with your support, I will go further.

Publications and Contributions

Chapter 2

Chapter 2 is a collaboration between myself and Ph.D. student Naoki Konno in the Graduate School of Science at the University of Tokyo.

- C.H., N.K. and S.A.W. conceived the study. C.H. conducted all the analyses. N.K. and S.A.W. participated in discussions and provided suggestions. C.H. wrote the Chapter. N.K. and S.A.W. helped revise the manuscript of the Chapter.

Chapter 3

- Hao, C., Dewar, A. E., West, S. A., & Ghoul, M. (2022). Gene transferability and sociality do not correlate with gene connectivity. *Proceedings of the Royal Society B*, 289(1987), 20221819.

Chapter 4

- C.H. conceived the study. C.H. conducted all the analyses. S.A.W. participated in discussions and provided suggestions. C.H. wrote the Chapter. S.A.W. helped revise the manuscript of the Chapter.

Appendix

The appendix includes a published paper, and two papers in review which I contributed to during my DPhil:

A. Laurence J Belcher, Anna E Dewar, **Chunhui Hao**, Melanie Ghoul, Stuart A West, Signatures of kin selection in a natural population of the bacteria *Bacillus subtilis*, *Evolution Letters*, Volume 7, Issue 5, October 2023, Pages 315–330, https://doi.org/10.1093/evlett/qrad029.

B. Dewar, A. E., **Hao, C.**, Belcher L.J., Ghoul, M., West, S. A., Bacterial lifestyle shapes pangenomes. *Nature Communications*. In review.

C. Belcher L.J., Dewar, A. E., **Hao, C.**, Katz, Z., Ghoul, M., West, S. A., SOCfinder: a genomic tool for identifying cooperative genes in bacteria. *Microbial Genomics*. In review.

Abstract

Bacteria can perform a range of cooperative behaviours, yet the significance of these cooperative behaviours in bacterial ecology and evolution is a burgeoning research area. In this thesis, I use a comparative genomics approach to elucidate the impact of cooperation on bacterial ecological and evolutionary processes. Specifically, I: (i) probe the effects of cooperation on the evolution of bacterial niche breadth; (ii) investigate the relationship between horizontal gene transfer and bacterial cooperation, with a specific emphasis on the interplay between gene connectivity and sociality in determining gene transfer potential. Additionally, I expand the scope by examining how ecological determinants and genomic characteristics shape the evolution of bacterial growth rates. In summary, this thesis offers novel perspectives on bacterial ecology and evolution through the lens of bacterial cooperation.

Contents

1 Introduction ……..…….……1

2 Cooperative mechanisms in determining bacterial niche breadth evolution……23

3 Gene transferability and sociality do not correlate with gene connectivity……...…..68

4 Environmental variability mediates the evolution of bacterial growth rate and genomic characteristics……..……..78

5 Discussion…….……..111

6 Supplementary materials……132

Appendices

A. Signatures of kin selection in a natural population of the bacteria Bacillus subtilis

B. Bacterial lifestyle shapes pangenomes.

C. SOCfinder: a genomic tool for identifying cooperative genes in bacteria.

Chapter 1: Introduction

Bacteria are an incredibly diverse and ancient group of single-celled organisms, which underpin myriad processes critical to both our personal health and the health of our planet (Figure 1). In the human body, a community of bacteria, known as the microbiome, participates in numerous essential physiological activities[1,2]. For instance, bacteria and their metabolic by-products can be detected by both haematopoietic and non-haematopoietic cells of the innate immune system, subsequently translating these signals into host physiological responses[3]. Additionally, some pathogenic bacteria release virulence factors, which can instigate intestinal inflammation, leading to diseases like colitis or gastroenteritis[4]. Beyond their impact on human health, bacteria also play a pivotal role in shaping broader environmental processes. They are central to numerous aspects of the world's biogeochemical cycles, engaging in processes such as anaerobic methane oxidation, photosynthesis, phosphorous absorption, organic pollutant biodegradation, and the nitrogen and sulphur cycles[5,6]. Moreover, the intricate interactions between the microbiome and ecosystems, such as soils and forests, have broader repercussions for agricultural systems and global ecological patterns[7–9].

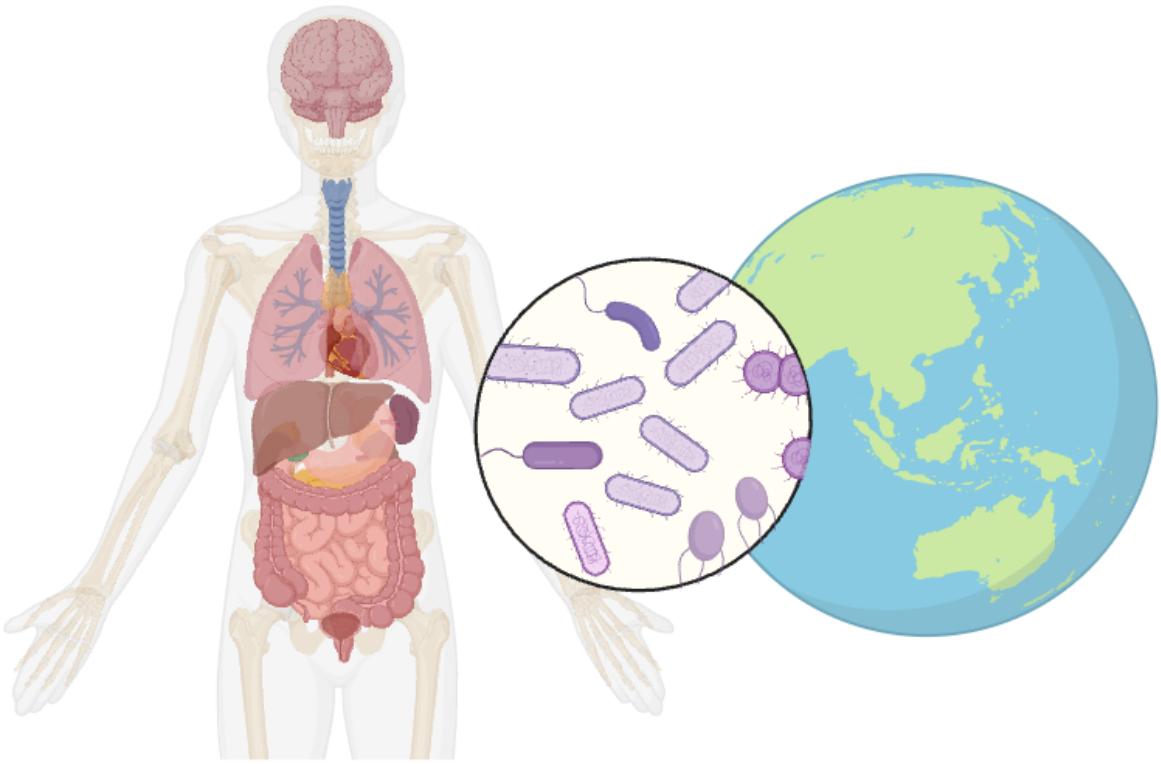


Figure 1. Bacteria is critical to both our personal health and the health of our planet. Bacteria are found to associate with multiple human diseases, including tuberculosis (caused by *Mycobacterium tuberculosis*), bacterial pneumonia (caused by pathogens like *Streptococcus pneumoniae*), and Lyme disease (caused by *Borrelia burgdorferi*). Bacteria also play a pivotal role in shaping broader environmental processes, such as nitrogen fixation (by genera like *Rhizobium* and *Azotobacter*), decomposition of organic matter (by countless saprophytic bacteria), and carbon cycling (via photosynthetic bacteria and those involved in carbon decomposition).

The ecology and evolution of bacteria

The profound influence of bacteria on our world underscores the importance of studying their ecology and evolution. First, investigating bacterial ecology and evolution offers insights into how these microorganisms engage with and adapt to fluctuating environments, how they interact amongst themselves, and how their characteristics influence the structure and dynamics of systems spanning from the human body to global ecosystems[10,11]. A notable instance illustrating bacterial evolution is the emergence of antibiotic resistance. When exposed to antibiotics, certain bacterial populations have evolved resistance mechanisms, resulting in the birth of formidable antibiotic-resistant strains. These resistance genes can further spread through horizontal gene transfer (HGT), exacerbating the challenge of bacterial infections and posing substantial threats to modern medical treatments[12–14].

Second, principles of community ecology, such as resilience, community disturbances and extinction, community assembly and succession offer lenses to comprehend the stability and dynamics of microbiomes, and the implications of these factors on human health[15–19]. For instance, research has shown that the assembly of the infant gut microbiota adheres to principles of community succession: Initially, pioneer taxa like *Staphylococcus* colonize the gut after birth. Subsequently, late-arriving members of the microbiome, such as *Klebsiella*, exploits the pioneer microorganism to gain a foothold within the gut[20].

Third, due to their rapid reproduction rates, suitability for laboratory experiments, and substantial genetic diversity, bacteria serve as ideal models to investigate fundamental phenomena in ecology and evolution. They enable scientists to test theories at scales and speeds that are challenging to achieve with more complex organisms[21]. Therefore, by studying the ecology and evolution of bacteria, not only do we gain a deeper understanding of bacterial life,

but we also equip ourselves with invaluable tools and insights to enrich the broader disciplines of ecology and evolution.

Cooperation in bacteria

a) Cooperation is common among bacteria

This thesis is particularly interested in understanding the ecology and evolution of bacterial cooperation. Historically, studies on cooperation predominantly focused on readily observable social behaviours, like the clear examples of altruism seen in social insects[22]. However, there's a recent shift in focus towards understanding the cooperative behaviours of microorganisms, such as biofilm formation, quorum sensing, and the compelling case of siderophore production[23,24]. Using siderophore production as a paradigm of "public good" cooperation among bacteria. Bacteria release these iron-binding molecules to scavenge iron from the environment. While they bear the cost of siderophore production, both the producers and their neighbouring non-producers reap the benefits, enhancing collective survival in environments with limited iron[25].

b) Evolutionary explanation for bacterial cooperation

To explain how and why cooperation emerges and is maintained in bacteria, an inevitable fact needs to be accounted for: cooperation is costly. Signalling molecules impose a metabolic burden on both the producing and recipient bacteria[10]. As a result, the emergence of "cheaters", those that benefit from cooperation without contributing, might gain an advantage and outnumber genuine cooperators[25,26].

With such evident costs, one might wonder why a bacterium would engage in cooperative behaviour that seemingly benefit others or the group over itself. Initially, the prevalent thought

among microbiologists was that cooperation was advantageous at the population or species scale, a concept referred to as the group selection perspective. Yet, extensive empirical and theoretical research has since challenged this perspective, suggesting it isn't the primary determinator behind bacterial cooperation[23]. Subsequent research then emphasized the necessity of examining both the direct and indirect benefits of cooperation. This line of thinking, consolidated into the inclusive fitness theory, has been instrumental in elucidating cooperative behaviours, not only in bacteria but across all living organisms[27,28] (Figure 2). In particular, the kin selection theory offers an explanation for altruistic cooperation, where the cooperating individual doesn't directly benefit. It suggests that by assisting a close relative to reproduce, an individual indirectly propagates its genes to future generations. This concept is concisely articulated in Hamilton's rule, which states that a behaviour will be evolutionarily favoured if the equation $rB - C > 0$ holds true. Here, C is the cost to the cooperating individual, B is the benefit to the recipient, and r denotes the genetic relatedness between the two[29].

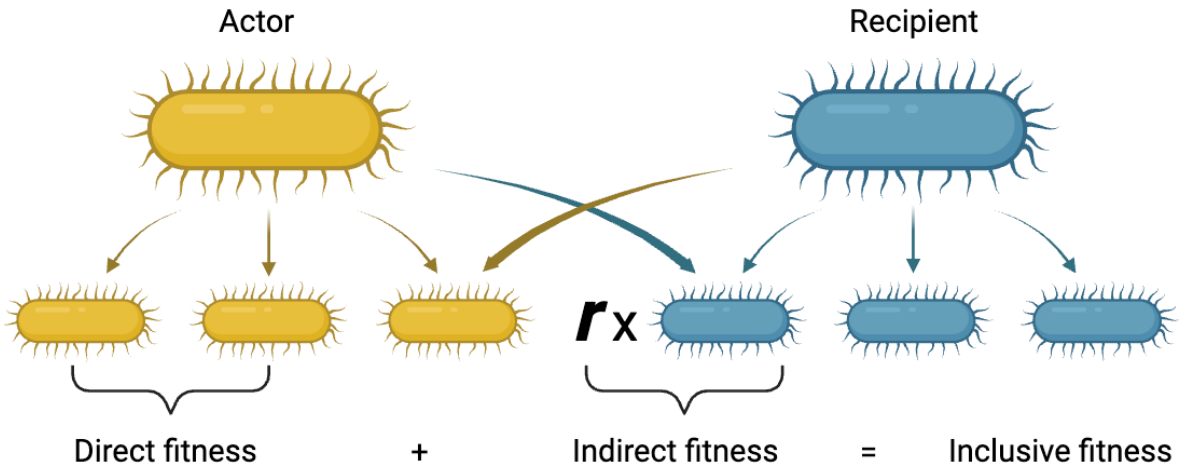


Figure 2: Understanding inclusive fitness in bacterial cooperation. Inclusive fitness comprises both direct and indirect fitness components. The prominent yellow bacterium exemplifies the actor exhibiting cooperative behaviour. The direct fitness is represented by the influence of the actor's behaviour (indicated by yellow arrows) on its reproductive success

(small yellow cells). The indirect fitness effect is showcased by how the actor's behaviour affects the reproductive output of its cooperative partners (small blue cells), taking into account the relatedness (*r*) between the actor and the recipient (large blue bacterium).

The role of cooperation in bacterial eco-evo processes

While our understanding of bacterial cooperation's evolutionary mechanisms is robust, its influence on bacterial ecological and evolutionary (eco-evo) processes remains an area of increasing research interest in recent decades.

a) Eco-evo processes shape bacterial cooperation

Various eco-evo dynamics can either foster or inhibit cooperative behaviours in bacteria. A notable example is the role of spatial organization and mobility on cooperation in bacterial populations. Within structured spaces, bacteria with cooperative tendencies often interact with genetically related individuals, enriching mutually advantageous behaviours[30]. This structured grouping fosters an environment conducive to cooperation, mainly benefiting those with high genetic relatedness[31]. However, increased migration disrupts this structure. As bacteria intermingle, the advantages of local cooperation diminish, offering opportunities for non-cooperative bacteria. These opportunists utilize the benefits generated by cooperators without reciprocating. Over time, as the number of cooperators dwindles, these cheaters can migrate to areas with a higher density of cooperators to repeat their tricks[32].

Another critical consideration is the impact of horizontal gene transfer (HGT) on bacterial cooperation. HGT might favour cooperation by disseminating cooperative genes among different bacterial strains. Under this premise, any "cheater" that lose the gene could potentially reacquire the cooperative gene through HGT, increasing relatedness and enhancing

cooperation[33–35]. Supporting this, some studies indicate that mobile genetic elements, like plasmids, may house a larger portion of genes for extracellular proteins (potential cooperative genes) compared to more stationary chromosomes[36–38]. However, counter-evidence emerged from recent studies. They suggested that while HGT can facilitate the initial spread of cooperative genes, its role in long-term cooperative maintenance is less pronounced[39]. Given the diverse findings regarding HGT's influence on cooperation, further nuanced studies are essential to delineate its exact role in varied bacterial species and contexts.

b) Cooperation as a catalyst in eco-evo processes

Cooperation can also significantly influence eco-evo processes within microbial populations and communities. For instance, the release of extracellular public goods can elevate the resource extraction and utilization efficiency in a microbial ecosystem (Figure 3a). This not only amplify community productivity but also enhance resilience against external perturbations. In environments with limited nutrients, resource-sharing through cooperation can be a lifeline for microbial communities[40].

Another example highlights how the production of public goods aids in facilitating coexistence, which is pivotal for maintaining diversity. Studies have uncovered intriguing non-transitive dynamics among strains or species with varying siderophore strategies (Figure 3b). For instance, a weak siderophore-producing strain might shield its resources from a non-producer, only to be outcompeted by a robust siderophore-producer. This stronger strain is, in turn, exploited by the non-producer. Such interactions embody a rock-paper-scissors dynamic, with strains cyclically outcompeting one another without a definitive dominant strain[41–43].

Furthermore, the stability of the microbiome is believed to be intertwined with cooperative behaviours. While the advantages of cooperation, such as enhanced resource utilization or fortified defence mechanisms, are evident, some theories argued that cooperation might actually destabilize the microbiome[44–46] (Figure 3c). This is attributed to the interconnected nature of cooperative entities, resulting in positive feedback loops. When one species diminishes in abundance, it can adversely impact its cooperative partners, potentially jeopardizing the system's stability[46–48].

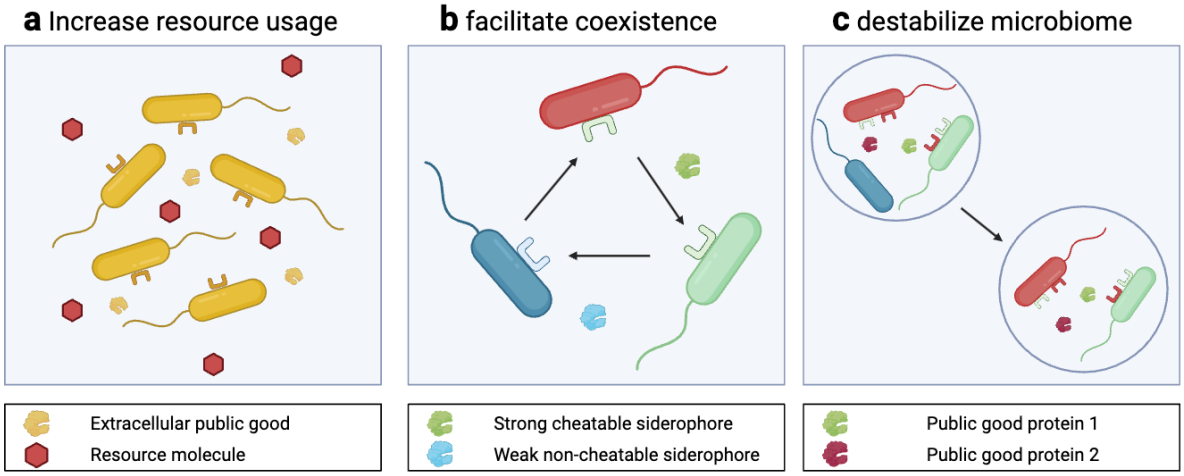


Figure 3. Cooperation as a catalyst in eco-evo processes. (a) Cooperation enhances resource efficiency. The release of extracellular public goods amplifies the effectiveness of resource consumption within a microbial community. (b) Strains with different siderophore strategies can engage in a rock-paper-scissors dynamic, fostering coexistence. As illustrated, a cheating strain (red cell) has an advantage over a siderophore producer (green cell) that releases a compatible siderophore (green protein). This producer then outcompetes a second siderophore producer (blue cell) which emits a less efficacious siderophore (blue protein). The less potent siderophore producer outcompetes the non-producing strain, as the latter lacks the appropriate receptor for the weak siderophore. (c) Bacterial cooperation destabilizes the microbiome. Cooperative pairs (red and green cells) proliferate and dominate. However, their dominance

can lead to the elimination of other strain (blue cell), thereby destabilizing the community equilibrium.

Unexplored Territory: Cooperation's influence on bacterial biogeography

a) What is bacterial biogeography?

While cooperation is recognized in various eco-evo processes, its impact on bacterial biogeographic patterns remains largely untapped. Biogeography studies biodiversity distribution across space and time. It seeks to understand where bacteria exist, at what abundance, and the reasons behind their distributions[49,50]. Bacteria exhibit biogeographic patterns temporally and spatially. Temporally, bacterial compositions change significantly within years or show recurring seasonal patterns[51,52]. Spatially, evidence for bacterial biogeographic patterns is more solid, with different taxa observed in diverse habitats[53–57].

b) Eco-evo processes behind bacterial biogeography

The study of bacterial biogeography offers insights into the mechanisms that generate and maintain spatial variations among bacteria. Some studies have aimed to separate the impacts of current environmental factors from historical influences on these patterns[49]. Another framework, rooted in general ecology and evolution, posited that bacterial diversity arises from four main processes: speciation, selection, dispersal, and drift[50,58]. Speciation introduces new species. Selection adjusts the abundance of certain species based on survival and reproductive outcomes, and may be influenced by biotic interactions like cooperation and competition. Lastly, the relocation and successful establishment of a species in a new place (dispersal), along with stochastic factors (drift), also play roles in determining species diversity across various locations.

A central element in bacterial biogeography is niche breadth – the range of conditions in which a species can thrive[59]. In the context of bacterial spatial distribution, niche breadth can denote the variety of unique habitats a species can occupy. Certain bacteria are "habitat generalists", adaptable to various environments, while others, "habitat specialists", are limited to specific niches[59–61]. Differences in niche breadth between bacterial species are subject to factors such as speciation, selection, dispersal, and drift. For instance, research indicated that generalist lineages have speciation rates 19 times higher than specialists, emphasizing generalists' role in introducing new species and maintaining biodiversity[62]. Furthermore, dispersal ability appears important in determining bacterial niche breadth, and generalists are more adept at dispersing than specialists[63–66].

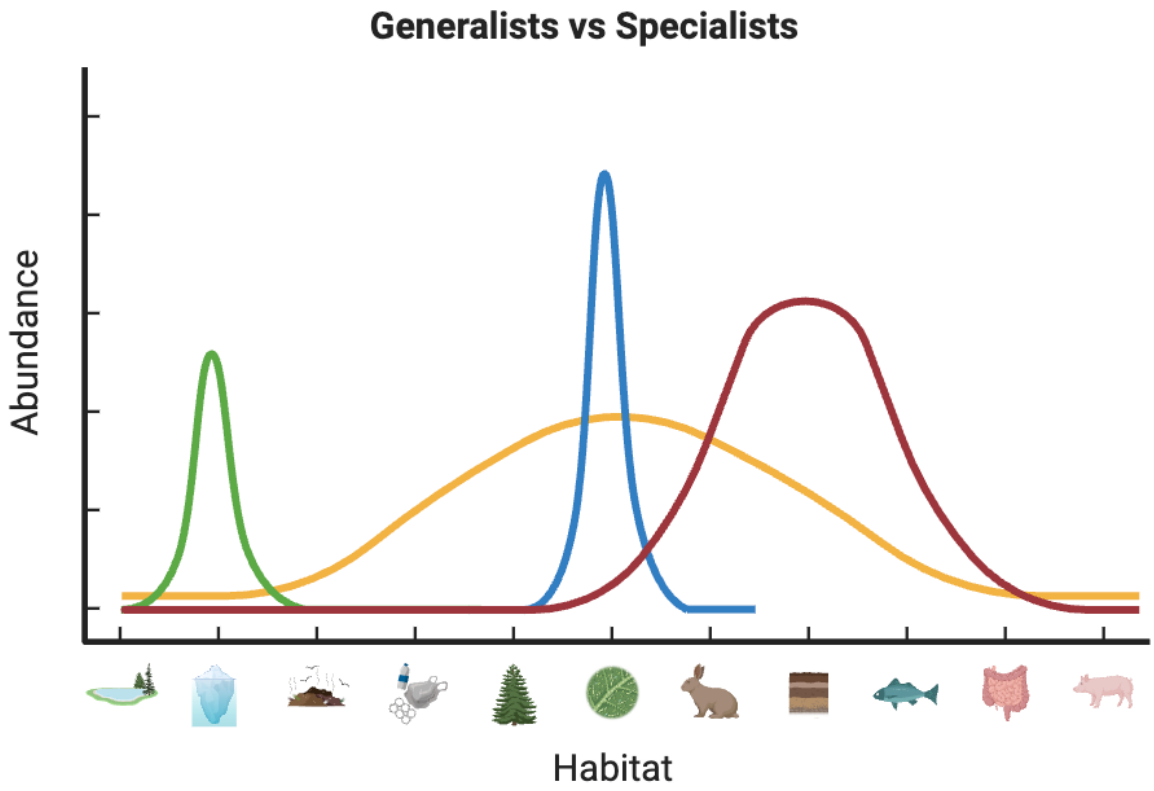


Figure 4. Generalists versus specialists. In this figure, we illustrate the contrast between habitat generalists (represented by yellow and red lines) and habitat specialists (represented by

green and blue lines). It is important to note that the abundance of habitat specialists is only significant in specific habitats, whereas generalists can thrive in multiple habitats, indicating their broader niche breadth when compared to habitat specialists.

c) The role of cooperation in the bacterial niche breath evolution is unclear

Regarding selection, it's intriguing to consider how cooperation-induced selective pressure may shape niche breadth variations. Cooperation often boosts efficiency in various tasks, enhancing growth and survival. As an illustration, siderophores produced by one individual can be used by another, maximizing iron uptake in environments where this mineral is scarce[25]. This collective use of resources suggests that cooperation might drive evolution towards generalization by optimizing resource use efficiency. Comparative genomics offers an ideal method to explore whether cooperation aids in niche expansion. A deeper look into the relationship between the carriage of cooperative genes and niche breadth across species might reveal if organisms with wider niches tend to have more cooperative genes. In supporting this, several genomic studies have found that generalists often possess a higher number of cooperative genes than specialists[67–69].

While it's often assumed that cooperation catalyses niche expansion, this might not always be the case. It is possible that causality is in the opposite direction — as bacteria diversify into new niches, the importance of cooperative genes intensifies, thereby seeing a heightened presence in generalists. Alternatively, the link between bacterial niche breadth and the carriage of cooperative genes might not be causative; a third factor, intertwined with both, could play the pivotal role[70,71]. To truly discern the nature of this relationship, a focused examination into the causal relationship between bacterial cooperation and niche breadth evolution is imperative.

Additionally, bacterial genes commonly experience simultaneous gains or losses. This trend is often due to coevolution of genes as bacteria adapt to different environments. For instance, when bacteria face a new habitat, they might need to acquire specific metabolic genes to metabolize new resources. These metabolic genes often work together in certain pathways, so when one gene is gained, it might prompt the acquisition of other related genes. This raises an important question: Are cooperative genes commonly gained or lost alongside other gene types? If so, the evolutionary implications could be intricate, since gains or losses in cooperative genes might be linked to gains or losses in other genes that also influence the bacteria's adaptability to different habitats. Consequently, to fully grasp how cooperation impacts bacterial niche breadth, we must look beyond just the cooperative genes.

The relationship between HGT and bacterial cooperation: A revisit through the lens of the Complexity Hypothesis

Previously, I highlighted the suggestion that cooperative genes may be more prone to horizontal transfer[34,36,72]. However, when considering the likelihood of a gene undergoing HGT, other factors come into play. A prime factor is gene connectivity, which measures how many links a gene's protein product has with products of other genes[73]. The "complexity hypothesis" posited that genes with high connectivity should not be present in genome regions prone to HGT. The rationale is straightforward: if a highly interconnected gene was transferred to a new host, its functionality would be compromised without the presence of other genes it is connected to[74]. This hypothesis has received support from various studies that have noted a negative correlation between HGT rates and gene connectivity[75,76].

In view of the complexity hypothesis, if certain genes have a higher propensity for horizontal transfer, they are expected to exhibit lower connectivity. This is because high connectivity

could hinder their functionality upon transferring horizontally and entering new hosts. Therefore, in the case of cooperative genes, there is an intriguing interplay between gene connectivity and gene sociality when considering their transferability. If cooperative genes are indeed primed for horizontal transfer, they might be characterized by notably lower connectivity levels.

Thesis Outline

Here I explore the role of cooperation in bacterial ecological and evolutionary processes using comparative genomics approach.

Chapter 2: Bacterial Cooperation and Niche Breadth Evolution

In this chapter, I used a comparative genomics approach to analyse the influence of bacterial cooperation on niche breadth evolution. I found a significant positive correlation between the proportion of cooperative genes and bacterial niche breadth across over 20,000 bacterial species. I then identified a causality of this correlation where a reduction in cooperative gene proportion can lead to niche contraction, suggesting a critical role for cooperative genes in sustaining the niche breadth of extant generalist species. Additionally, while cooperative genes undergo swift gains or losses in the short term, they remain relatively stable over extended periods, hinting at their non-primary role in niche expansion due to their transitory nature. Finally, genes that co-evolve with cooperative genes mainly narrow the niche breadth of extant generalists, contrasting with the more stabilizing influence of cooperative genes.

Chapter 3: Gene Transferability and Sociality Do Not Correlate with Gene Connectivity

This chapter delves into the interplay between gene connectivity and sociality, and their collective impact on gene transferability. Through analysis of chromosomal and plasmid genes

across 134 diverse prokaryotic species, I asked three questions: (i) Are chromosomal genes more connected than plasmids genes? (ii) Do genes on plasmids with higher transfer rates have lower connectivity? (iii) Does gene connectivity vary between cooperative genes versus private genes, and does this vary depending upon whether a gene is on a plasmid or the chromosome? We found that genes on plasmids tend to be less connected than their chromosomal counterparts. Moreover, there was no discernible correlation between a plasmid gene's connectivity and its mobility, and intriguingly, the sociality of genes (cooperative or private) was not correlated with gene connectivity.

Chapter 4: Driving Factors Behind Bacterial Growth Rate Evolution

This chapter examines the complex interplay between environmental factors and genomic characteristics in determining bacterial growth rates across species. Prevailing hypotheses propose a positive correlation between environmental variability and growth rates while suggesting a trade-off between growth rates and factors like genome size and metabolic features. However, my analysis of genomic data from 171 bacterial species painted a more nuanced picture. I revealed that species thriving in variable environments indeed tend to have accelerated growth rates. However, instead of the anticipated trade-off, genomic features were found to positively correlate with growth rates. Delving deeper, our data revealed a strong positive relationship between environmental variability and genomic characteristics. This suggests that the effect of environmental variability on both growth rates and genomic characteristics might offset the traditionally expected trade-offs between genomic features and growth rates.

Appendix: Supplementary Contributions During DPhil

During my DPhil studies, I also contributed to three additional manuscripts: (1) examining the kin selection patterns in the cooperative genes of *Bacillus subtilis* (Belcher et al., Evolution Letters); (2) probing the ways in which bacterial lifestyles influence pangenome fluidity (Dewar et al., under review, Nature Communications); (3) developing bioinformatic pipeline, *SocFinder*, to mine cooperative genes in genomes (Belcher et al., under review, Microbial Genomics).

References

1. Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease. *Curr Opin Gastroenterol* **31**, 69–75 (2015).
2. Hou, K. *et al.* Microbiota in health and diseases. *Sig Transduct Target Ther* **7**, 1–28 (2022).
3. Thaiss, C. A., Zmora, N., Levy, M. & Elinav, E. The microbiome and innate immunity. *Nature* **535**, 65–74 (2016).
4. Bäumler, A. J. & Sperandio, V. Interactions between the microbiota and pathogenic bacteria in the gut. *Nature* **535**, 85–93 (2016).
5. Madsen, E. L. Microorganisms and their roles in fundamental biogeochemical cycles. *Current Opinion in Biotechnology* **22**, 456–464 (2011).
6. Hawley, A. K. *et al.* Diverse Marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients. *Nat Commun* **8**, 1507 (2017).
7. Lladó, S., López-Mondéjar, R. & Baldrian, P. Forest Soil Bacteria: Diversity, Involvement in Ecosystem Processes, and Response to Global Change. *Microbiology and Molecular Biology Reviews* **81**, 10.1128/mmbr.00063-16 (2017).

8. Trivedi, P., Leach, J. E., Tringe, S. G., Sa, T. & Singh, B. K. Plant–microbiome interactions: from community assembly to plant health. *Nat Rev Microbiol* **18**, 607–621 (2020).
9. Hartmann, M. & Six, J. Soil structure and microbiome functions in agroecosystems. *Nat Rev Earth Environ* 1–15 (2022) doi:10.1038/s43017-022-00366-w.
10. Keller, L. & Surette, M. G. Communication in bacteria: an ecological and evolutionary perspective. *Nat Rev Microbiol* **4**, 249–258 (2006).
11. Herold, M. *et al.* Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nat Commun* **11**, 5281 (2020).
12. Davies, J. & Davies, D. Origins and Evolution of Antibiotic Resistance. *Microbiology and Molecular Biology Reviews* **74**, 417–433 (2010).
13. MacLean, R. C. & Millan, A. S. The evolution of antibiotic resistance. *Science* **365**, 1082–1083 (2019).
14. Larsson, D. G. J. & Flach, C.-F. Antibiotic resistance in the environment. *Nat Rev Microbiol* **20**, 257–269 (2022).
15. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat Rev Genet* **13**, 260–270 (2012).
16. Foster, K. R., Schluter, J., Coyte, K. Z. & Rakoff-Nahoum, S. The evolution of the host microbiome as an ecosystem on a leash. *Nature* **548**, 43–51 (2017).
17. Gonze, D., Coyte, K. Z., Lahti, L. & Faust, K. Microbial communities as dynamical systems. *Current Opinion in Microbiology* **44**, 41–49 (2018).
18. Coyte, K. Z. *et al.* Horizontal gene transfer and ecological interactions jointly control microbiome stability. *PLOS Biology* **20**, e3001847 (2022).
19. Gilbert, J. A. & Lynch, S. V. Community ecology as a framework for human microbiome research. *Nat Med* **25**, 884–889 (2019).

20. Rao, C. *et al.* Multi-kingdom ecological drivers of microbiota assembly in preterm infants. *Nature* **591**, 633–638 (2021).
21. Jessup, C. M. *et al.* Big questions, small worlds: microbial model systems in ecology. *Trends in Ecology & Evolution* **19**, 189–197 (2004).
22. West, S. A., Cooper, G. A., Ghoul, M. B. & Griffin, A. S. Ten recent insights for our understanding of cooperation. *Nat Ecol Evol* **5**, 419–430 (2021).
23. West, S. A., Griffin, A. S., Gardner, A. & Diggle, S. P. Social evolution theory for microorganisms. *Nat Rev Microbiol* **4**, 597–607 (2006).
24. West, S. A., Diggle, S. P., Buckling, A., Gardner, A. & Griffin, A. S. The Social Lives of Microbes. *Annual Review of Ecology, Evolution, and Systematics* **38**, 53–77 (2007).
25. Griffin, A. S., West, S. A. & Buckling, A. Cooperation and competition in pathogenic bacteria. *Nature* **430**, 1024–1027 (2004).
26. 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).
27. Queller, D. C. & Goodnight, K. F. ESTIMATING RELATEDNESS USING GENETIC MARKERS. *Evolution* **43**, 258–275 (1989).
28. Abbot, P. *et al.* Inclusive fitness theory and eusociality. *Nature* **471**, E1–E4 (2011).
29. Hamilton, W. D. The genetical evolution of social behaviour. II. *Journal of Theoretical Biology* **7**, 17–52 (1964).
30. Nadell, C. D., Drescher, K. & Foster, K. R. Spatial structure, cooperation and competition in biofilms. *Nat Rev Microbiol* **14**, 589–600 (2016).
31. West, S. A., Griffin, A. S. & Gardner, A. Evolutionary Explanations for Cooperation. *Current Biology* **17**, R661–R672 (2007).
32. Funk, F. & Hauert, C. Directed migration shapes cooperation in spatial ecological public goods games. *PLOS Computational Biology* **15**, e1006948 (2019).

33. Lee, I. P. A., Eldakar, O. T., Gogarten, J. P. & Andam, C. P. Bacterial cooperation through horizontal gene transfer. *Trends in Ecology & Evolution* **0**, (2021).
34. Smith, J. The social evolution of bacterial pathogenesis. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268**, 61–69 (2001).
35. Ginty, S. E. M., Rankin, D. J. & Brown, S. P. Horizontal Gene Transfer and the Evolution of Bacterial Cooperation. *Evolution* **65**, 21–32 (2011).
36. Nogueira, T. *et al.* Horizontal Gene Transfer of the Secretome Drives the Evolution of Bacterial Cooperation and Virulence. *Current Biology* **19**, 1683–1691 (2009).
37. Dimitriu, T. *et al.* Genetic information transfer promotes cooperation in bacteria. *Proceedings of the National Academy of Sciences* **111**, 11103–11108 (2014).
38. Wang, Q., Wei, S., Silva, A. F. & Madsen, J. S. Cooperative antibiotic resistance facilitates horizontal gene transfer. *ISME J* **17**, 846–854 (2023).
39. Dewar, A. E. *et al.* Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range. *Nat Ecol Evol* **5**, 1624–1636 (2021).
40. Dai, T. *et al.* Nutrient supply controls the linkage between species abundance and ecological interactions in marine bacterial communities. *Nat Commun* **13**, 175 (2022).
41. Inglis, R. F., Biernaskie, J. M., Gardner, A. & Kümmerli, R. Presence of a loner strain maintains cooperation and diversity in well-mixed bacterial communities. *Proceedings of the Royal Society B: Biological Sciences* **283**, 20152682 (2016).
42. Leinweber, A., Fredrik Inglis, R. & Kümmerli, R. Cheating fosters species co-existence in well-mixed bacterial communities. *ISME J* **11**, 1179–1188 (2017).
43. Kramer, J., Özkaya, Ö. & Kümmerli, R. Bacterial siderophores in community and host interactions. *Nat Rev Microbiol* **18**, 152–163 (2020).

44. Van den Abbeele, P., Van de Wiele, T., Verstraete, W. & Possemiers, S. The host selects mucosal and luminal associations of coevolved gut microorganisms: a novel concept. *FEMS Microbiology Reviews* **35**, 681–704 (2011).

45. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-Bacterial Mutualism in the Human Intestine. *Science* **307**, 1915–1920 (2005).

46. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science* **350**, 663–666 (2015).

47. Oliveira, N. M., Niehus, R. & Foster, K. R. Evolutionary limits to cooperation in microbial communities. *Proceedings of the National Academy of Sciences* **111**, 17941–17946 (2014).

48. Coyte, K. Z. & Rakoff-Nahoum, S. Understanding Competition and Cooperation within the Mammalian Gut Microbiome. *Current Biology* **29**, R538–R544 (2019).

49. Martiny, J. B. H. *et al.* Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**, 102–112 (2006).

50. Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C. & Martiny, J. B. H. Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Microbiol* **10**, 497–506 (2012).

51. Gilbert, J. A. *et al.* Defining seasonal marine microbial community dynamics. *ISME J* **6**, 298–308 (2012).

52. Caporaso, J. G., Paszkiewicz, K., Field, D., Knight, R. & Gilbert, J. A. The Western English Channel contains a persistent microbial seed bank. *ISME J* **6**, 1089–1093 (2012).

53. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences* **104**, 11436–11440 (2007).

54. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

55. Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).

56. Lima-Mendez, G. *et al.* Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).

57. Wu, L. *et al.* Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat Microbiol* **4**, 1183–1195 (2019).

58. Vellend, M. Conceptual Synthesis in Community Ecology. *The Quarterly Review of Biology* **85**, 183–206 (2010).

59. Sexton, J. P., Montiel, J., Shay, J. E., Stephens, M. R. & Slatyer, R. A. Evolution of Ecological Niche Breadth. *Annual Review of Ecology, Evolution, and Systematics* **48**, 183–206 (2017).

60. Sexton, J. P., McIntyre, P. J., Angert, A. L. & Rice, K. J. Evolution and Ecology of Species Range Limits. *Annual Review of Ecology, Evolution, and Systematics* **40**, 415–436 (2009).

61. Muller, E. E. L. Determining Microbial Niche Breadth in the Environment for Better Ecosystem Fate Predictions. *mSystems* **4**, 10.1128/msystems.00080-19 (2019).

62. Sriswasdi, S., Yang, C. & Iwasaki, W. Generalist species drive microbial dispersion and evolution. *Nat Commun* **8**, 1162 (2017).

63. Pandit, S. N., Kolasa, J. & Cottenie, K. Contrasts between habitat generalists and specialists: an empirical extension to the basic metacommunity framework. *Ecology* **90**, 2253–2262 (2009).

64. Monard, C., Gantner, S., Bertilsson, S., Hallin, S. & Stenlid, J. Habitat generalists and specialists in microbial communities across a terrestrial-freshwater gradient. *Sci Rep* **6**, 37719 (2016).

65. Kneitel, J. M. Occupancy and environmental responses of habitat specialists and generalists depend on dispersal traits. *Ecosphere* **9**, e02143 (2018).

66. Qiao, H., Saupe, E. E., Soberón, J., Peterson, A. T. & Myers, C. E. Impacts of Niche Breadth and Dispersal Ability on Macroevolutionary Patterns. *The American Naturalist* **188**, 149–162 (2016).

67. von Meijenfeldt, F. A. B., Hogeweg, P. & Dutilh, B. E. A social niche breadth score reveals niche range strategies of generalists and specialists. *Nat Ecol Evol* **7**, 768–781 (2023).

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

69. 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).

70. Dillard, J. R. & Westneat, D. F. Disentangling the Correlated Evolution of Monogamy and Cooperation. *Trends in Ecology & Evolution* **31**, 503–513 (2016).

71. Cornwallis, C. K. *et al.* Cooperation facilitates the colonization of harsh environments. *Nat Ecol Evol* **1**, 1–10 (2017).

72. 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).

73. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101–113 (2004).

74. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: The complexity hypothesis. *PNAS* **96**, 3801–3806 (1999).

75. 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).

76. 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).

Chapter 2: Cooperative Mechanisms in Determining Bacterial Niche Breadth Evolution

Abstract

Bacteria exhibit varied habitat preferences, with generalist species thriving in diverse habitats while specialists are confined to specific niches. We explored the role of bacterial cooperation in shaping the evolution of niche breadth using causal inference. Our findings revealed a positive correlation between the proportion of cooperative genes and bacterial niche breadth. This correlation results from a causality where a decreased proportion of cooperative genes promotes niche contraction, suggesting that cooperative genes mainly serve to maintain the niche breadth of generalists. Further analysis indicated that cooperative genes undergo rapid gains or losses on short-term scales, suggesting they may not be primary drivers for niche expansion due to their transitory nature on microevolutionary scales. By extending our analytical framework to encompass genes that co-evolved with cooperative genes, we discerned that they primarily narrow the niche breadth of existing generalists, in contrast to the stabilizing role of cooperative genes.

Introduction

In the field of microbial ecology, a key objective is to decipher the global distribution patterns of bacteria and identify the driving forces behind them[1–5]. One key aspect of bacterial distribution is niche breadth. Certain bacterial species are "habitat generalists", capable of adapting to a diverse range of environments, while others are "habitat specialists", restricted to very specific ecological niches[6–8]. While both abiotic and biotic factors have been suggested to shape bacterial niche breadth evolution, many studies have been descriptive or relied on straightforward comparisons, overlooking the underlying mechanisms [9–13].

One relatively unexplored factor that may play a role in promoting bacterial niche expansion is cooperation. Cells produce and excrete a variety of factors that benefit the local population, acting as cooperative "public goods". Examples include iron-scavenging siderophores[14], proteases that digest extracellular proteins[15] and exotoxins that disintegrate host cell membranes[16]. Cooperation provides mechanism for increasing efficiencies across various tasks, favouring growth and survival[17–20]. Therefore, it is possible that cooperation could facilitate the evolution towards generalization by enhancing resource utilization efficiency. This hypothesis has been supported by several genomic studies, which showed that cooperative genes were more frequently found in generalists than in specialists[10,21,22].

The idea that cooperation favours niche expansion could however be incorrect. An alternative explanation for the data is that causation is in the opposite direction — an initial expansion into new niches could increase the utility of cooperative genes, making them more prevalent in generalists. For example, mechanisms like metabolic flexibility or bacterial motility could actually be the main drivers for niche expansion, with cooperation emerging as a by-product[23 13,24]. Alternatively, the relationship between bacterial habitat preferences and cooperative gene

carriage might not be causal at all, but instead could be influenced by a third, unidentified variable correlated with both[25,26]. These alternative explanations can only be distinguished between by testing for a causal link between bacterial cooperation and niche breadth evolution.

Another complication is that bacterial genes often experience simultaneous gains or losses. This pattern is frequently driven by coevolution among different genes and associated with microbial adaptation to varied environments[27–29]. For instance, when bacteria encounter a new environment, they may need to acquire new metabolic genes to process novel substances. Since these metabolic genes often interact in pathways, gaining one gene could trigger the acquisition of additional interacting genes[30–32]. Are cooperative genes frequently gained or lost in conjunction with other types of genes? If this is true, the evolutionary consequences could be more complex, as the gain or loss of cooperative genes could be accompanied by the gain or loss of other genes that play a different role in determining bacterial niche breadth.

We tested the causal link between bacterial niche breadth evolution and the prevalence of cooperative genes. We began by characterizing the niche breadths of over 20,000 bacterial species globally, spanning multiple taxonomic groups. We then categorized cooperative genes into five types based on their functional annotations: biofilm formation, quorum-sensing, secretion systems, iron-scavenging siderophores, and antibiotic degradation. To unravel the causal relationship between bacterial niche breadth and the presence of cooperative genes, we employed ancestral state reconstruction techniques. By reconstructing the cooperative gene profiles and habitat preferences of ancestral species, we were able to infer causality by examining the chronological order of relevant evolutionary events. Finally, we extended our analysis to encompass genes that have co-evolved with cooperative genes, aiming to better understand their collective impact on shaping bacterial niche breadths.

Results & Discussion

Characterising habitat generalists and specialists

To globally assess bacterial niche breadth and categorize bacteria as either generalists or specialists, we initially inferred the habitat preferences for each of the 32,262 species in our dataset using their 16S ribosomal RNA (rRNA) gene sequences. These habitat annotations were performed using ProkAtlas, a database containing environmental 16S rRNA gene sequences from 114 distinct habitats like soil, human gut, and seawater[33] (Table S2).

A potential problem with such habitat classifications is that they can bear the imprint of human biases. For example, what we delineate as distinct environments, like freshwater and marine, could be functionally similar habitats for bacteria[34]. To address this, we adopted a previously developed method that grouped similar habitats into "habitat clusters" based on species composition[9]. Under this system, if two habitats, say A and B, are inhabited by the same set of species, they are grouped together as a single habitat cluster. Conversely, if no species are shared between habitats A and B, they are treated as independent habitats. Using this approach, we reorganized the original 114 habitats into 26 habitat clusters (Figure 1a; Table S2). This habitat clustering confirmed some intuitive groupings, such as clustering chicken gut, human gut, and pig gut into a single habitat cluster, but it also revealed some new associations. For example, the method grouped paper pulp and insect gut into a single habitat cluster, suggesting similarities between these two environments in terms of microbial adaptation (Figure 1a).

To assess the niche breadths of each species, we quantified their environmental variability by counting the number of habitat clusters in which they were found (Figure 1b). To differentiate between habitat generalists and specialists, we generated an expected distribution of species' environmental variability using permutation tests (n = 10,000) and compared this to the

observed distribution. Thresholds for classifying species as specialists or generalists were set at points where the observed count of species with a particular level of environmental variability significantly exceeded the expected count (Figure 1b). For specialists, we added another condition that they must also have the narrowest niche breadths. Accordingly, species found in only one habitat cluster were identified as specialists, those found in 8 or more clusters were termed generalists, and those in between were classified as intermediate species (Figure 1b).

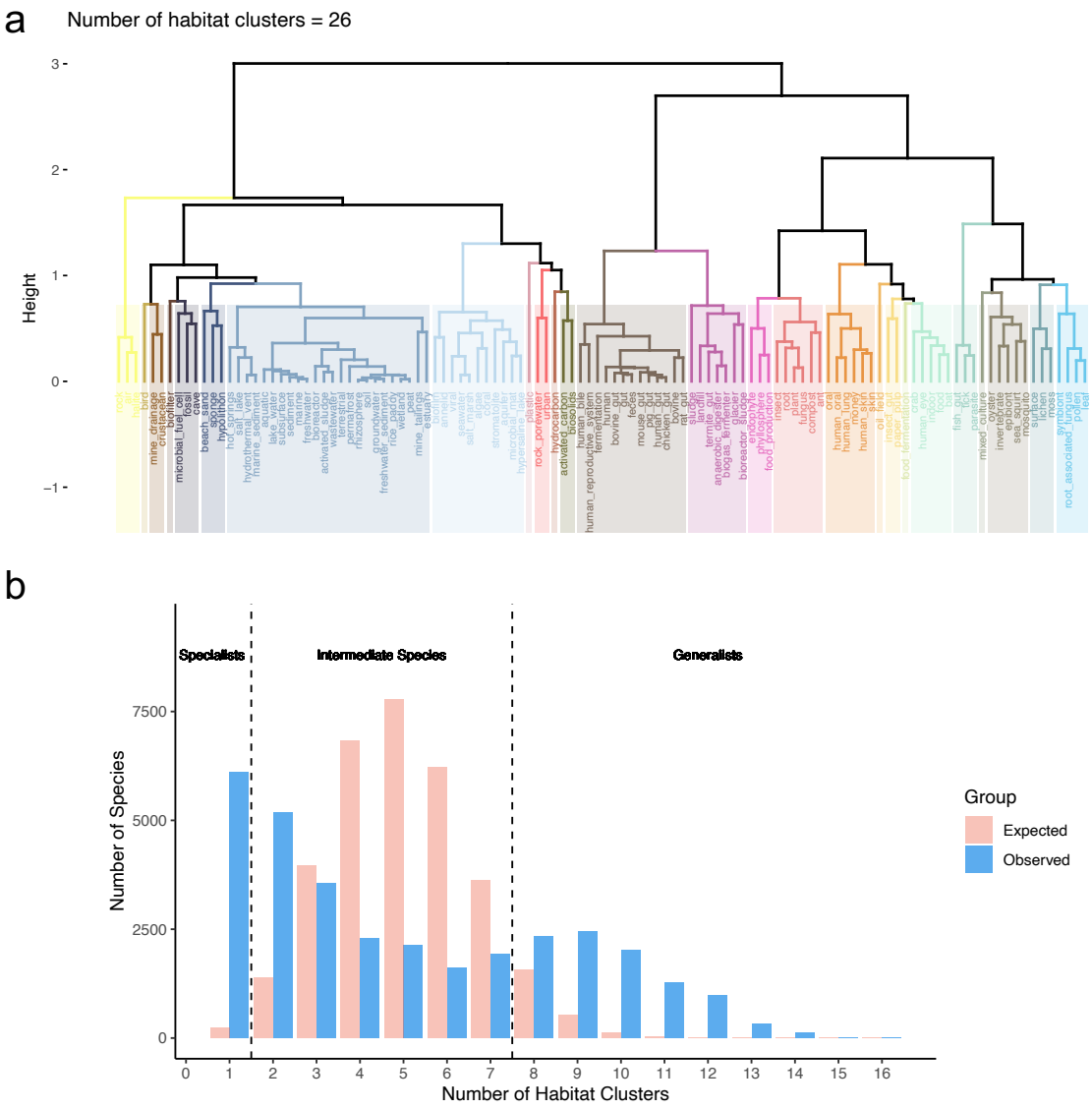


Figure 1. Assessment of bacterial niche breadths. Initially, habitat preferences for each of the 32,262 species in our dataset were inferred using ProkAtlas, which utilizes 16S rRNA gene sequences from 114 distinct habitats. These habitats were subsequently grouped into "habitat clusters" based on similarities in species composition, and species' habitat preference annotations were updated accordingly. (a) The 114 original habitats from the ProkAtlas database were clustered into 26 habitat clusters. (b) Niche breadth for each species was quantified by their environmental variability, defined as the number of habitat clusters in which they were found. The observed environmental variability was then compared against an expected distribution from 10,000 permutations to categorize species as generalists or

specialists. Species present in only one habitat cluster were deemed specialists, those in 8 or more clusters were identified as generalists, and species found in between were classified as having intermediate niche breadth.

The challenge of precisely defining and measuring microbial niche breadths is considerable, primarily due to the complex and often invisible array of environmental and biological conditions that can influence a species' ability to live and reproduce[7]. Classical ecological theories, such as Hutchinson's concept of the "fundamental niche," call for a comprehensive accounting of these conditions — an endeavour that is often impractical in empirical studies[35].

In the practice of microbial ecology, various methods have been proposed to estimate microbial niche breadth, each capturing specific facets of the "fundamental niche" [10,11,36]. Our approach aimed to balance the constraints of data availability with the need for ecological significance. We leveraged 16S rRNA gene sequence data, which is both abundant and readily accessible, to estimate niche breadth on a global scale across various taxonomic groups[37]. In doing so, our definition primarily encapsulated a species' ability to adapt to ecologically distinct environments—a criterion that is both intuitive and relatively straightforward to assess. Alternative approaches often require more specialized data that is not universally available. For instance, some methodologies relied on community composition data, which is often limited to high-quality metagenomic samples[10]. Others necessitated a deep inventory of multi-dimensional abiotic factors, such as pH and temperature or nutrient availability, restricting them to specific study systems[11,38]. Still others require metatranscriptomic data to dissect a species' metabolic niche, a dataset that is frequently inaccessible for many taxa[8,36].

Habitat generalists had higher proportion of cooperative genes than specialists

To explore the relationship between bacterial niche breadth and the prevalence of cooperative genes, we adopted an approach from prior research[39] to classify cooperative genes based on functional annotations. Genes were designated as "cooperative" if they were tagged with one or more of the following cooperative behaviours commonly found in bacteria: biofilm formation, quorum-sensing, secretion systems, siderophore production and usage, and antibiotic degradation. We successfully annotated habitat preferences and identified cooperative genes for 25,785 species, categorizing them into 4,230 specialists, 12,975 intermediate species, and 8,580 generalists. The proportion of cooperative genes within each species' representative genome were calculated to serve as an indicator of the level of cooperative gene carriage for that species, ranging from 0 to 0.05 (Figure S1).

Using MCMC Generalized Linear Mixed Models (MCMCglmm), we found that habitat generalists carried a higher proportion of cooperative genes compared to both specialists and intermediate species (MCMCglmm[40]; n = 25,785 species; generalist vs. intermediate species: pMCMC < 0.001; generalist vs. specialist: pMCMC < 0.001; Figure 2a).

Alternatively, in examining environmental variability (defined by the number of habitat clusters a species occupies) as an indicator of niche breadth, our analysis revealed a positive correlation between environmental variability and the proportion of cooperative genes across species (MCMCglmm; n = 25,785 species; proportion of cooperative genes ~ environmental variability: pMCMC < 0.001; Figure 2b).

In our analysis examining various types of cooperative genes, we consistently observed a significant positive relationship between the proportion of these genes and bacterial niche

breadth (PGLS-ANOVA[41]; n = 25,785 species; antibiotic degradation: F-value = 126.198, P-value < 0.001; biofilm: F-value = 7.526, P-value < 0.001; quorum sensing: F-value = 4.291, P-value = 0.014; secretion system: F-value = 8.573, P-value < 0.001; siderophores: F-value = 21.657, P-value < 0.001; Figure 2c; Table S2).

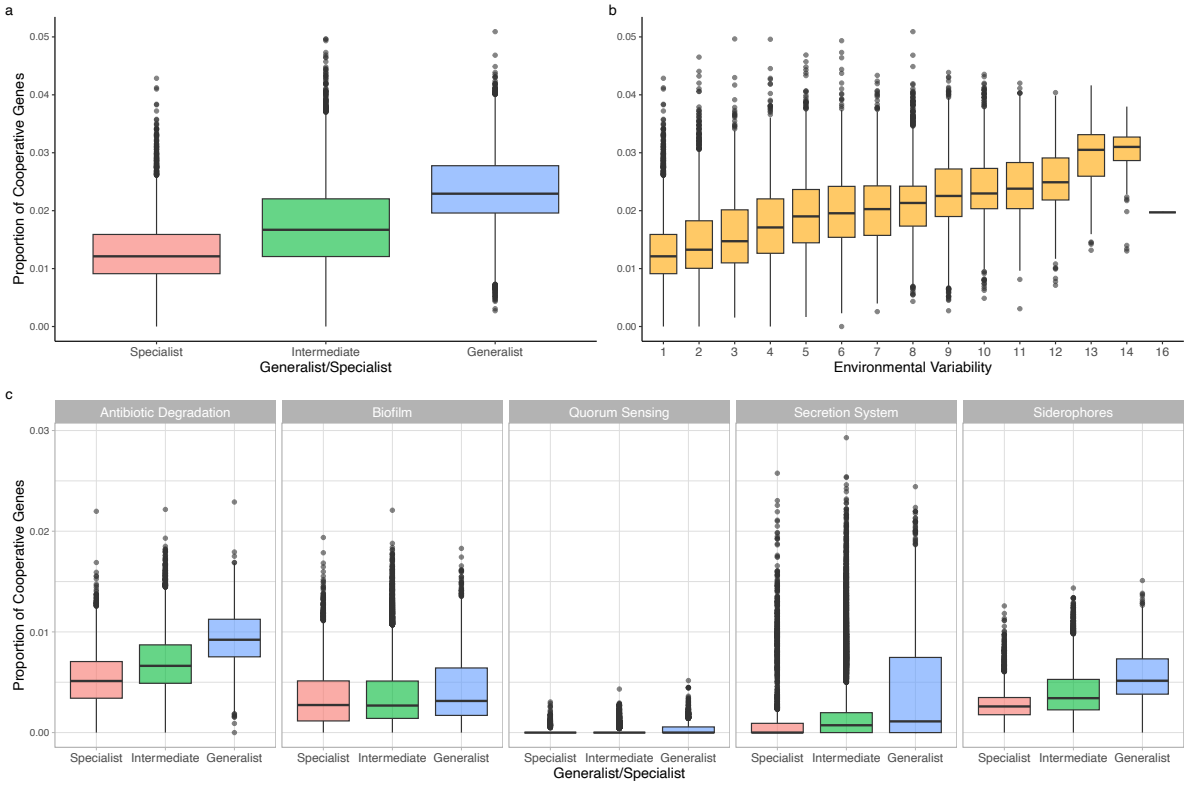


Figure 2. Correlation between cooperative gene proportion and bacterial niche breadth. (a) Habitat generalists consistently exhibited a higher proportion of cooperative genes compared to both specialists and species classified as intermediate. (b) This upward trend was maintained when utilizing environmental variability as a metric for niche breadth. (c) Across varying functional categories of cooperative genes, a consistent positive correlation with bacterial niche breadth was observed.

Our observation that habitat generalists possess a higher proportion of cooperative genes is consistent with earlier studies. For example, one study defined cooperative genes as those encoding extracellular proteins and found a similar pattern[22]. Moreover, research has shown that genes associated with quorum-sensing, biofilm formation, protein secretion systems, and siderophore production are more prevalent in generalist bacteria when their classification is based on the biotic interactions[10].

This pattern can be rationalized by considering the fitness advantages that cooperation offers. First, in a cooperative population, various individual cells can specialize in different tasks or in the acquisition of different resources, thereby allowing the collective to exploit a broader array of resources than any single cell could alone. This makes the species more adaptable to various habitats, aligning with the definition of a generalist. Second, study has proposed that bacteria could use cooperative secretions that modify their environment to extend their host range and infect multiple host species[21]. However, this interpretation assumes a causal relationship, suggesting that cooperation facilitates niche expansion. Alternative causal pathways could also explain the observed correlation. For instance, an initial expansion into new niches could increase the utility of cooperative genes, making them more prevalent in generalists. In such scenarios, other factors like bacterial motility might be the driving force behind the niche expansion, with cooperation being a secondary outcome[13,24]. Therefore, investigating the causal link between cooperation and niche expansion is key to clarifying their roles in microbial evolution.

Exploring the causal link between niche breadth and cooperation

To probe the causal relationship between bacterial niche breadth and the prevalence of cooperative genes, we employed the Granger causality framework, a method that allows for causal inference via the chronological examination of events[42,43]. Ancestral state reconstruction enabled us to recreate the cooperative gene profiles and habitat preferences for ancestral (node) species, based on data from extant (tip) species and their known phylogenetic relationships[41]. We successfully reconstructed the habitat preferences for 24,912 ancestral nodes, which included 3,994 specialists, 11,683 species with intermediate niche breadth, and 9,235 generalists (Figure S2).

Four potential causal scenarios were formulated: (a) an increased number of cooperative genes may lead to a broader niche (generalization); (b) a reduced number of cooperative genes may lead to a narrower niche (specialization); (c) a broader niche may drive the acquisition of more cooperative genes, and (d) a narrower niche may result in the loss of cooperative genes. Each of these scenarios was examined through the lens of evolutionary chronology (Figure 3 a-d). For example, to test the hypothesis that carrying more cooperative genes facilitates generalization, we compared the cooperative gene profiles of specialist ancestral species whose descendants expanded their niche with those whose descendants remained specialists. If the former group exhibits a higher proportion of cooperative genes, it would suggest that an increase in such genes likely preceded the transition towards a broader habitat preference, thereby confirming that causal relationship (Figure 3e).

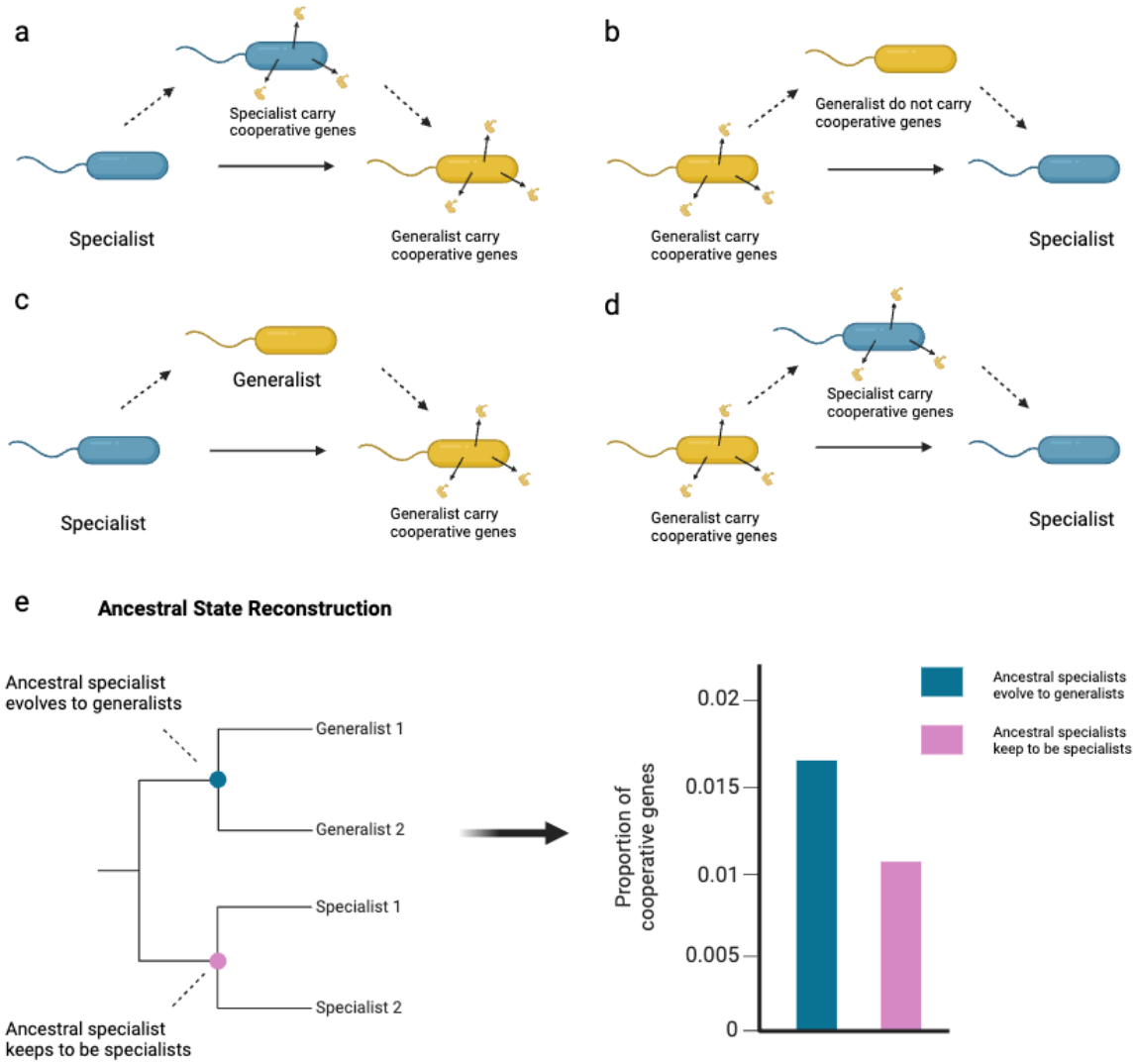


Figure 3. Causality inference using the Granger causality framework: examining the chronological order of evolutionary events. (a-d) Illustration of four potential causal relationships between the carriage of cooperative genes and the evolution of niche breadth, differentiated by the sequence of evolutionary events. Blue cells signify specialists, while yellow cells represent generalists. Cells secreting golden molecules represent those carrying cooperative genes. (e) To test the scenario that a higher proportion of cooperative genes promotes niche expansion, ancestral state reconstruction is employed for both niche breadths and cooperative gene carriage. Support for this causal relationship would come from observing

that specialist ancestors with a higher proportion of cooperative genes more often give rise to descendants that become generalists, as compared to those that remain specialists.

We first investigated whether having a higher proportion of cooperative genes facilitated generalization (Figure 3a). Each specialist ancestral species (parents) had two direct descendants (children), which could be specialists, intermediate species, or generalists. To quantify the niche breadth of each parent's children, we assigned numerical values: specialist = 0, intermediate species = 0.5, and generalist = 1. Summing these for both children yielded possible niche breadth scores of 0 (two specialists), 0.5 (one specialist, one intermediate species), 1 (two intermediate species, or one specialist one generalist), 1.5 (one generalist, one intermediate species), 2 (two generalists). We focused solely on direct descendants to maintain analytical clarity. The proportion of cooperative genes in specialist parents did not significantly affect whether their children expanded their niches (MCMCglmm; n = 3832 specialist ancestral species; 0 vs. 0.5: pMCMC = 0.153; 0 vs 1: pMCMC = 0.229; Figure 4a, Table S1). This suggested that a higher proportion of cooperative genes did not promote generalization.

We then turned our attention to whether a lower proportion of cooperative genes favoured specialization, and found that a lower proportion of cooperative genes facilitated specialization (Figure 3b). Using a similar scoring method for quantifying the niche breadths of the children of generalist ancestral species, we found that generalist parents whose children underwent significant niche contraction had a lower proportion of cooperative genes (MCMCglmm; n = 9173 generalist ancestral species; 1 vs. 0.5: posterior mean = 0.004, pMCMC = 0.003; 1.5 vs 1: posterior mean = 1.005e-04, pMCMC = 0.711; 2 vs 1.5: posterior mean = 0.0004, pMCMC < 2e-04; Figure 4b, Table S1).

Lastly, we simultaneously examined whether generalization promoted an increase in cooperative genes or whether specialization facilitated a decrease in such genes, using a direct regression approach. We found that neither becoming a generalist nor becoming a specialist influenced the proportion of cooperative genes in the offspring, negating causality in both these directions. Specifically, we calculated the change in the proportion of cooperative genes between each ancestral species and the average for its two children. Our analyses showed no significant correlation between the ancestral species' habitat preferences and the change in proportion of cooperative genes in their descendants (MCMCglmm; n = 24590 ancestral species; generalists vs. intermediate species: pMCMC = 0.361; generalists vs specialists: pMCMC = 0.374; Figure 4c, Table S1).

In summary, our analyses suggest that the observed correlation between bacterial niche breadth and the prevalence of cooperative genes was primarily driven by the causality that a lower proportion of cooperative genes facilitated specialization (Figure 3b). We extended our causality analysis to include each specific category of cooperative genes, such as those involved in quorum sensing, biofilm formation, secretion systems, siderophores, and antibiotic degradation. Interestingly, we observed variable patterns across these categories in their role in niche expansion or contraction (see in the Supplementary Material). For example, we found that having a higher proportion of biofilm formation genes actually hindered, not facilitated niche expansion (Figure S3, Table S4). This suggests that cooperative genes with different functions may have distinct influences in shaping or responding to the change in ecological niches occupied by bacterial species.

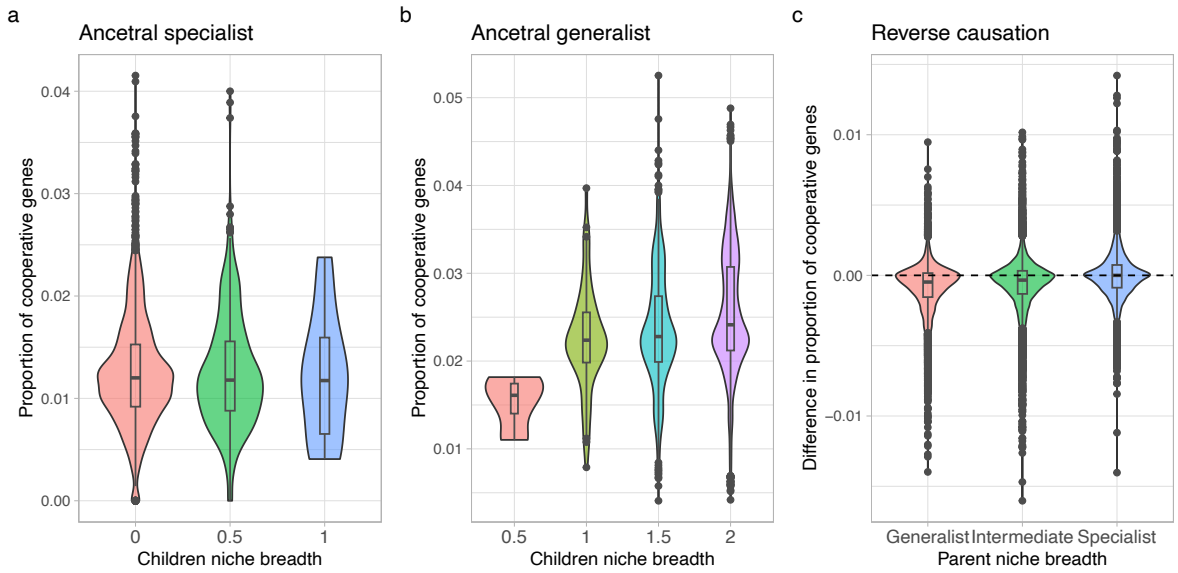


Figure 4. Results of causal inference for various scenarios. (a) The prevalence of cooperative genes in specialist ancestors did not significantly influence the niche expansion of their descendants, indicating that higher levels of cooperative genes are not a driving force for becoming generalists. (b) Generalist ancestors with descendants experiencing substantial niche contraction had fewer cooperative genes, suggesting that a reduced prevalence of cooperative genes might facilitate specialization. (c) No significant relationship was observed between the niche breadth of ancestral species and the change in the proportion of cooperative genes among their descendants, suggesting that transitions to either generalist or specialist do not affect the cooperative gene carriage in the offspring.

Why did we observe a causality where a lower proportion of cooperative genes led to specialization, but not the reverse? We proposed that cooperation serves not as a catalyst for generalization but as a stabilizing factor. While the potential benefits of cooperative genes, such as enhanced resource utilization, might theoretically encourage their rapid acquisition by specialists aiming to expand their niches, the associated costs, such as the energy expenditure for synthesizing these cooperative genes or the vulnerability to exploitation by ‘cheaters’, also

act as selective pressures for their swift loss[18,44,45]. This rapid turnover in the acquisition and loss of cooperative genes may occur too quickly to have a meaningful impact on the longer-term evolutionary transition from being a specialist to a generalist. On the other hand, for existing generalists, cooperative genes act more like a maintenance mechanism that keeps them in their broad ecological niches. Losing these genes undermines this stability, pushing the organism towards specialization. In the absence of direct measurements for the costs and benefits of cooperation, investigating the dynamics of cooperative gene acquisition and loss becomes crucial for testing this hypothesis. Therefore, we proceeded to examine the rates of gain and loss for cooperative genes in the following step.

Cooperative genes only undergo more frequent gains and losses over short terms

Our investigation into the dynamics of cooperative gene gain and loss examined two distinct evolutionary timescales: long-term (relatively macroevolutionary) and short-term (relatively microevolutionary)[46]. Using ancestral state reconstruction, we estimated the average rate of gene gains and losses —referred to as state transitions — for each ortholog group (OG) per phylogenetic branch in the long-term scale. Specifically, gene gains were marked by transitions from absence to presence, while gene losses were denoted by transitions from presence to absence.

We found no significant differences between the average rates of state transitions in cooperative and non-cooperative OGs in the long-term scale (permutated Kruskal-Wallis test; chi-squared = 0.624, p-value = 0.433; Figure 5a). However, we observed distinct patterns when we broke down the cooperative genes by function. Biofilm formation-related OGs had significantly lower rates of state transitions compared to non-cooperative OGs, while siderophore production-related OGs displayed significantly higher rates (Dunns Test; biofilm vs. non-

cooperative: p-value = 0.012; siderophores vs. non-cooperative: p-value = 0.046; Figure 5a). Additionally, cooperative OGs were more prone to gene loss than to gene gain, as demonstrated by their significantly higher rates of gene loss (permutated Kruskal-Wallis test; chi-squared = 390.29, p-value < 1e-04; Figure 5b).

in the short-term scale (relative microevolutionary), where direct estimations of gene gain or loss rates are not feasible, we employed an alternative strategy by leveraging the concept of bacterial pangenomes. A pangenome comprises all the genes found across different strains of a species[47]. Within this, some genes are termed "accessory" because they are only present in a subset of strains, making them more prone to gains and losses as compared to core genes, which are ubiquitous across all strains[30] (Figure 5c). To quantify the susceptibility of cooperative genes to gains and losses, we introduced an index Π, which measures the propensity of a cooperative gene being an accessory gene in a pangenome relative to what would be expected by chance (see in the Methods).

To implement this approach, we reconstructed pangenomes for 171 bacterial species for which at least 100 high-quality genomes were available in GTDB (Figure S6). Our analyses revealed that cooperative genes were significantly more likely to be accessory genes (MCMCglmm; n = 171 species; pMCMC = 0.0004; Figure 5c; Table S1). This trend was notably stronger for genes associated with secretion systems and antibiotic degradation (MCMCglmm; n = 171 species; secretion systems: pMCMC = 0.016; antibiotic degradation: pMCMC = 0.017; Figure 5c; Table S1), but not for other functional categories of cooperative genes.

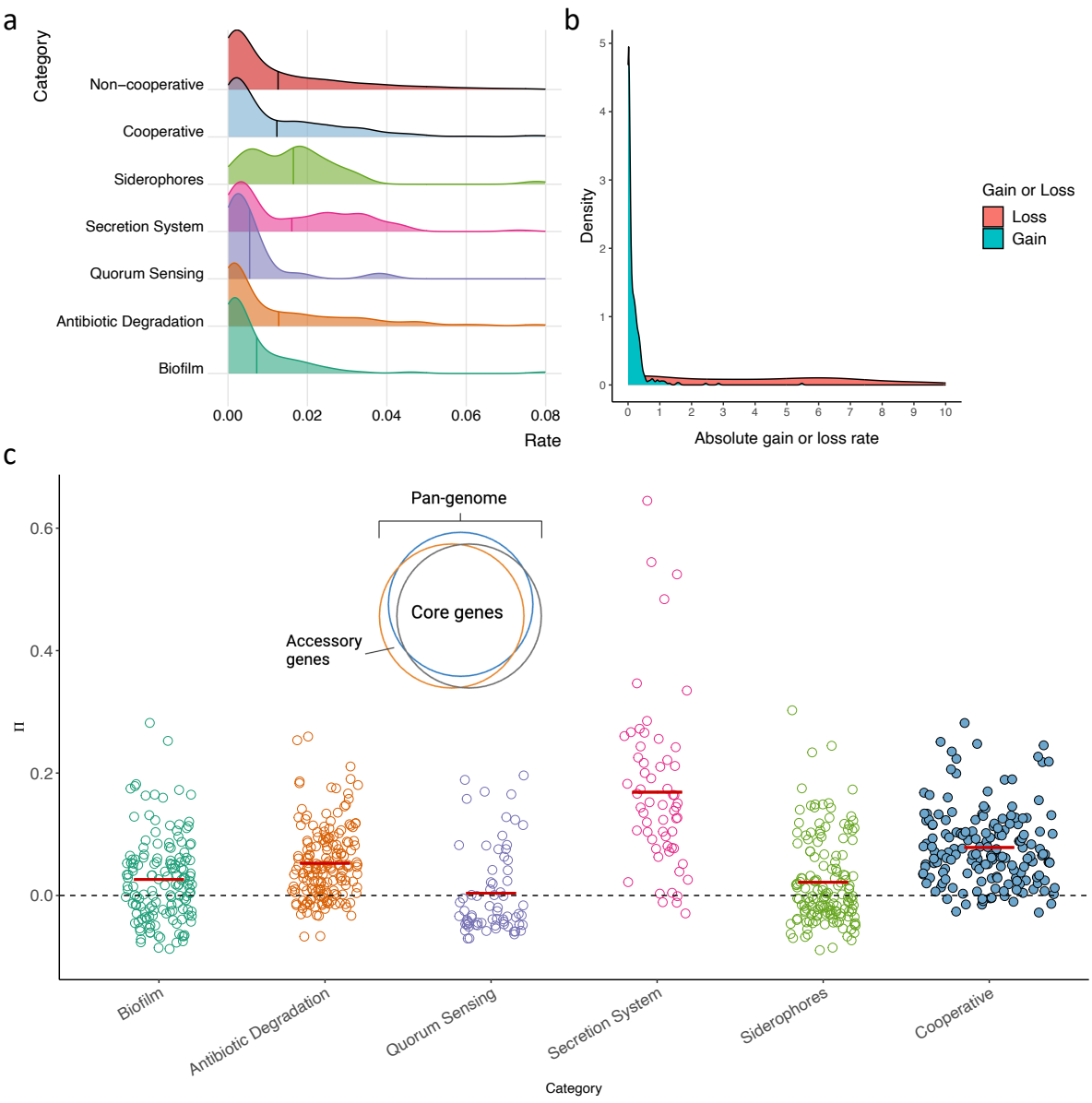


Figure 5. Dynamics of cooperative gene gain and loss. (a) At the macroevolutionary level, the rates of gene gain and loss were represented by average state transition rates. Cooperative and non-cooperative orthologous groups (OGs) showed no significant differences in these rates. And cooperative OGs with distinct functions exhibited varying rates of gene gain and loss. (b) The cooperative OGs exhibited higher loss rates compared to their gain rates. (c) At the microevolutionary level, the rates of gene gain and loss were indicated by the propensity of a cooperative gene being an accessory gene in a pangenome, compared to random expectation.

Cooperative genes were found to be significantly more likely to serve as accessory genes, suggesting a higher frequency of short-term gain and loss events for these genes.

Our results offer empirical support for our initial hypothesis by revealing distinct temporal dynamics in the turnover of cooperative genes. Specifically, we found that while cooperative genes don't exhibit increased rates of gain or loss on long-term scales, they are subject to rapid gains and losses on short-term scales. This observation aligned well with our hypothesis that the swift turnover of cooperative genes, driven by a balancing act between costs and benefits, occurred too quickly to substantially affect long-term evolutionary shift from being specialists to generalists. This insight could clarify why we did not observe that a greater proportion of cooperative genes facilitated the transition to generalization. Furthermore, our data revealed that cooperative genes are more frequently lost than gained, which reinforced the direction of causality we identified that the loss of these genes could serve as a driving force for specialization. This finding was consistent with previous research suggesting that gene loss rates generally exceed gain rates[32,44,48].

Identification of 357 KOs that co-evolved with cooperative KOs

In exploring the role of cooperative genes in bacterial habitat preferences, it's crucial to consider that genes often work in concert with others[49]. Accordingly, we expanded our focus to include genes that co-evolved with cooperative genes. We defined co-evolved genes (KOs) as those sharing similar phylogenetic distribution patterns with cooperative genes (KOs) across the tree of life. This analysis resulted in identifying 357 co-evolved KOs corresponding to 372 cooperative KOs[50] (Table S8). To elucidate the reasons behind the co-evolution of these KOs with cooperative KOs, we examined two hypotheses: the functional association hypothesis and the habitat selection hypothesis.

The functional association hypothesis suggests that these co-evolved KOs participate in similar metabolic pathways as the cooperative KOs, implying that the gain or loss of one gene could directly affect the other via functional association[28]. To assess this, we calculated the Jaccard Index between the KEGG metabolic pathway vectors of cooperative and co-evolved KOs and utilized permutation tests to establish the statistical significance of the similarity. Our results revealed that cooperative KOs are involved in 28 metabolic pathways, while their co-evolved counterparts are involved in 132 pathways. With a Jaccard Index of 0.11 and a permutation P-value of 0.958, we found no significant functional similarity between the two KO sets, thereby refuting the functional association hypothesis (Figure 6a).

Next, we turned our attention to the habitat selection hypothesis, which proposes that these co-evolved KOs are favoured along with cooperative KOs by similar habitats[32]. To investigate this, we calculated the average Spearman's rank correlation coefficient for the habitat preference vectors of both the cooperative and co-evolved KOs. We used a permutation testing approach, similar to the one employed for the functional association hypothesis, to assess statistical significance. The analysis revealed a high Spearman's rank correlation coefficient of 0.957 between the habitat preference vectors, with a permutation P-value of 0 (Figure 6b). This finding suggested that cooperative KOs and their co-evolved counterparts are likely favoured by similar habitats, thus supporting the habitat selection hypothesis.

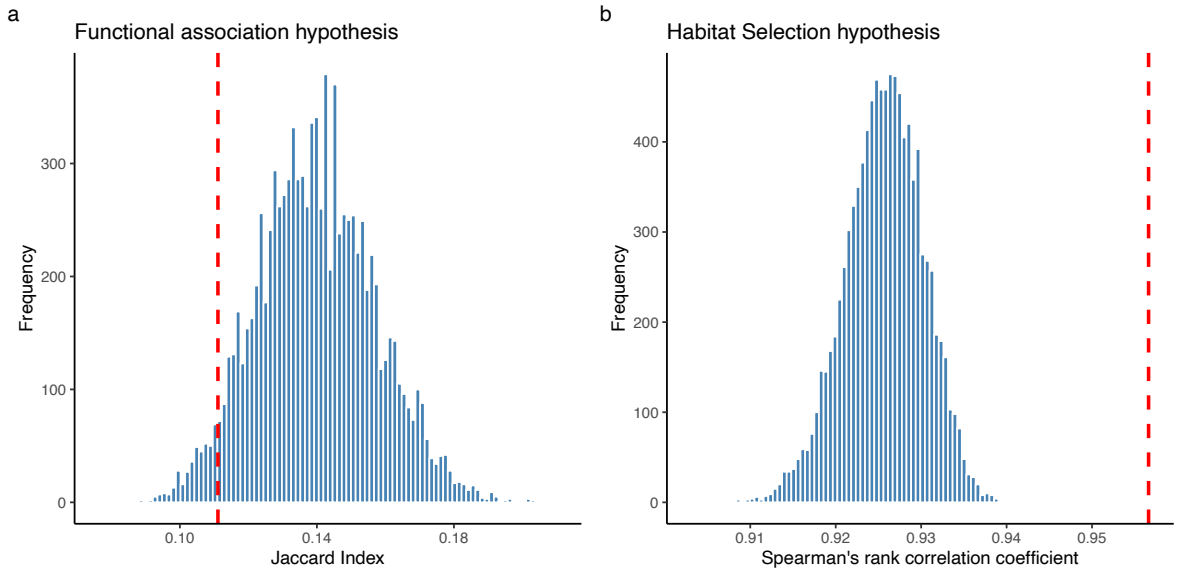


Figure 6. Testing hypotheses for co-evolution with cooperative genes. (a) The functional association hypothesis posits that co-evolved KOs participate in similar metabolic pathways to the cooperative KOs. The Jaccard Index between the KEGG metabolic pathway vectors for cooperative and co-evolved KOs (indicated by the red dashed line) was not significantly higher than expected, rejecting this hypothesis. (b) The habitat selection hypothesis suggests that both co-evolved and cooperative KOs are selected by similar habitats. The average Spearman's rank correlation coefficient for the habitat preference vectors of both cooperative and co-evolved KOs (indicated by a red dashed line) was significantly higher than expected, lending support to this hypothesis.

The significant correlation between the habitat preferences of cooperative and co-evolved genes suggest that these sets of genes might collectively contribute to the bacteria's ability to survive and thrive in specific habitats, thereby affecting the long-term habitat preferences and habitat transition of bacteria. Given these findings, future steps should aim to investigate the roles of both cooperative and co-evolved genes in the evolution of bacterial niche breadth, to

gain a more comprehensive understanding of how bacterial habitat preferences was influenced by the prevalence of cooperative genes and their co-evolved genes.

Co-evolved genes play opposite roles in shaping bacterial niche breadth

We began our analysis by assessing the individual influence of co-evolved genes on bacterial niche breadth. Our analysis revealed that the proportion of co-evolved genes in the representative genomes of species ranged from 0.11 to 0.71, making them more prevalent than cooperative genes (Figure S7).

Our analysis then uncovered substantial variation in the proportion of co-evolved genes among bacterial species with different niche breadths (PGLS-ANOVA; n = 25,785 species; proportion of co-evolved genes ~ generalist, intermediate, or specialist: F-value = 276.562, P-value < 0.001; Figure 7a; Table S2). Specifically, we found that generalist species had fewer co-evolved genes in comparison to both intermediate and specialist species (MCMCglmm; n = 25,785 species; generalist vs. intermediate species: pMCMC < 0.001; generalist vs. specialist: pMCMC < 0.001; Figure 7a; Table S1). This pattern is in contrast to what we observed for cooperative genes (Figure 2a). Furthermore, we found a negative correlation between environmental variability and the proportion of co-evolved genes across species (MCMCglmm; n = 25,785 species; proportion of co-evolved genes ~ environmental variability: pMCMC < 0.001; Figure 7b; Table S1).

Our analysis demonstrated that cooperative and co-evolved genes play diverging roles in shaping bacterial niche breadth, thereby underscoring the necessity of examining their interactive effects. Therefore, we examined how the proportion of co-evolved genes might modulate the relationship between cooperative genes and bacterial niche breadth. Given the

different scales at which two types of genes are represented in the genome, we normalized the data for a more equitable comparison.

Focusing on environmental variability as a proxy for niche breadth, we discovered that both types of genes exert significant yet opposite independent effects. Specifically, the proportion of cooperative genes positively correlated with environmental variability, whereas the proportion of co-evolved genes showed a negative correlation. Notably, the absolute value of the coefficient for the proportion of co-evolved genes was higher, indicating they impose a stronger influence on niche breadth compared to cooperative genes. Furthermore, we observed a significant negative interaction between the two types of genes, suggesting that these types of genes counteract each other's influence when jointly shaping the bacterial niche (MCMCglmm; n = 25,785 species; proportion of cooperative genes ~ environmental variability: posterior mean = 0.118, pMCMC < 0.001; proportion of co-evolved genes ~ environmental variability: posterior mean = -0.810, pMCMC < 0.001; interaction term: posterior mean = -0.063, pMCMC < 0.001; Figure 7b; Table S1).

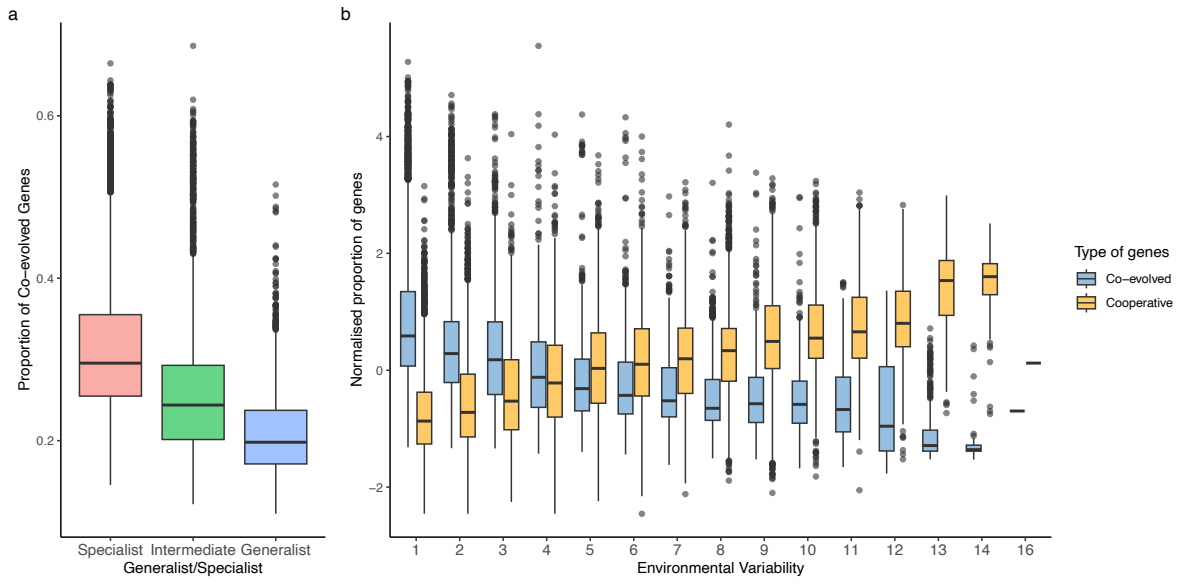


Figure 7. Opposing patterns of co-evolved and cooperative genes. (a) In contrast to cooperative genes, habitat generalists showed a lower proportion of co-evolved genes compared to both specialists and intermediate species. (b) This negative correlation persisted when environmental variability was used as a measure for niche breadth.

These findings underscored the complexity in how cooperative and co-evolved genes collectively influence bacterial niche breadth. The counteracting effects revealed by our analysis point to an intriguing biological dynamic: the presence of co-evolved genes seems to mitigate the positive impact of cooperative genes on niche expansion. This finding set the stage for further steps using causal inference methodologies to delve deeper into the mechanisms by which co-evolved genes limit the scope of bacterial habitats.

Causal inference of co-evolved genes

We extended our causal inference approach, previously applied to cooperative genes, to examine the causal relationship between the proportion of co-evolved genes and niche breadth. Specifically, we first assessed whether a lower proportion of co-evolved genes in specialist parents facilitates generalization in their offspring. Our results showed that the proportion of co-evolved genes in specialist parents did not significantly impact the niche breadth of their offspring, indicating that a reduced presence of co-evolved genes did not appear to promote generalization (MCMCglmm; n = 3,832 specialist ancestral species; pMCMC = 0.929; Figure 8a, Table S1). Additionally, when we considered the potential interactive effects between cooperative and co-evolved genes, our analysis did not reveal any significant individual or interactive effects.

In the context of evolutionary transitions from generalists to specialists, we examined whether a higher proportion of co-evolved genes in generalist parents facilitates specialization in their offspring. We observed that a higher proportion of co-evolved genes in the generalist parents was associated with a narrower niche breadth in their offspring (MCMCglmm; n = 9,173 generalist ancestral species; posterior mean = -10.530, pMCMC < 0.001; Figure 8b, Table S1). Moreover, when analysing the potential interactive effects of cooperative and co-evolved genes, we found a significant negative interaction between the two. Both types of genes were found to have significant but opposite independent effects on the niche breadth of offspring, and the absolute value of the coefficient for the proportion of co-evolved genes was greater than that for cooperative genes (MCMCglmm; n = 9173 generalist ancestral species; ancestral proportion of cooperative genes ~ children niche breadth: posterior mean = 0.099, pMCMC = 0.012; ancestral proportion of co-evolved genes ~ children niche breadth: posterior mean -0.557, pMCMC < 0.001; interaction term: posterior mean = -0.117, pMCMC < 0.001; Figure 8b; Table S1).

Finally, we sought to discern whether generalization led to a decrease in co-evolved genes or whether specialization facilitated an increase in these genes, using a direct regression approach. However, we did not find any significant evidence to support a causal relationship in either direction (MCMCglmm; n = 24,590 ancestral species; ancestral niche breadth ~ proportion of co-evolved genes in children: pMCMC = 0.154; Figure 8c, Table S1). Additionally, when examining the potential interactive effects between cooperative and co-evolved genes in this context, our analysis failed to detect any significant individual or interactive influences.

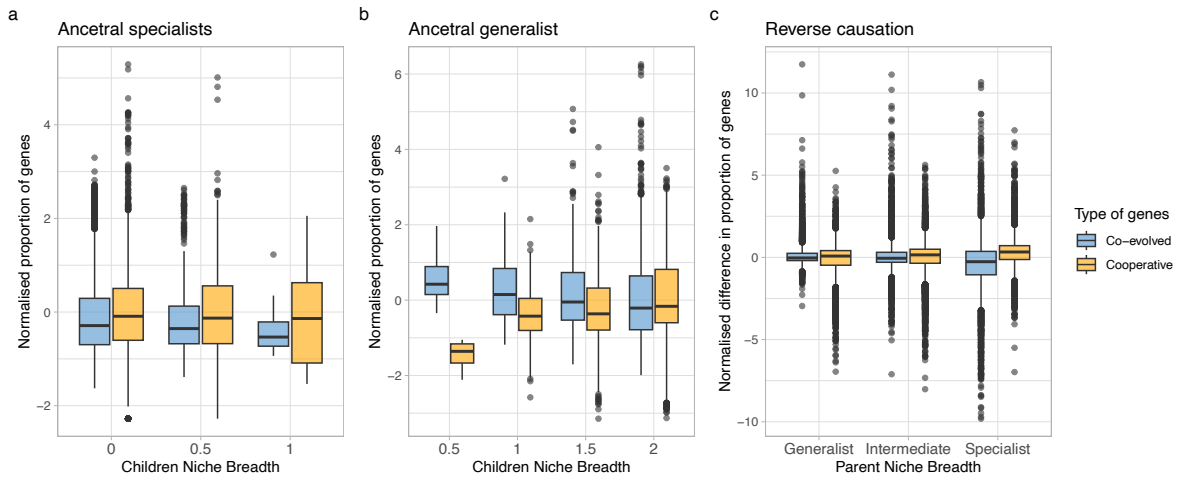


Figure 8. Causal inference for co-evolved genes. (a) The proportion of co-evolved genes in specialist ancestral species did not significantly affect the trend of their descendants becoming generalists, suggesting that co-evolved genes do not play role in facilitating niche expansion. (b) In contrast to cooperative genes, a higher proportion of co-evolved genes in generalist ancestors was associated with descendants that experienced niche contraction, indicating that an increased prevalence of co-evolved genes might contribute to specialization. (c) No significant correlation was found between the ancestral species' niche breadth and the proportion of co-evolved genes in their descendants, implying that the transition between specialist and generalist states does not influence the co-evolved gene carriage in the offspring.

In summary, our causal inference analyses shed light on the influential role of co-evolved genes in bacterial niche breadth evolution. We found that a higher proportion of co-evolved genes in generalist parents corresponded to a narrower niche in their offspring, indicating that co-evolved genes predominantly function in narrowing niche breadth of existing generalists. Interestingly, when we studied both cooperative and co-evolved genes, a significant negative interaction was observed when generalist species evolved into specialists. This suggested that co-evolved genes counterbalance the influence of cooperative genes in such evolutionary shifts.

Ultimately, both cooperative and co-evolved genes contribute to the evolutionary trajectory from generalists to specialists, albeit in opposite directions, with co-evolved genes exerting a stronger impact.

Conclusion

Our study has provided key insights into the causality between bacterial niche breadth and the prevalence of cooperative genes. We've demonstrated that a lower proportion of cooperative genes drives the habitat transition towards specialization (Figure 4). While one might expect a higher prevalence of cooperative genes to drive generalization, the rapid turnover of these genes, due to a balance of associated costs and benefits, may be too swift to impact long-term shifts from specialization to generalization. Evidence for this came from our observation that cooperative genes, while not necessarily showing elevated gains or losses over long-term evolutionary scales, experience rapid gains or losses in shorter timescales (Figure 5). Additionally, we identified 357 ortholog groups that were gained or lost simultaneously with cooperative ortholog groups, primarily driven by overlapping habitat preferences (Figure 6). By assessing the combined effects of cooperative genes and their co-evolved counterparts on bacterial niche evolution, we found that the presence of co-evolved genes tends to temper the advantageous influence of cooperative genes in promoting niche expansion, and co-evolved genes primarily act to constrict the niche breadth of extant generalists (Figure 7 & 8). These results deepen our comprehension of niche breadth evolution in bacteria. For future endeavours, a detailed examination of bacterial cooperation, coevolution, and habitat preferences will further illuminate the patterns of bacterial diversification and enhance our understanding of bacterial evolution.

Methods

Species selection and genome collection

We obtained a list of species and their corresponding genomes from the Genome Taxonomy Database (GTDB[51]; release 207) on 1st October 2022. At the first step, we selected species with representative genomes in GTDB that enabled the identification of full-length 16S rRNA gene sequences, yielding a list of 38,474 species. To ensure high-quality of genomes, we applied a Minimum Information about a Metagenome-Assembled Genome (MIMAG) criterion[52], considering genomes with completeness ≥ 95% and contamination ≤ 5% as high-quality genomes. We then used ncbi-genome-download scripts (version 0.3.1) to download the available amino acid sequences for all genomes (files named as "*_protein.faa.gz," where the asterisk represents an arbitrary string) from either GenBank[53] or RefSeq[54] database using the assembly entries specified in GTDB. This resulted in a dataset of 237,354 genomes with corresponding amino acid sequences for subsequent analysis.

Habitat preferences annotation

The annotation of habitat preferences followed an approach modified from a previous study[9]. The procedure comprised three steps: (a) inferring the preliminary habitat preferences for each species using their 16S rRNA gene sequences; (b) clustering the preliminary habitats based on species compositions; (c) re-annotating the habitat preferences of each species by replacing the preliminary habitats with the newly defined habitat clusters.

a) Preliminary habitat estimation

To estimate the preliminary habitat preferences for each species, we first retrieved the full-length 16S rRNA gene sequences for their representative genomes from GTDB (https://data.gtdb.ecogenomic.org/releases/release207/207.0/genomic_files_reps/bac120_ssu_

reps_r207.tar.gz). We then used BLAST to search these representative full-length 16S rRNA sequences against the 16S rRNA sequences in metagenome datasets containing various environmental information through the ProkAtlas online platform[33]. The sequence similarity threshold was set to 97%, and the sequence coverage threshold was set to 150bp for conducting the ProkAtlas annotation. A species was considered present in an environment if its habitat preference score, reported by ProkAtlas, was greater than zero.

b) Clustering habitats based on species compositions

To avoid potential bias arising from human intuition in the preliminary habitat definitions, we adopted a more objective biological approach by clustering habitats based on their species assemblages[9]. This method aimed to identify situations where two habitats, though differently defined by humans, might actually host the same microbial species. For instance, if two habitats (A and B) have identical sets of species, they would be clustered together. Conversely, if habitats A and B share no species, they would be considered independent habitats.

We initially collected a comprehensive dataset of environmental 16S rRNA sequences from the ProkAtlas database[33]. This dataset comprises 361,474 sequence fragments derived from 114 distinct habitats (see Table S2 for the list of ProkAtlas habitats). To determine the species-level habitat preferences of these 16S rRNA fragments, we then mapped the fragmental sequences to full-length 16S rRNA sequences available in the SILVA database[37] (https://www.arb-silva.de/fileadmin/silva_databases/release_138.1/Exports/SILVA_138.1_SSURef_NR99_tax_silva.fasta.gz). For the mapping process, we utilized VSEARCH[55] (v2.22.1_linux_x86_64) to perform an all-against-all BLAST search between the ProkAtlas and SILVA sequences, setting the sequence identity threshold at 98%. As a result, we successfully matched 248,295 out of 361,474 ProkAtlas fragmental sequences to 71,574 full-length 16S rRNA sequences present in

the SILVA database. Each of these full-length sequences corresponded to a unique bacterial species. This enabled us to create presence/absence profiles of bacterial species associated with every ProkAtlas habitat.

To evaluate the similarities between ProkAtlas habitats and prepare for subsequent clustering steps, we calculated environmental similarity scores based on a previous study[9]. Specifically, for any pair of habitats H_1 and H_2, and for a sequence identity threshold T (ranging from 70% to 98%), the similarity score between H_1 and H_2 was calculated as the geometric mean of two proportions: (i) The proportion of full-length 16S rRNA sequences (species) in H_1 that were similar to any sequence (species) in H_2 with a sequence identity above the threshold T; (ii) The proportion of full-length 16S rRNA sequences in H_2 that were similar to any sequence in H_1 with a sequence identity above the threshold T. The sequence matching was performed with BLASTn (version 2.13.0) in an all-against-all mode between each pair of sequence sets at the threshold T. We allowed T to vary because species do not necessarily need to be identical (T = 98%) to exhibit similar habitat preferences. Bacterial species from higher taxonomic ranks might possess adaptive traits that enable them to thrive in the same environments, even with some genetic divergence[56]. As a result of this process, we obtained matrices of similarity scores between each pair of ProkAtlas habitats, each calculated at different sequence identity thresholds.

Finally, we conducted hierarchical clustering to group ProkAtlas habitats using the dissimilarity score (1- similarity score) at different thresholds T. To objectively determine the optimal clustering methods (between “ward.D”, “ward.D2”, “single”, “complete”, “UPGMA”, “WPGMA”, “WPGMC”, or “UPGMC”), the number of clusters (range from 2 to 113), and the sequence identity threshold (range from 70% to 98%), we calculated the Silhouette index

for various clustering results. The Silhouette index ranges from −1 to +1, with a higher value indicating better clustering performance, where objects are well-matched to their own cluster and poorly matched to neighbouring clusters. We found the optimal Silhouette index was achieved when we using "ward.D2" method with a threshold T = 77% and 26 habitat clusters (Fig S8). To visualize the clustering result, the dendrogram of this optimal clustering can be checked in Fig 1a. The analyses here was performed in R (version 4.2.1), using package *NbClust*[57] and package *Factoextra*[58].

c) Re-annotating the habitat preferences

Using the 26 newly generated environmental clusters, we replaced the previous 114 habitats, allowing us to re-annotate habitat preferences for species in our dataset. This re-annotation provided a more refined representation of species' habitat preferences, based on a relatively objective clustering of similar environments. We successfully annotated 32,262 species from our initial list with 25 environmental clusters. We calculated the environmental variability of each species as the number of environmental clusters in which they were found. This measurement of species environmental variability allowed us to understand the range of ecological niches that each species can inhabit. The distribution of species' environmental variability is depicted in Figure 1a.

Definition of habitat generalists/specialists

To identify habitat generalists and specialists, we compared the observed distribution of species' environmental variability to an expected distribution generated through permutations (Figure 1b). We conducted 10,000 random permutations of the species-environmental cluster associations, preserving the species count in each cluster, to establish the expected environmental variability distribution (Figure 1b). Cut-offs for defining specialists and

generalists were set where the observed species count at a particular environmental variability significantly exceeded the expected count. We also sought to identify habitat specialists as species occupying the minimum number of environmental clusters. As a result, we set cut-offs at environmental variabilities of 1 and 8 for specialists and generalists, respectively. Species with environmental variability equal to 1 were defined as specialists, those found in 8 or more environmental clusters as generalists, and those in between as intermediate species (Figure 1b).

Defining cooperative genes

We adapted a method from a previous study[39] to define cooperative genes based on functional annotations. Specifically, genes were considered "cooperative" if they were annotated with at least one of the five well-known forms of bacterial cooperative behaviours (biofilm formation, quorum-sensing, secretion systems, siderophores production and usage, and antibiotic degradation). We employed the KEGG Orthology (KO) database[59] for functional annotations and used KofamScan[60] to assign KO identifiers to protein sequences.

To create a list of "cooperative KO identifiers", we first generated a list of cooperative keywords by referring to prior studies[39]. We then downloaded the complete KO identifier list (https://www.genome.jp/ftp/db/kofam/ko_list.gz) and filtered it based on these keywords. This list was manually curated to ensure specificity and relevance to prokaryotes, resulting in a final list of 407 cooperative KO identifiers (biofilm, 86; quorum sensing, 24; secretion systems, 86; siderophores, 44; and antibiotic degradation, 167; Table S7).

Subsequently, we calculated the proportion of cooperative genes for each genome. In calculating the proportions, a potential issue arose when some genes lack confident functional annotations, which could influence the accuracy of the estimates. To address this, we directly

used KOs to represent gene content, specifically considering only genes that could be matched with at least one significant KO identifier. As a result, the proportion of cooperative genes in each genome was calculated as the ratio of the number of cooperative KOs to the total number of KOs. At this stage, 25,785 species had both generalist/specialist annotations and calculated proportions of cooperative genes. Our primary analysis focused on this subset of species.

Phylogeny

The phylogenetic reference tree of bacterial species was downloaded from GTDB[51] (release 207, https://data.gtdb.ecogenomic.org/releases/release207/207.0/bac120_r207.tree). To focus on the 25,785 species relevant to our study, we pruned the phylogenetic tree using the "*keep.tip*" function in the R package "*ape*" [61]. The dendrogram of the pruned tree, displaying the evolutionary relationships among the selected species, can be found in the supplementary material (Figure S2).

Ancestral state reconstruction for habitat preference and gene content

To infer causality and calculate gene gain/loss rates, we estimated ancestral habitat preferences and gene content for each species' representative genome. For habitat preference, extant species had three discrete states: generalist, intermediate, and specialist. We applied an empirical Bayesian method (the marginal posterior probability approximation (MPPA) method with the F81-like model) in PastML[62] (version 1.9.15) to reconstruct ancestral states for habitat preferences of each internal node in our species' phylogeny. Of the 25,784 internal nodes, 9,235 were estimated as generalists, 3,994 as specialists, and 11,683 as intermediates. Ancestral states for 872 nodes remained undetermined due to uncertainty of the method.

For gene content, we used a similar approach to estimate the ancestral presence or absence of each ortholog group (OG) for every internal node in the phylogeny. The extant species' OG status was determined through KEGG Orthologs (KO). Consequently, the reconstructed "genome" for each internal node (ancestral species) consisted of all "present" OG (KO)s.

Causality inference

We applied the Granger causality framework to infer causal relationships between two evolutionary events[42]. Specifically, if events A and B are correlated and event A precedes event B, we would suggest that A "Granger causes" B. For instance, if carrying more cooperative genes (Event A) is found to precede becoming a generalist (Event B), we could propose that a higher proportion of cooperative genes may facilitate the transition to a generalist habitat preference (Figure 3).

To determine the chronological order of these evolutionary events, we compared the states of ancestral nodes (parents) with their immediate descendants (children). For instance, to investigate whether possessing more cooperative genes promotes the transition to being a generalist, we examined the proportion of cooperative genes in specialist parents whose descendants broaden their niche versus those whose descendants remain specialists. If the former group displays a higher proportion of cooperative genes, it would suggest that the increased presence of cooperative genes likely preceded the transition to a more generalist habitat preference. In such a case, we could propose that having more cooperative genes facilitates the evolution toward a generalist habitat preference (Figure 3).

Estimation of gene gain and loss rates

We analysed the rate of gene gains and losses on two evolutionary timescales: relative macroevolutionary (long-term) and relative microevolutionary (short-term). On the relative macroevolutionary scale, the rate of gains or losses for each ortholog group (OG) was calculated as the average rate of state changes per phylogenetic branch — either presence (coded as 1) to absence (coded as 0) or vice versa. Gene gains were indicated by zero-to-one transitions, while gene losses were signified by one-to-zero transitions. These state change rates were determined through ancestral state reconstruction using PastML[62] (version 1.9.15).

To assess whether cooperative genes experienced more frequent gains or losses compared to non-cooperative genes, we examined the difference in average rates of state changes per branch between cooperative and non-cooperative OGs. We further computed the absolute rates of gene gains and losses as the difference between the gene gain rate and the gene loss rate across various species. These rates were estimated using the "*asr_mk_model*" function in the R package "*castor*", employing an "*ARD (all rates different)*" rate model[63].

At the relative microevolutionary level, direct estimation of gene gain or loss rates is unfeasible. We used the extent to which a gene is categorized as accessory in a pangenome can to serve as an indirect indicator of the rate of gene gains or losses on a microevolutionary scale, because accessory genes are more prone to being gained or lost compared to core genes[47]. We reconstructed pangenomes for species that had at least 100 high-quality genomes in GTDB, resulted in a list comprising 171 species (Figure S6). The typical process of pangenome clustering can be broken down into three primary steps: 1) gene prediction and annotation, 2) homologous genes identification, 3) categorizing genes as core or accessory based on their presence or absence across different genomes[64,65]. In our study, we employed the protocol

outlined in the KO database for gene prediction and ortholog identification[59]. KofamScan[60] was used to assign KO identifiers to protein sequences, thereby grouping genes with identical KO identifiers into gene families. Consequently, the presence or absence profile of each gene family across the genomes of each species was represented by the presence or absence of each KO.

To assess whether cooperative genes are more likely to be accessory genes in pangenomes, we introduced an index Π to gauge the extent to which cooperative genes are accessory. For each species, we randomly selected two strains to create a minimum pangenome based on these pairwise genomes. We then estimated π, which represents the difference between the observed and expected proportions of accessory cooperative genes. The π is,

$$\pi = \frac{N(accessory, cooperative)}{N(cooperative)} - \frac{N(accessory)}{N(total)}$$

Here, $N(accessory, cooperative)$ refers to the number of cooperative KOs that are accessory, $N(cooperative)$ is the total number of cooperative KOs, $N(accessory)$ is the number of accessory KOs, and $N(total)$ is the total number of KOs in this minimum pangenome. This index allows us to quantify how much more (or less) likely a cooperative gene is to be an accessory gene compared to what would be expected by chance in the minimum pangenome.

The index Π is then denoted as the average π calculated across all possible minimum pangenomes consisting of two genomes:

$$\Pi = \frac{\sum_{i=1}^{\eta} \pi}{\eta},$$

where $\eta = \binom{N}{2} = \frac{N(N-1)}{2}$, which is the total number of unique combinations of two genomes that can be chosen from N available strains for a species. The higher the value of Π for cooperative genes, the more they tend to be accessory genes within bacterial pangenomes, indicating a greater frequency of gains and losses.

For species with an extensive number of genomes, such as *Escherichia coli* which has 25,981 genomes in our dataset, the computation of π for all selected genome pairs would be resource-intensive. To address this, we applied Central Limit Theorem (CLT) to estimate Π for these species. This approach involved three steps: 1) randomly sample 20 genomes of a species for 10,000 times, 2) calculating the Π for each sample, and 3) using the mean value of all the sample Π as an estimation for Π of the species. According to the CLT, the average of sample means will converge to the population mean when the sample size is sufficiently large.

Identification of genes that co-evolved with cooperation genes

To pinpoint genes that may have co-evolved with cooperative genes, we adopted the PhyloCorrelate tool to identify OGs that share similar phylogenetic distribution patterns with cooperative OGs throughout the tree of life[50]. For each cooperative KO, we computed the runs-adjusted-Jaccard coefficient (rJC) to quantify the similarity in the presence/absence patterns between every non-cooperative KO across the phylogenetic tree. Additionally, we calculated a hypergeometric P-value (rHyperP) to assess the statistical significance of the similarity between two runs-adjusted phylogenetic profiles. The 'runs-adjustment' method served as a heuristic strategy to mitigate phylogenetic redundancy, effectively eliminating overrepresented patterns within specific lineages[66].

We then focused on KOs that exhibited a statistically significant co-occurrence with cooperative KOs (rHyperP < 0.001). It should be noted that a non-cooperative KO can co-occur with multiple cooperative KOs. We analysed the distribution of the number of cooperative KOs with which each non-cooperative KO co-occurred and identified an enrichment threshold at 270 cooperative KOs (Figure S9). Consequently, we defined KOs as having co-evolved with cooperative KOs if they were significantly co-occurred (rHyperP < 0.001) with at least 270 cooperative KOs. This analysis resulted in a list of 357 KOs that likely co-evolved with cooperative KOs, as summarized in Table S8.

Exploring reasons for co-evolution with cooperative genes

To delve into the reasons why certain KOs have likely co-evolved with cooperative KOs, we formulated two hypotheses: the functional association hypothesis and the habitat selection hypothesis.

a) Functional association hypothesis

This hypothesis posits that the 357 identified KOs co-evolved with cooperative KOs due to their involvement in similar metabolic pathways. In such circumstances, the gain or loss of one gene might directly influence the gain or loss of another gene via functional associations[28]. To test this, we sourced metabolic pathway data for both the cooperative and co-evolved KOs from the KEGG PATHWAY Database[42]. We used the KEGG API (https://rest.kegg.jp/link/ko/pathway) to get the associations data between KOs (denoted by KO identifiers starting with "ko") and their corresponding pathways (as represented by pathway identifiers starting with "map"). Subsequently, we computed the Jaccard Index to compare the pathway vectors of cooperative KOs and their co-evolved KOs. This step aimed to assess whether these two groups of KOs are involved in similar metabolic pathways. We

then calculated the P-value using a permutation method. Specifically, we computed Jaccard indices between the pathway vector of cooperative KOs and the pathway vectors of 10,000 sets of 357 randomly selected KOs, thereby creating a null distribution. The p-value was then calculated as the proportion of these permuted Jaccard indices that exceeded the observed Jaccard index between the pathway vectors of cooperative KOs and their co-evolved KOs.

b) Habitat selection hypothesis

This hypothesis proposes that the 357 identified KOs co-evolved with cooperative KOs due to their shared habitat preferences[32]. To test this hypothesis, we constructed a map linking KO presence/absence to habitat clusters based on the existing maps of species-KO presence/absence associations and species-habitat cluster associations. We then calculated the average Spearman's rank correlation coefficient between the habitat preference vectors of cooperative KOs and their co-evolved KOs to assess the degree of similarity in their habitat selections.

To estimate statistical significance, we used a permutation method similar to the one applied for the functional association hypothesis. However, to achieve a biologically unbiased null distribution, we shuffled the species-habitat cluster associations map for each round of permutation. Then, we calculated the average Spearman's rank correlation coefficients between the habitat preference vectors of cooperative KOs and those of 357 randomly selected non-cooperative KOs. The P-value was determined based on this permuted distribution.

Statistics

a) PGLS-ANOVA

We used the phylogenetic generalized linear model (PGLS) to determine if cooperative gene proportions differ among generalists, specialists, and intermediate species[41]. As these groups represent different niche breadth categories, we applied PGLS-ANOVA to investigate these variations. F-statistics and P-values from all PGLS-ANOVA models were reported.

b) MCMCglmm

R package *MCMCglmm* allows us to fit generalized linear mixed-effects models (GLMMs) using a Markov chain Monte Carlo approach with a Bayesian statistical framework[40]. To account for potential phylogenetic nonindependence among the species in our dataset, we incorporated the phylogeny as a random effect in the model. To ensure that the phylogenetic tree used in the MCMCglmm fitting had an ultrametric structure, we applied the "*force.ultrametric*" function in the R package "*phytools*" (with an option *"method = extend"*) to transfer our tree into an ultrametric form [67]. To assess the goodness of fit for each model, we reported the *pMCMC* value (referred to as "p-value").

References

1. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences* **104**, 11436–11440 (2007).
2. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
3. Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).
4. Lima-Mendez, G. *et al.* Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).

5. Wu, L. *et al.* Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat Microbiol* **4**, 1183–1195 (2019).
6. Sexton, J. P., McIntyre, P. J., Angert, A. L. & Rice, K. J. Evolution and Ecology of Species Range Limits. *Annual Review of Ecology, Evolution, and Systematics* **40**, 415–436 (2009).
7. Sexton, J. P., Montiel, J., Shay, J. E., Stephens, M. R. & Slatyer, R. A. Evolution of Ecological Niche Breadth. *Annual Review of Ecology, Evolution, and Systematics* **48**, 183–206 (2017).
8. Muller, E. E. L. Determining Microbial Niche Breadth in the Environment for Better Ecosystem Fate Predictions. *mSystems* **4**, 10.1128/msystems.00080-19 (2019).
9. Sriswasdi, S., Yang, C. & Iwasaki, W. Generalist species drive microbial dispersion and evolution. *Nat Commun* **8**, 1162 (2017).
10. von Meijenfeldt, F. A. B., Hogeweg, P. & Dutilh, B. E. A social niche breadth score reveals niche range strategies of generalists and specialists. *Nat Ecol Evol* **7**, 768–781 (2023).
11. Hernandez, D. J., Kiesewetter, K. N., Almeida, B. K., Revillini, D. & Afkhami, M. E. Multidimensional specialization and generalization are pervasive in soil prokaryotes. *Nat Ecol Evol* **7**, 1408–1418 (2023).
12. Carscadden, K. A. *et al.* Niche Breadth: Causes and Consequences for Ecology, Evolution, and Conservation. *The Quarterly Review of Biology* **95**, 179–214 (2020).
13. Jaffe, A. L., Castelle, C. J. & Banfield, J. F. Habitat Transition in the Evolution of Bacteria and Archaea. *Annual Review of Microbiology* **77**, null (2023).
14. Griffin, A. S., West, S. A. & Buckling, A. Cooperation and competition in pathogenic bacteria. *Nature* **430**, 1024–1027 (2004).
15. Häse, C. C. & Finkelstein, R. A. Bacterial extracellular zinc-containing metalloproteases. *Microbiological Reviews* **57**, 823–837 (1993).
16. Hale, T. L. Genetic basis of virulence in Shigella species. *Microbiological Reviews* **55**, 206–224 (1991).
17. Hall-Stoodley, L., Costerton, J. W. & Stoodley, P. Bacterial biofilms: from the Natural environment to infectious diseases. *Nat Rev Microbiol* **2**, 95–108 (2004).

18. West, S. A., Griffin, A. S., Gardner, A. & Diggle, S. P. Social evolution theory for microorganisms. *Nat Rev Microbiol* **4**, 597–607 (2006).
19. West, S. A. & Cooper, G. A. Division of labour in microorganisms: an evolutionary perspective. *Nat Rev Microbiol* **14**, 716–723 (2016).
20. West, S. A., Cooper, G. A., Ghoul, M. B. & Griffin, A. S. Ten recent insights for our understanding of cooperation. *Nat Ecol Evol* **5**, 419–430 (2021).
21. McNally, L., Viana, M. & Brown, S. P. Cooperative secretions facilitate host range expansion in bacteria. *Nat Commun* **5**, 4594 (2014).
22. 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).
23. Chen, Y.-J. *et al.* Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. *ISME J* **15**, 2986–3004 (2021).
24. Di Rienzi, S. C. *et al.* The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *eLife* **2**, e01102 (2013).
25. Dillard, J. R. & Westneat, D. F. Disentangling the Correlated Evolution of Monogamy and Cooperation. *Trends in Ecology & Evolution* **31**, 503–513 (2016).
26. Cornwallis, C. K. *et al.* Cooperation facilitates the colonization of harsh environments. *Nat Ecol Evol* **1**, 1–10 (2017).
27. Arnold, B. J., Huang, I.-T. & Hanage, W. P. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol* 1–13 (2021) doi:10.1038/s41579-021-00650-4.
28. Tassia, M. G., Whelan, N. V. & Halanych, K. M. Toll-like receptor pathway evolution in deuterostomes. *Proceedings of the National Academy of Sciences* **114**, 7055–7060 (2017).
29. Whelan, F. J., Hall, R. J. & McInerney, J. O. Evidence for Selection in the Abundant Accessory Gene Content of a Prokaryote Pangenome. *Molecular Biology and Evolution* **38**, 3697–3708 (2021).
30. Domingo-Sananes, M. R. & McInerney, J. O. Mechanisms That Shape Microbial Pangenomes. *Trends in Microbiology* **29**, 493–503 (2021).

31. Hall, R. J. *et al. Gene-gene relationships in an Escherichia coli accessory genome are linked to function and mobility*. 2021.03.26.437181 https://www.biorxiv.org/content/10.1101/2021.03.26.437181v1 (2021) doi:10.1101/2021.03.26.437181.
32. Konno, N. & Iwasaki, W. Machine learning enables prediction of metabolic system evolution in bacteria. *Science Advances* **9**, eadc9130 (2023).
33. Mise, K. & Iwasaki, W. Environmental Atlas of Prokaryotes Enables Powerful and Intuitive Habitat-Based Analysis of Community Structures. *iScience* **23**, 101624 (2020).
34. Coelho, L. P. *et al.* Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256 (2022).
35. Hutchinson, G. E. Concluding Remarks. *Cold Spring Harb Symp Quant Biol* **22**, 415–427 (1957).
36. Malard, L. A. & Guisan, A. Into the microbial niche. *Trends in Ecology & Evolution* **0**, (2023).
37. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**, D590–D596 (2013).
38. Opulente, D. A. *et al.* Genomic and ecological factors shaping specialism and generalism across an entire subphylum. 2023.06.19.545611 Preprint at https://doi.org/10.1101/2023.06.19.545611 (2023).
39. Simonet, C. & McNally, L. Kin selection explains the evolution of cooperation in the gut microbiota. *PNAS* **118**, (2021).
40. Hadfield, J. D. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software* **33**, 1–22 (2010).
41. Revell, L. J. & Harmon, L. J. *Phylogenetic Comparative Methods in R*. (Princeton University Press, 2022).
42. Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **37**, 424–438 (1969).
43. Detto, M. *et al.* Causality and Persistence in Ecological Systems: A Nonparametric Spectral Granger Causality Approach. *The American Naturalist* **179**, 524–535 (2012).
44. Nogueira, T. *et al.* Horizontal Gene Transfer of the Secretome Drives the Evolution of Bacterial Cooperation and Virulence. *Current Biology* **19**, 1683–1691 (2009).

45. Kramer, J., Özkaya, Ö. & Kümmerli, R. Bacterial siderophores in community and host interactions. *Nat Rev Microbiol* **18**, 152–163 (2020).
46. Rolland, J. *et al.* Conceptual and empirical bridges between micro- and macroevolution. *Nat Ecol Evol* **7**, 1181–1193 (2023).
47. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat Microbiol* **2**, 1–5 (2017).
48. Press, M. O., Queitsch, C. & Borenstein, E. Evolutionary assembly patterns of prokaryotic genomes. *Genome Res.* **26**, 826–833 (2016).
49. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101–113 (2004).
50. Tremblay, B. J.-M., Lobb, B. & Doxey, A. C. PhyloCorrelate: inferring bacterial gene–gene functional associations through large-scale phylogenetic profiling. *Bioinformatics* **37**, 17–22 (2021).
51. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* **50**, D785–D794 (2022).
52. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**, 725–731 (2017).
53. Benson, D. A. *et al.* GenBank. *Nucleic Acids Research* **41**, D36–D42 (2013).
54. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733–D745 (2016).
55. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
56. Silby, M. W., Winstanley, C., Godfrey, S. A. C., Levy, S. B. & Jackson, R. W. Pseudomonas genomes: diverse and adaptable. *FEMS Microbiology Reviews* **35**, 652–680 (2011).
57. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software* **61**, 1–36 (2014).

58. A, K. Factoextra: extract and visualize the results of multivariate data analyses. *R Package Version* **1**, (2016).
59. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457–D462 (2016).
60. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
61. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
62. Ishikawa, S. A., Zhukova, A., Iwasaki, W. & Gascuel, O. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Molecular Biology and Evolution* **36**, 2069–2085 (2019).
63. Louca, S. & Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* **34**, 1053–1055 (2018).
64. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
65. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020).
66. Cokus, S., Mizutani, S. & Pellegrini, M. An improved method for identifying functionally linked proteins using phylogenetic profiles. *BMC Bioinformatics* **8**, S7 (2007).
67. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**, 217–223 (2012).

PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Hao C, Dewar AE, West SA, Ghoul M. 2022 Gene transferability and sociality do not correlate with gene connectivity. *Proc. R. Soc. B* **289**: 20221819. https://doi.org/10.1098/rspb.2022.1819

Received: 13 September 2022
Accepted: 8 November 2022

Subject Category:
Evolution

Subject Areas:
evolution, microbiology, bioinformatics

Keywords:
gene connectivity, horizontal transfer, plasmid mobility, cooperative genes

Author for correspondence:
Chunhui Hao
e-mail: chunhui.hao@biology.ox.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6302904.

THE ROYAL SOCIETY PUBLISHING

Gene transferability and sociality do not correlate with gene connectivity

Chunhui Hao, Anna E. Dewar, Stuart A. West and Melanie Ghoul

Department of Biology, University of Oxford, Oxford OX1 3SZ, UK

CH, 0000-0002-5634-3446

The connectivity of a gene, defined as the number of interactions a gene's product has with other genes' products, is a key characteristic of a gene. In prokaryotes, the complexity hypothesis predicts that genes which undergo more frequent horizontal transfer will be less connected than genes which are only very rarely transferred. We tested the role of horizontal gene transfer, and other potentially important factors, by examining the connectivity of chromosomal and plasmid genes, across 134 diverse prokaryotic species. We found that (i) genes on plasmids were less connected than genes on chromosomes; (ii) connectivity of plasmid genes was not correlated with plasmid mobility; and (iii) the sociality of genes (cooperative or private) was not correlated with gene connectivity.

1. Introduction

Genes interact with one another through the associations between their protein products. Such protein products and their interactions form a protein–protein interaction (PPI) network, where the proteins are the nodes, and the interactions are the edges [1–3]. The connectivity of a gene, usually defined as the number of links a gene's product has to other genes' products, is one of the most elementary characteristics of a gene in its corresponding PPI network [2,4–6]. Gene connectivity varies significantly among different genes. For instance, in an *Escherichia coli* PPI network, the most connected gene has 175 interactions with other genes, while 785 genes have only one interaction [7,8]. In many cases, such variations reflect differences in the essentiality of genes. Genes with higher connectivity are often more essential for the reproductive success of a cell or organism [4,5,9–11].

In prokaryotic microorganisms, gene essentiality may not be the only factor linked to gene connectivity. Prokaryotic genes are often subjected to frequent horizontal gene transfer [12–16], and one therefore might expect that this could also affect their gene connectivity. It has been suggested that highly connected genes should not be carried in parts of the genome that can undergo horizontal gene transfer. This is because if a highly connected gene was transferred to a new host, it would likely be non-functional without the other genes it relies on. This idea, known as the 'complexity hypothesis', has been supported by several studies, which found an inverse correlation between the rate of horizontal gene transfer and gene connectivity [17–20]. In these studies, the rate of horizontal gene transfer was estimated by inferring gene gain and loss events from phyletic patterns, and so combined genes that could be gained or lost in several different ways, such as duplications, deletion and inversions [17–20].

An alternative approach is to compare genes which are maintained or transferred in different ways. We focus on the comparison between genes which are housed on the chromosome, and so likely to have low rates of horizontal gene transfer, versus genes which are housed on plasmids, which represent key vectors of horizontal gene transfer [12,13,21]. Plasmids are self-replicating genetic structures, many of which can move between cells horizontally via a process

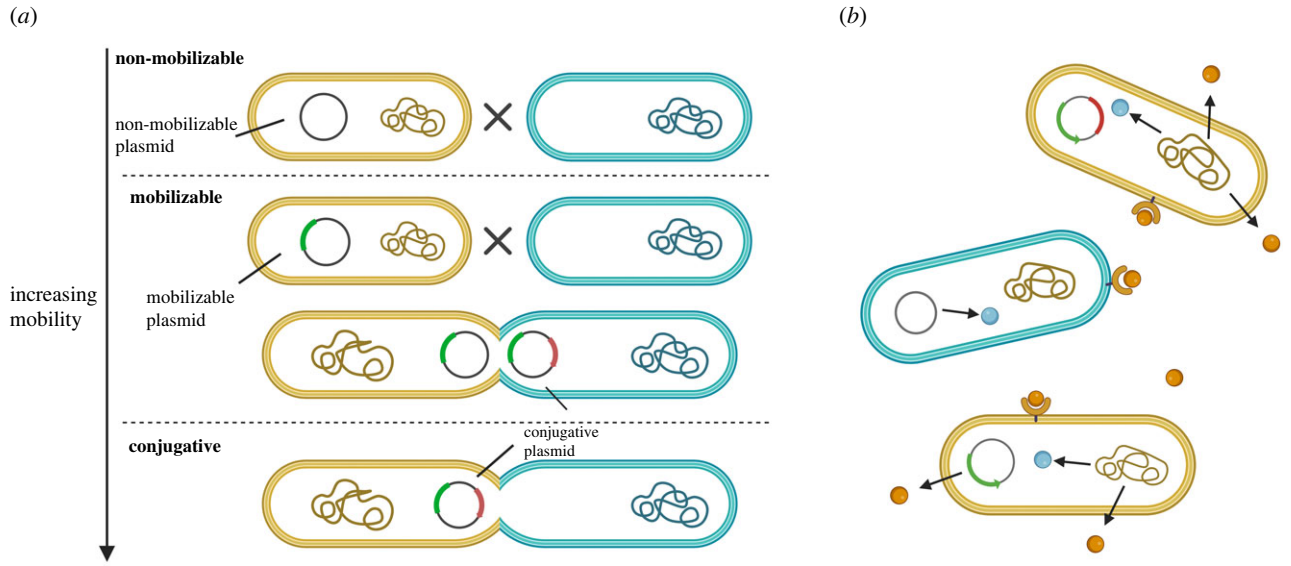


Figure 1. Plasmid mobility and extracellular proteins. (*a*) Plasmid mobility. There are three mobility types of plasmids: non-mobilizable (lowest or no mobility); mobilizable (intermediate mobility) and conjugative (highest mobility). Yellow cells are plasmid donors, while blue cells are plasmid recipients. Each section shows when plasmid transfer can be performed for one of the three plasmid mobility types. Non-mobilizable plasmids cannot be transferred via conjugation. Mobilizable plasmids cannot be transferred alone but can 'hijack' the machinery produced by conjugative plasmids. Conjugative plasmids can be transferred independently. (*b*) Extracellular and intracellular proteins. Orange molecules are proteins, which be released outside of cells (extracellular proteins); blue molecules are proteins that only act within cells (intracellular proteins). Yellow cells produce both extracellular proteins and intracellular proteins, while blue cell only produces intracellular proteins, but can receive extracellular proteins produced by other cells. Extracellular proteins provide benefits not only to the cells that produced them, but also to their neighbours (public goods). Created with BioRender.com. (Online version in colour.)

called conjugation, in addition to vertically via offspring. However, not all plasmids can transfer via conjugation, and so plasmids can be divided into three broad mobility types: non-mobilizable (lowest or no mobility); mobilizable (intermediate mobility) and conjugative (highest mobility) (figure 1*a*) [22,23]. Conjugative plasmids carry all the genes necessary for their transfer [24]. Mobilizable plasmids cannot be transferred alone, but they carry enough genes to 'hijack' the machinery of a conjugative plasmid in the same cell [22]. Non-mobilizable plasmids cannot be transferred by conjugation, but only by transduction and transformation like all genes in the genome [22,25].

Another possible factor that may influence gene connectivity is gene function. If certain types of genes are more likely to be transferred horizontally, these types of genes are less likely to have high connectivity, since high connectivity will decrease the likelihood that they are functional when they enter new hosts. For example, it has been suggested that genes which code for extracellular proteins (public goods) are more likely to be transferred horizontally [26,27]. This prediction can arise for two reasons. First, extracellular factors, though not all of them, could provide benefits to other cells, not just those that produced them, and so can represent cooperative helping traits (figure 1*b*). Horizontal gene transfer could favour cooperation, by allowing cooperative genes to reinfect 'cheats' that don't produce the extracellular factors (non-cooperators) [28–31]. Second, extracellular factors can allow adaptation to different environments, favouring the ability to gain and/or lose them in different environments. For both scenarios, the complexity hypothesis would predict that genes coding for extracellular proteins may have particularly low connectivity [32].

We carried out an across-species comparative analyses, examining connectivity in 140 genomes from 134 prokaryote species. We asked the following three questions. (i) Are chromosomal genes more connected than plasmids genes? (ii) Do genes on plasmids with higher transfer rates have lower connectivity? (iii) Does gene connectivity vary across genes coding for the production of extracellular versus intracellular factors, and does this vary depending upon whether a gene is on a plasmid or the chromosome?

2. Methods

(a) Network data collection

We extracted PPI networks from the STRING database version 11.0 [33] (https://string-db.org/) and used PPI networks to calculate connectivity. We chose STRING because it covers a large number of organisms (5090), allowing for across-species comparative analysis. In addition, STRING is a comprehensive PPI network database: unlike other databases based on either experimental [34–37], or computational prediction interactions [38], STRING integrates both of these and includes direct (physical) and indirect (functional) associations. This allowed us to include as many prokaryotic species as possible in our analyses.

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. Proteins can functionally interact without touching, such as when a transcription factor helps to regulate the expression and production of another protein, or when two enzymes exchange a specific substrate via diffusion [33]. Such indirect interactions could be inferred, for example, from co-expression or co-occurrence patterns between genes. For each pair of interactions, a separate score is given per channel. A combined confidence score ranging from 0 to 1000 is given by combining and adjusting the scores from the different channels [39]. The larger the threshold, the higher the confidence score, but it also means fewer proteins, interactions and species. In our main analysis, we specified a threshold of 400 for the combined scores of the interactions, meaning any interaction below

this threshold would not be considered. 400 is a medium confidence threshold according to the STRING database. We chose this threshold for our main analysis to gain a balance between confidence and sample size. To check the reproducibility of our results, we also repeated our analysis by setting three other thresholds: 150 (low confidence), 700 (high confidence) and 900 (highest confidence). The results at different thresholds are presented in the electronic supplementary material, tables. To match with other databases, we retrieved all the available PPI networks by using the STRINGdb package (version 2.4.0) in R [33] (see 'Database Matching and Genome Collection' below).

(b) Categorization of genes and annotations of replicons

To select genes that were putatively 'cooperative', we followed the methods of previous studies which have considered genes coding for extracellular proteins as a proxy for 'cooperative' genes [40–42]. This is because extracellular proteins often act as public goods, whose benefits are shared with neighbouring cells [27]. Although not all cooperative genes produce extracellular proteins and not all extracellular proteins are cooperative, any strong effect of sociality is likely to be captured by using this proxy [32]. We determined protein subcellular localization for each protein included in our analysis with PSORTdb 4.0 (https://db.psort.org/) [43]. PSORTdb was selected for its reliability and validity in systematically deducing both bacterial and archaeal protein subcellular localization.

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 [43]. For Gram-negative bacteria, proteins are assigned to one of five locations, where the cell wall has been replaced by the 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 analysed.

The PSORTdb outputs we used also included whether each gene was carried on a plasmid or a chromosome. We initially collected precomputed PSORTdb results for all available genomes, including 73 136 replicons belonging to 8416 bacterial and archaeal strains, and kept 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.

(c) Database matching and genome collection

To compare the connectivity (see below) 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. We used NCBI Entrez Direct Command Line Tools (https://www.ncbi.nlm.nih.gov/books/NBK179288/) to transfer NCBI taxonomy IDs used by STRING, to RefSeq genome/replicon accessions used by PSORTb. By doing so, we were able 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. A total of 1570 strains were retrieved originally, of which 462 strains had gene connectivity data on chromosomes and plasmids, and 167 strains had connectivity data on genes encoding extracellular proteins on chromosomes and plasmids. To make sure all species in our dataset were unique, we only included species with a complete Latin binomial name. This gave us a list of 134 species (140 genomes), which included 5 archaeal species (5 genomes) and 129 bacterial species (135 genomes), with 358 plasmid genomes in total (electronic supplementary material, tables S1 and S2).

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 2.6.1 [33]. This unique 'STRING_id' was used to calculate the connectivity for every individual gene. Genes that could not be mapped with 'STRING_id' were not included in our dataset. Details on the number of genes per genome (strain) per species included in this analysis could be found in the electronic supplementary material, table S8.

(d) Connectivity

The complexity hypothesis suggests that highly connected genes are less likely to undergo horizontal transfer, because if a highly connected gene was transferred to a new host, it would probably be non-functional without the other genes it relies on. To examine the connectivity of genes, we used the term 'gene connectivity' to mean the same as the 'protein connectivity' of its protein product in a PPI network. We followed previous studies by defining protein connectivity as the number of PPIs in which the protein is embedded in the PPI network [17,44]. 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 functional when transferred to a new host [20].

The connectivity of each gene was different in networks with different thresholds. Because when a specific network threshold was used, any PPIs below this threshold were not considered, resulting in some genes being omitted if their protein products did not interact with any other proteins above this threshold. As the network threshold increased, the number of genes and interactions included in the analyses decreased. The connectivity of each gene was also reduced. Details on the number of genes per genome per species included in our analyses of networks with different thresholds, and the average connectivity of all genes per genome per species could be found in the electronic supplementary material, table S8.

Network size (the total number of proteins in a PPI network) could influence gene connectivity [45]. Genes with the same connectivity have different impacts in networks of different sizes. To control for the possible influence of network size in our analyses, we also examined whether network size was correlated with the connectivity of genes in our dataset. We performed all calculations of connectivity using the R package 'igragh' [46].

(e) Predicting plasmid mobility and host range

To predict the mobility of every plasmid in our dataset, we used the MOB-typer tool of the software MOB-suite [47]. This tool is designed to provide *in silico* predictions of the origin of transfer (oriT), relaxase type and mate-pair formation (MPF) type for each plasmid based on its sequence. Afterwards, each plasmid is assigned to one of three mobility types: (i) conjugative, where the plasmid contains the complete set of genes and DNA features needed for transfer; (ii) mobilizable, where the plasmid encodes either a relaxase or an oriT but is missing the MPF marker and (iii) non-mobilizable, where plasmid is missing a relaxase and an oriT [47]. We classified the mobility of 358 plasmids for subsequent analysis (electronic supplementary material, table S2). MOB-suite also provided information on plasmid's host range, which 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 is found.

For example, a plasmid found only in genomes of *Yersinia* sp. would have a host range of 'genus', while a plasmid found in a number of Gammaproteobacteria species would have a host range of 'class'. In general, the higher the taxonomic rank of genomes carrying the plasmid, the larger the plasmid's host range. Each plasmid was assigned one of six plasmid's host ranges: genus, family, order, class, phylum and multi-phyla (electronic supplementary material, table S2).

(f) Statistics

We carried out all statistical analyses and graph plotting in R (v. 4.0.2). For all comparisons between groups that included all our species, we used the R package MCMCglmm [48]. MCMCglmm fits generalized linear mixed-effects models (GLMMs) using a Markov chain Monte Carlo approach under a Bayesian statistical framework [48]. Species share traits descended from their common ancestor, and so cannot be considered as independent data points. We thus used MCMCglmm to control for this, with a phylogeny as a random effect in our models (see 'Phylogeny' below) [49]. For each analysis, we used 1 100 000 model iterations with a starting burn-out phase of 100 000, 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 our main analyses using methods described in [50,51]. DIC is a hierarchical modelling generalization of the Akaike information criterion, which balances model fit and model complexity simultaneously. Like other information criteria, smaller values of DIC are preferred [48].

(g) Phylogeny

To control for phylogenetic relationships between our species, we used a phylogenetic tree including all 134 species in our dataset (electronic supplementary material, figure S6). We put together this phylogeny using the methods of a recent study [32]. The tree was based on a recently published maximum-likelihood tree of life using 16 ribosomal protein sequences data [52]. 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' [53]. 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 [54]. Where there were more than two species within a genus in our dataset, we consulted phylogenies from the literature to add any within-genus clustering of species' branches.

3. Results

(a) Chromosomal genes are more connected than plasmid genes

We first compared the gene connectivity between chromosomal genes and plasmids genes. For our main analysis using networks with a medium threshold of confidence, we found that genes located on chromosomes had significantly higher levels of connectivity compared to genes on plasmids (figure 2). Specifically, across species, the difference in connectivity between chromosomal genes and plasmid genes was significantly different from zero (MCMCglmm [48]; posterior mean = 15.127, 95% CI = 12.061 to 18.169, pMCMC < 0.001, n = 134 species, R^2 of phylogeny = 0.201; figure 2; electronic supplementary material, table S3).

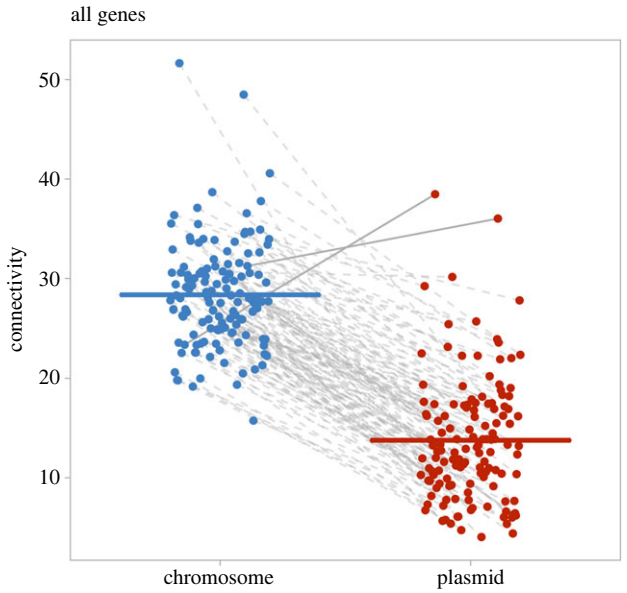


Figure 2. The relative connectivity between chromosomal genes and plasmid genes. Each dot represents the average connectivity of all genes in either the chromosome or plasmid(s) of one species. Chromosome and plasmid values of the same species are linked by a line. A solid line means the average connectivity 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 (132/134), chromosomal genes have a higher level of connectivity than plasmid genes. (Online version in colour.)

Our result was robust to alternative analyses. First, we found the same pattern when we instead analysed the ratio of connectivity between chromosomal genes and plasmid genes instead of the difference. Chromosomal genes on average interacted with 2.6 times more genes than plasmid genes (MCMCglmm; posterior mean = 2.58, 95% CI = 2.002 to 3.142, pMCMC < 0.001, n = 134 species, R^2 of phylogeny = 0.36; electronic supplementary material, table S3). Second, when we looked at the individual species level, chromosomal genes had higher connectivity than plasmid genes in 98.5% of species (132/134); only *Beijerinckia indica* and *Ralstonia pickettii* had plasmid genes with higher connectivity (electronic supplementary material, figure S1 and table S4). Third, we obtained the same qualitative result when we used networks with different confidence thresholds of PPIs (electronic supplementary material, table S3). Increasing the threshold reduced the posterior mean of the differences in chromosome and plasmid connectivity and also reduced the number of species included (electronic supplementary material, table S3).

(b) Plasmid mobility does not affect the relative gene connectivity between chromosomes and plasmids

We then examined whether the mobility of plasmids was correlated with connectivity. We assigned each plasmid in our dataset with one of three mobility types using MOB-suite [47]: non-mobilizable (lowest mobility); mobilizable (intermediate mobility) and conjugative (highest mobility). We found genes on plasmids with different mobilities did not differ from each other in gene connectivity when compared

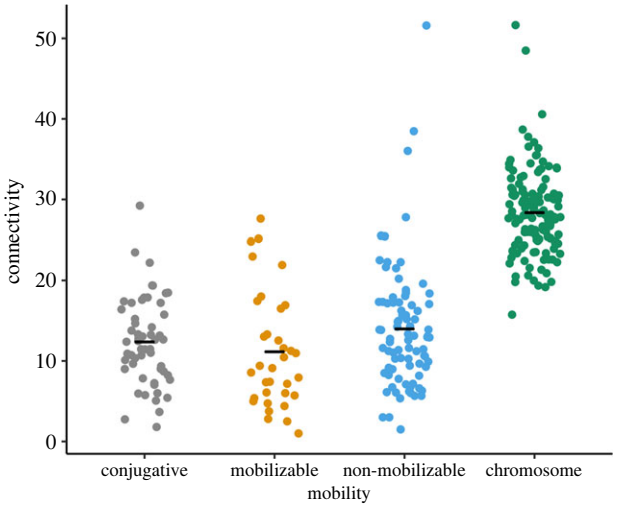


Figure 3. Connectivity of genes on plasmids with different mobilities. We classified plasmids into three mobility types: conjugative (highest mobility); mobilizable (intermediate mobility) and non-mobilizable (lowest mobility). Each dot represents the mean connectivity of all genes on certain replicons for one species. The black horizontal line represents the mean for each group. Plasmid mobility does not influence relative gene connectivity between chromosomes and plasmids. (Online version in colour.)

to genes on chromosomes. Specifically, we first compared the gene connectivity between chromosomes and plasmids across these three mobility types, and found that genes on all three types of plasmids had significantly lower connectivity than genes on chromosomes (MCMCglmm; conjugative plasmids compared to chromosomes: posterior mean = −17.03, 95% CI = −18.60 to −15.24, pMCMC < 0.001; mobilizable plasmids compared to chromosomes: posterior mean = −16.87, 95% CI = −19.03 to −14.81, pMCMC < 0.001; non-mobilizable plasmids compared to chromosomes: posterior mean = −14.15, 95% CI = −15.61 to −12.69, pMCMC < 0.001; conjugative plasmids compared to mobilizable plasmids: posterior mean = 0.20, 95% CI = −2.23 to 2.46, pMCMC = 0.87; mobilizable plasmids compared to non-mobilizable plasmids: posterior mean = 2.75, 95% CI = 0.59 to 5.30, pMCMC = 0.028; n = 134 species; R^2 of fixed effect = 0.535; DIC = 1943.26; figure 3; electronic supplementary material, table S3). Second, we generated a minimum adequate model by combining all genes on plasmids with different mobilities, and compared their gene connectivity with genes on chromosomes. We found that plasmids genes were significantly less connected than chromosomal genes (MCMCglmm; posterior mean = −14.67, 95% CI = −15.85 to −13.67, pMCMC < 0.001; R^2 of fixed effect = 0.572; DIC = 1698.59; figure 2; electronic supplementary material, table S3). Finally, by comparing R^2 of fixed effect and DIC between the two models, we preferred the second minimum adequate model with a higher R^2 of fixed effect and lower DIC. These results suggested that plasmid mobility did not influence relative gene connectivity between chromosomes and plasmids. This result was robust to alternative analysis when we looked at the patterns in networks with different confidence thresholds of PPIs (electronic supplementary material, table S3).

Additionally, we also carried out analysis to examine whether the host range of plasmids was correlated with connectivity. We assigned each plasmid in our dataset with one of the six host ranges using MOB-suite [47]: genus, family, order, class, phylum and multi-phyla. We then compared the gene connectivity of genes on plasmids with different host ranges and genes on chromosomes. We found that genes on plasmids with different host ranges did not differ from each other in gene connectivity when compared to genes on chromosomes (MCMCglmm; multi-phyla compared to phylum: posterior mean = 2.30, 95% CI = −7.39 to 12.66, pMCMC = 0.66; multi-phyla compared to class: posterior mean = −0.90, 95% CI = −8.70 to 8.05, pMCMC = 0.83; multi-phyla compared to order: posterior mean = 8.04, 95% CI = −0.12 to 16.21, pMCMC = 0.06; multi-phyla compared to family: posterior mean = 2.37, 95% CI = −6.71 to 10.72, pMCMC = 0.57; multi-phyla compared to genus: posterior mean = 6.22, 95% CI = −1.63 to 14.36, pMCMC = 0.13; multi-phyla compared to chromosomes: posterior mean = 21.69, 95% CI = 13.53 to 29.59, pMCMC < 0.001; electronic supplementary material, figure S7 and table S3). This result was robust to alternative analysis using networks with different thresholds (electronic supplementary material, table S3).

(c) Genes connectivity did not consistently differ between genes encoding extracellular or intracellular proteins

We then tested if genes coding for extracellular (cooperative) and intracellular (private) proteins were different from each other in terms of their relative gene connectivity between chromosomes and plasmids. We found genes coding for extracellular proteins (cooperative) were significantly more connected on chromosomes than on plasmids (MCMCglmm; posterior mean = 9.325, 95% CI = 0.108 to 17.203, pMCMC = 0.032; n = 134 species; figure 4*a*; electronic supplementary material, figure S4 and table S3). At the individual species level, 81.3% (109/134) of species showed this pattern, with genes coding for extracellular proteins having higher connectivity on chromosomes than plasmids (electronic supplementary material, figure S2 and table S5). Genes coding for intracellular proteins (private) were also significantly more connected on chromosomes than on plasmids (MCMCglmm; posterior mean = 15.325, 95% CI = 12.190 to 18.393, pMCMC < 0.001; n = 134 species; figure 4*b*; electronic supplementary material, figure S4 and table S3). At the individual species level, 97.8% (131/134) of species showed this pattern in that direction (electronic supplementary material, figure S3 and table S6).

We then examined sociality (extracellular or intracellular) and location (chromosome or plasmid) simultaneously. We found that the difference in gene connectivity between chromosomal genes and plasmids genes was smaller for genes coding for extracellular proteins compared to intracellular proteins (MCMCglmm; Interaction term: posterior mean = −4.410, 95% CI = −8.567 to −0.331, pMCMC = 0.042; n = 134 species; R^2 of interaction term = 0.021, electronic supplementary material, table S3). The small R^2 of the interaction term suggested that the effect size of the interaction between sociality and location is small. In addition, the significance of this result was not robust to different network thresholds, with the interactions between sociality and location being significant at the lower threshold (150), but not at higher thresholds (700 and 900). Although the interaction between sociality and location was significant in networks with

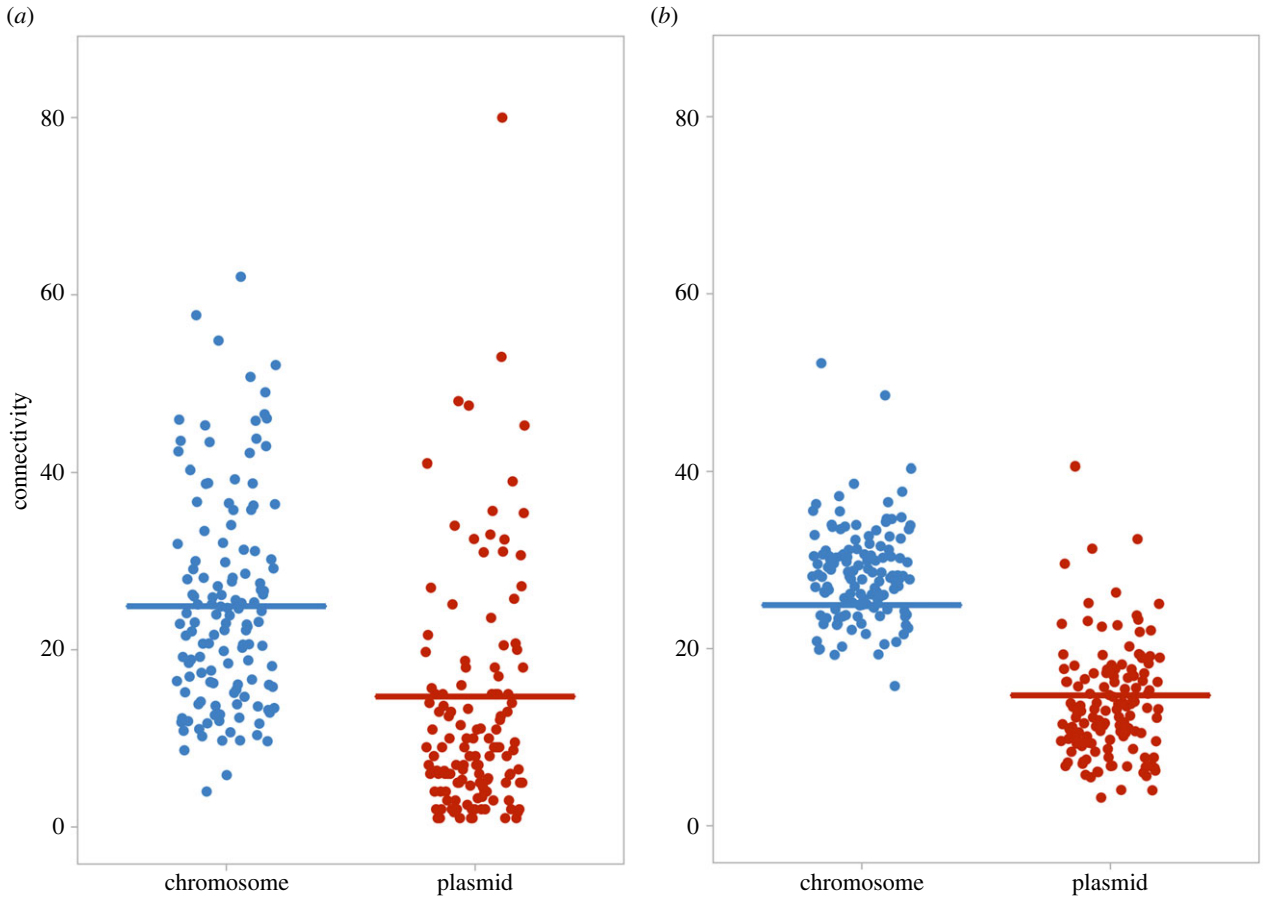


Figure 4. Comparison of chromosomal and plasmid connectivity for genes with different sociality (*a*) chromosome versus plasmid comparisons for genes encoding extracellular proteins (cooperative); (*b*) chromosome versus plasmid comparisons for genes encoding intracellular proteins (private). Each dot 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. Two outlying species have been removed from figure 4*b*, where genes on plasmids have much higher levels of connectivity (greater than 85). The complete version of figure 4*b*, with these two outlying data points, is in the electronic supplementary material, figure S4. Chromosomal genes were more connected than plasmid genes, for both genes encoding extracellular proteins and genes encoding intracellular proteins. (Online version in colour.)

a threshold of 150, the effect size of the interaction remained small (MCMCglmm; Interaction term: posterior mean = −23.096, 95% CI = −43.832 to −2.354, pMCMC = 0.042; $n = 134$ species; R^2 of interaction term = 0.004, electronic supplementary material, table S3). Overall, these results suggest that there is either a small or no interaction between the influence of sociality and location on gene connectivity (electronic supplementary material, table S3). If we ignored the potential interactions between sociality and location when in associate with gene connectivity, and directly compared the connectivity between genes coding for extracellular and intracellular proteins, we found no significant differences between them (MCMCglmm; posterior mean = −6.06, 95% CI = −12.72 to −0.68, pMCMC = 0.052; $n = 134$ species; figure 4).

We also found a suggestive result that the variance of gene connectivity was greater for extracellular proteins, relative to intracellular proteins. This result is only suggestive because we were only able to examine with a standard *F*-test (*F*-test; chromosomes: $F_{133,133} = 0.194$, p-value $< 2.2 \times 10^{-16}$; plasmids: $F_{133,133} = 0.084$, p-value $< 2.2 \times 10^{-16}$), rather than a MCMCglmm analysis that controls for phylogeny. Species often share characteristics inherited through common descent, rather than through independent evolution and so cannot be considered independent data points [55]. Therefore, given that we cannot control for phylogeny, this result should only be seen as suggestive.

(d) Network size did not affect our results

Network size (the total number of proteins in a PPI network) could affect gene connectivity. In the extreme, gene connectivity could be constrained by small network sizes. There is evidence that genes with the same connectivity have different impacts in networks of different sizes [45]. To examine whether network size influenced the connectivity of genes in our dataset, we tested whether the two were correlated. We found no significant correlation between network size and gene connectivity (MCMCglmm; posterior mean = 0.000212, 95% CI = −0.000480 to 0.000916, pMCMC = 0.578, $n = 134$ species; electronic supplementary material, figure S5 and table S7). We also found no significant correlation between network size and either the difference or the ratio of connectivity between chromosomes and plasmids (electronic supplementary material, table S7). The same results were also applied to genes coding for extracellular proteins (electronic supplementary material, table S7). These results suggest that network size did not affect our finding that connectivity of genes is higher for chromosomes than plasmids.

(e) Gene connectivity was higher in core versus accessory genes

We also compared the gene connectivity between core genes, that are found in all (100%) genomes of a species, and

accessory genes, which are only found in some genomes. The pangenome information was retrieved from the PanX database (https://pangenome.org/) [56]. We were only able to analyse seven species (nine strains) in our dataset where there was also data available on core and accessory genes in the PanX database (electronic supplementary material, table S9). We found that when on chromosomes, core genes had significantly higher levels of connectivity compared to accessory genes (MCMCglmm; posterior mean = 12.63, 95% CI = 7.22 to 17.68, pMCMC < 0.001, $n = 7$ species; electronic supplementary material, figure S8 and table S3). This result was also robust to networks with other thresholds (electronic supplementary material, table S3).

4. Discussion

We found that: (i) as predicted by the complexity hypothesis, plasmid genes had consistently lower connectivity compared to chromosome genes (figure 2); (ii) contrary to the prediction of the complexity hypothesis, there was no correlation between plasmid mobility and gene connectivity (figure 3); and (iii) genes encoding extracellular proteins and genes encoding intracellular proteins did not differ in relative gene connectivity between chromosomes and plasmids (figure 4).

Our finding that genes on plasmids had lower levels of connectivity than genes on the chromosome is consistent with previous studies looking at horizontally transferred genes [17–19] (figure 2). The explanation of this pattern from the perspective of the complexity hypothesis is that if genes on plasmids are likely to undergo frequent horizontal gene transfer, they will be less likely to be malfunctioning in a new host if they have lower connectivity [17,20]. However, alternative explanations are also possible, because plasmids are not merely gene delivery platforms and not all plasmids are transferable [57]. Around 53% of plasmids lack relaxases (i.e. non-mobilizable plasmids), thus cannot be mobilized by conjugation, but only by transformation or transduction, similar to chromosomal genes [24]. Even for mobilizable plasmids that are capable of using the genetic cassette of other conjugative plasmids, the rate of conjugation of these plasmids is still likely to be lower than that of conjugative plasmids [58–60]. Therefore, if plasmid genes were less connected because they undergo more frequent horizontal gene transfer, as predicted by the complexity hypothesis, we would expect plasmid gene connectivity to decrease with increasing plasmid mobility.

Our next finding, however, suggested that in contrast with the complexity hypothesis, the mobility of plasmids was not correlated with gene connectivity (figure 3). This suggests that other evolutionary forces beyond the horizontal transfer of plasmids may contribute to the pattern we observed. This is also in line with a recent review which highlighted characteristics of plasmids, in addition to their potential for horizontal transfer, that could drive specific evolutionary dynamics of plasmid-encoded genes, and have been largely overlooked until recently [57].

Instead of horizontal gene transfer, another possible factor that could explain the difference between plasmid and chromosome gene connectivity is the inheritance stability of genes on plasmids. Plasmids are suggested to be less stable than chromosomes, because (i) plasmids usually confer a cost to the host, causing a competitive disadvantage that may select for plasmids to be lost in the absence of a benefit to the cell, and (ii) plasmids can be lost during cell division if they are incompletely segregated between daughter cells [21,61–65]. Therefore, genes on plasmids may not be as stable as genes on chromosomes.

A theoretical study suggested that essential genes, which are usually highly connected, were more likely to be found on chromosomes rather than on plasmids because the inheritance of chromosomes is more stable than that for plasmids [66]. A recent empirical study provided further evidence that plasmid inheritance instability is responsible for essential genes not being carried on plasmids, by observing that inserting an essential chromosomal gene into a plasmid makes the plasmid more likely to be lost in *E. coli* [67]. A variety of mechanisms have been suggested to help stabilize plasmid persistence, such as host-plasmid co-adaptation, compensatory evolution and high plasmid transfer rates [65,68–71]. Theory predicts that even low rates of plasmid loss can make essential genes more likely to be on chromosomes than plasmids. Specifically, the only case where essential genes can be found on plasmids is when essential genes are more likely to be lost when on chromosomes compared to on plasmids [66]. However, this is very unlikely to be the case in the real world. Consequently, given our results that plasmid mobility has a limited effect on plasmid gene connectivity, the instability of plasmid inheritance may instead be a more likely explanation for why highly connected genes are frequently absent from plasmids.

Examining chromosomal genes, we found that core genes, present in every genome, had a higher connectivity than accessory genes, found in only a subset of genomes (electronic supplementary material, figure S8 and table S3). This result was predicted because core genes are likely to encode more essential functions than accessory genes, and thus tend to have higher connectivity [4,5]. An alternate explanation, that cannot be separated, is that horizontal gene transfer is more common in the accessory genome. Horizontal gene transfer is just one process for shaping the accessory genome, which accounts for 15.5% of accessory genomes, alongside with other processes such as gene deletion [72,73]. Although these rates of horizontal gene transfer are much lower than with plasmids, and we cannot separate out different rates of horizontal gene transfer as we can with plasmids (figure 1).

Chromosomal genes can undergo horizontal gene transfer via mechanisms such as integrative conjugative elements (ICEs) and prophages [74,75]. However, there are three reasons that this is unlikely to have had a major influence on the broader patterns we examined comparing plasmids and chromosomes. First, most chromosomal genes can be regarded as immobile. A study of 80 bacteria found that horizontally transferred genes were concentrated in only a small fraction of chromosomal regions (approx. 1%) [72]. Second, plasmids have evolved to have fitness interests distinct from chromosomes, therefore, chromosomal and plasmid genes differ not only in their rate of horizontal gene transfer, but also in how selection operates to shape their traits [76]. Regarding gene connectivity, our findings demonstrated that selection was not necessarily associated with the level of horizontal gene transfer across plasmids. Third, rates of horizontal transfer can be much higher for plasmids compared to chromosomes [77,78]. What matters is not just whether horizontal transfer occurs, but how frequently over

evolutionary time. Nonetheless, a useful future direction would be to extend analyses to other routes of horizontal transfer, such as ICEs.

We also compared the connectivity of genes encoding cooperative (extracellular) to genes encoding private (intracellular) factors. We examined this factor because genes coding for cooperative factors have been hypothesized to undergo more frequent horizontal transfer and may have particularly low connectivity when on plasmids to allow for easier transfer [29,30]. However, we found no difference in the connectivity of cooperative and private genes, and we also found that the relative difference between plasmids and chromosome genes was the same for cooperative genes compared to private genes (figure 4). This lack of difference in connectivity between genes coding for extracellular and intracellular proteins could be due to two reasons. (i) Genes coding for extracellular proteins are not more likely to be transferred horizontally. A recent study found that genes coding for extracellular proteins were not overrepresented on plasmids compared to chromosomes, or on more mobile plasmids compared to less mobile plasmids [32]. (ii) Other factors could affect gene connectivity of plasmid genes in addition to any effect due to horizontal gene transfer, such as gene essentiality or plasmid instability, which would act similarly on all genes, not just those coding for extracellular proteins. Either way, these results suggested that the factors determining which genes are carried on plasmids, particularly which genes might bear the potential costs of plasmid instability and loss, are not affected by whether the gene codes for a cooperative public good.

To conclude, our results provide mixed support for the complexity hypothesis, suggesting that there are other factors at play. While genes on plasmids have lower connectivity than genes on chromosomes, their connectivity did not correlate with the rate at which different plasmids are likely to transfer horizontally. This suggests that other factors, particularly the stability of gene inheritance, might be more important than gene mobility in explaining the variation in gene connectivity across prokaryotic genomes. Key tasks for the future include (i) examining gene connectivity on other mobile genetic elements (such as phages and ICEs [79]), (ii) directly examining the role of stability and (iii) using different or additional methods for identifying cooperative traits.

Data accessibility. The data used to generate all results and figures are provided in the electronic supplementary material [80].

Authors' contributions. C.H.: conceptualization, data curation, formal analysis, methodology, visualization, writing—original draft and writing—review and editing; A.E.D.: conceptualization, methodology, writing—original draft and writing—review and editing; S.A.W.: conceptualization, funding acquisition, project administration, supervision and writing—review and editing; M.G.: conceptualization, project administration, supervision and writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. The authors declare no competing interests.

Funding. This work was supported by European Research Council Grants 834164 (to A.E.D, S.A.W. and M.G).

Acknowledgements. We thank Craig MacLean, Josh Firth, Laurence Belcher, Zheren Zhang and Zhijie Liao for their helpful comments. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work. http://dx.doi.org/10.5281/zenodo.22558.

References

1. Newman M. 2018 *Networks*. Oxford, UK: Oxford University Press.
2. Barabási AL, Oltvai ZN. 2004 Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113. (doi:10.1038/nrg1272)
3. Snider J, Kotlyar M, Saraon P, Yao Z, Jurisica I, Stagljar I. 2015 Fundamentals of protein interaction network mapping. *Mol. Syst. Biol.* **11**, 848. (doi:10.15252/msb.20156351)
4. Jeong H, Mason SP, Barabási AL, Oltvai ZN. 2001 Lethality and centrality in protein networks. *Nature* **411**, 41–42. (doi:10.1038/35075138)
5. Coulomb S, Bauer M, Bernard D, Marsolier-Kergoat MC. 2005 Gene essentiality and the topology of protein interaction networks. *Proc. R. Soc. B* **272**, 1721–1725. (doi:10.1098/rspb.2005.3128)
6. Yu H *et al*. 2008 High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110. (doi:10.1126/science.1158684)
7. Rajagopala SV *et al*. 2014 The binary protein-protein interaction landscape of Escherichia coli. *Nat. Biotechnol.* **32**, 285–290. (doi:10.1038/nbt.2831)
8. Cong Q, Anishchenko I, Ovchinnikov S, Baker D. 2019 Protein interaction networks revealed by proteome coevolution. *Science* **365**, 185–189. (doi:10.1126/science.aaw6718)
9. He X, Zhang J. 2006 Why do hubs tend to be essential in protein networks? *PLoS Genet.* **2**, e88. (doi:10.1371/journal.pgen.0020088)
10. Khuri S, Wuchty S. 2015 Essentiality and centrality in protein interaction networks revisited. *BMC Bioinf.* **16**, 109. (doi:10.1186/s12859-015-0536-x)
11. Rancati G, Moffat J, Typas A, Pavelka N. 2018 Emerging and evolving concepts in gene essentiality. *Nat. Rev. Genet.* **19**, 34–49. (doi:10.1038/nrg.2017.74)
12. Soucy SM, Huang J, Gogarten JP. 2015 Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* **16**, 472–482. (doi:10.1038/nrg3962)
13. Syvanen M. 2012 Evolutionary implications of horizontal gene transfer. *Annu. Rev. Genet.* **46**, 341–358. (doi:10.1146/annurev-genet-110711-155529)
14. Gogarten JP, Townsend JP. 2005 Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* **3**, 679–687. (doi:10.1038/nrmicro1204)
15. Koonin EV, Makarova KS, Aravind L. 2001 Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* **55**, 709–742. (doi:10.1146/annurev.micro.55.1.709)
16. Brito IL. 2021 Examining horizontal gene transfer in microbial communities. *Nat. Rev. Microbiol.* **19**, 442–453. (doi:10.1038/s41579-021-00534-7)
17. Cohen O, Gophna U, Pupko T. 2011 The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* **28**, 1481–1489. (doi:10.1093/molbev/msq333)
18. Davids W, Zhang Z. 2008 The impact of horizontal gene transfer in shaping operons and protein interaction networks—direct evidence of preferential attachment. *BMC Evol. Biol.* **8**, 23. (doi:10.1186/1471-2148-8-23)
19. Lercher MJ, Pál C. 2008 Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol. Biol. Evol.* **25**, 559–567. (doi:10.1093/molbev/msm283)
20. Jain R, Rivera MC, Lake JA. 1999 Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 3801–3806. (doi:10.1073/pnas.96.7.3801)
21. Sørensen SJ, Bailey M, Hansen LH, Kroer N, Wuertz S. 2005 Studying plasmid horizontal transfer in situ: a critical review. *Nat. Rev. Microbiol.* **3**, 700–710. (doi:10.1038/nrmicro1232)
22. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, Cruz F de la. 2010 Mobility of Plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434–452. (doi:10.1128/MMBR.00020-10)
23. Ramsay JP, Firth N. 2017 Diverse mobilization strategies facilitate transfer of non-conjugative

mobile genetic elements. *Curr. Opin. Microbiol.* **38**, 1–9. (doi:10.1016/j.mib.2017.03.003)

24. Coluzzi C, Garcillán-Barcia M del P, Cruz F de la, Rocha EPC. 2022 Evolution of plasmid mobility: origin and fate of non-conjugative plasmids. *Mol. Biol. Evol.* **39**, msac115. (doi:10.1101/2021.12.10.472114)
25. Acman M, van Dorp L, Santini JM, Balloux F. 2020 Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* **11**, 2452. (doi:10.1038/s41467-020-16282-w)
26. West SA, Griffin AS, Gardner A, Diggle SP. 2006 Social evolution theory for microorganisms. *Nat. Rev. Microbiol.* **4**, 597–607. (doi:10.1038/nrmicro1461)
27. West SA, Diggle SP, Buckling A, Gardner A, Griffin AS. 2007 The social lives of microbes. *Ann. Rev. Ecol. Evol. Syst.* **38**, 53–77. (doi:10.1146/annurev.ecolsys.38.091206.095740)
28. Smith J. 2001 The social evolution of bacterial pathogenesis. *Proc. R. Soc. Lond. Ser. B* **268**, 61–69. (doi:10.1098/rspb.2000.1330)
29. Ginty SEM, Rankin DJ, Brown SP. 2011 Horizontal gene transfer and the evolution of bacterial cooperation. *Evolution* **65**, 21–32. (doi:10.1111/j.1558-5646.2010.01121.x)
30. Mc Ginty SÉ, Lehmann L, Brown SP, Rankin DJ. 2013 The interplay between relatedness and horizontal gene transfer drives the evolution of plasmid-carried public goods. *Proc. R. Soc. B* **280**, 20130400. (doi:10.1098/rspb.2013.0400)
31. Lee IPA, Eldakar OT, Gogarten JP, Andam CP. 2021 Bacterial cooperation through horizontal gene transfer. *Trends Ecol. Evol.* **37**, 223–232. (doi:10.1016/j.tree.2021.11.006)
32. Dewar AE, Thomas JL, Scott TW, Wild G, Griffin AS, West SA, Ghoul M. 2021 Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range. *Nat. Ecol. Evol.* **5**, 1624–1636. (doi:10.1038/s41559-021-01573-2)
33. Szklarczyk D *et al*. 2019 STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613. (doi:10.1093/nar/gky1131)
34. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. 2004 The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451. (doi:10.1093/nar/gkh086)
35. Orchard S *et al*. 2014 The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363. (doi:10.1093/nar/gkt1115)
36. Ammari MG, Gresham CR, McCarthy FM, Nanduri B. 2016 HPIDB 2.0: a curated database for host–pathogen interactions. *Database* **2016**, baw103. (doi:10.1093/database/baw103)
37. Chatr-aryamontri A *et al*. 2017 The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45**, D369–D379. (doi:10.1093/nar/gkw1102)
38. Zhang QC, Petrey D, Garzón JI, Deng L, Honig B. 2013 PrePPI: a structure-informed database of protein–protein interactions. *Nucleic Acids Res.* **41**, D828–D833. (doi:10.1093/nar/gks1231)
39. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. 2005 STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**, D433–D437. (doi:10.1093/nar/gki005)
40. Nogueira T, Touchon M, Rocha EPC. 2012 Rapid evolution of the sequences and gene repertoires of secreted proteins in bacteria. *PLoS ONE* **7**, e49403. (doi:10.1371/journal.pone.0049403)
41. Nogueira T, Rankin DJ, Touchon M, Taddei F, Brown SP, Rocha EPC. 2009 Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Curr. Biol.* **19**, 1683–1691. (doi:10.1016/j.cub.2009.08.056)
42. Garcia-Garcera M, Rocha EPC. 2020 Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. *Nat. Commun.* **11**, 758. (doi:10.1038/s41467-020-14572-x)
43. Lau WYV, Hoad GR, Jin V, Winsor GL, Madyan A, Gray KL, Laird MR, Lo R, Brinkman FSL. 2021 PSORTdb 4.0: expanded and redesigned bacterial and archaeal protein subcellular localization database incorporating new secondary localizations. *Nucleic Acids Res.* **49**, D803–D808. (doi:10.1093/nar/gkaa1095)
44. Park C, Zhang J. 2012 High expression hampers horizontal gene transfer. *Genome Biol. Evol.* **4**, 523–532. (doi:10.1093/gbe/evs030)
45. Li X, Li W, Zeng M, Zheng R, Li M. 2020 Network-based methods for predicting essential genes or proteins: a survey. *Brief. Bioinform.* **21**, 566–583. (doi:10.1093/bib/bbz017)
46. Csardi G, Nepusz T. 2005 The Igraph Software Package for Complex Network Research. *InterJournal Complex Syst.* **1695**, 1–9.
47. Robertson J, Nash JHE. 2018 MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genom* **4**, e000206. (doi:10.1099/mgen.0.000206)
48. Hadfield JD. 2010 MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R Package. *J. Stat. Softw.* **33**, 1–22. (doi:10.18637/jss.v033.i02)
49. Grafen A, Hamilton WD. 1989 The phylogenetic regression. *Phil. Trans. R. Soc. Lond. B* **326**, 119–157. (doi:10.1098/rstb.1989.0106)
50. Nakagawa S, Schielzeth H. 2013 A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **4**, 133–142. (doi:10.1111/j.2041-210x.2012.00261.x)
51. Nakagawa S, Johnson PCD, Schielzeth H. 2017 The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R. Soc. Interface* **14**, 20170213. (doi:10.1098/rsif.2017.0213)
52. Hug LA *et al*. 2016 A new view of the tree of life. *Nat. Microbiol.* **1**, 1–6. (doi:10.1038/nmicrobiol.2016.48)
53. Paradis E, Schliep K. 2019 ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528. (doi:10.1093/bioinformatics/bty633)
54. Revell LJ. 2012 phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223. (doi:10.1111/j.2041-210X.2011.00169.x)
55. Felsenstein J. 1985 Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15. (doi:10.1086/284325)
56. Ding W, Baumdicker F, Neher RA. 2018 panX: pan-genome analysis and exploration. *Nucleic Acids Res.* **46**, e5. (doi:10.1093/nar/gkx977)
57. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. 2021 Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat. Rev. Microbiol.* **19**, 347–359. (doi:10.1038/s41579-020-00497-1)
58. Pérez-Mendoza D, Lucas M, Muñoz S, Herrera-Cervera JA, Olivares J, de la Cruz F, Sanjuán J. 2006 The relaxase of the *Rhizobium etli* symbiotic plasmid shows nic site cis-acting preference. *J. Bacteriol.* **188**, 7488–7499. (doi:10.1128/JB.00701-06)
59. Blanca-Ordóñez H, Oliva-García JJ, Pérez-Mendoza D, Soto MJ, Olivares J, Sanjuán J, Nogales J. 2010 pSymA-dependent mobilization of the *Sinorhizobium meliloti* pSymB megaplasmid. *J. Bacteriol.* **192**, 6309–6312. (doi:10.1128/JB.00549-10)
60. Klümper U, Droumpali A, Dechesne A, Smets BF. 2014 Novel assay to measure the plasmid mobilizing potential of mixed microbial communities. *Front. Microbiol.* **5**, 730. (doi:10.3389/fmicb.2014.00730)
61. Baltrus DA. 2013 Exploring the costs of horizontal gene transfer. *Trends Ecol. Evol.* **28**, 489–495. (doi:10.1016/j.tree.2013.04.002)
62. Nordström K, Austin SJ. 1989 Mechanisms that contribute to the stable segregation of plasmids. *Annu. Rev. Genet.* **23**, 37–69. (doi:10.1146/annurev.ge.23.120189.000345)
63. San Millan A, MacLean RC. 2017 Fitness costs of plasmids: a limit to plasmid transmission. *Microbiol. Spect.* **5**, 5.5.02. (doi:10.1128/microbiolspec.MTBP-0016-2017)
64. Hall JPJ, Brockhurst MA, Dytham C, Harrison E. 2017 The evolution of plasmid stability: are infectious transmission and compensatory evolution competing evolutionary trajectories? *Plasmid* **91**, 90–95. (doi:10.1016/j.plasmid.2017.04.003)
65. Carroll AC, Wong A. 2018 Plasmid persistence: costs, benefits, and the plasmid paradox. *Can. J. Microbiol.* **64**, 293–304. (doi:10.1139/cjm-2017-0609)
66. Tazzyman SJ, Bonhoeffer S. 2015 Why there are no essential genes on plasmids. *Mol. Biol. Evol.* **32**, 3079–3088. (doi:10.1093/molbev/msu293)
67. Wein T, Wang Y, Barz M, Stücker FT, Hammerschmidt K, Dagan T. 2021 Essential gene acquisition destabilizes plasmid inheritance. *PLoS Genet.* **17**, e1009656. (doi:10.1371/journal.pgen.1009656)

68. Harrison E, Brockhurst MA. 2012 Plasmid-mediated horizontal gene transfer is a coevolutionary process. *Trends Microbiol.* **20**, 262–267. (doi:10.1016/j.tim.2012.04.003)
69. Dimitriu T, Matthews AC, Buckling A. 2021 Increased copy number couples the evolution of plasmid horizontal transmission and plasmid-encoded antibiotic resistance. *Proc. Natl Acad. Sci. USA* **118**, e2107818118. (doi:10.1073/pnas.2107818118)
70. Hall JPJ, Wright RCT, Harrison E, Muddiman KJ, Wood AJ, Paterson S, Brockhurst MA. 2021 Plasmid fitness costs are caused by specific genetic conflicts enabling resolution by compensatory mutation. *PLoS Biol.* **19**, e3001225. (doi:10.1371/journal.pbio.3001225)
71. Brockhurst MA, Harrison E. 2021 Ecological and evolutionary solutions to the plasmid paradox. *Trends Microbiol.* **30**, 534–543. (doi:10.1016/j.tim.2021.11.001)
72. Oliveira PH, Touchon M, Cury J, Rocha EPC. 2017 The chromosomal organization of horizontal gene transfer in bacteria. *Nat. Commun.* **8**, 841. (doi:10.1038/s41467-017-00808-w)
73. Domingo-Sananes MR, McInerney JO. 2021 Mechanisms that shape microbial pangenomes. *Trends Microbiol.* **29**, 493–503. (doi:10.1016/j.tim.2020.12.004)
74. Canchaya C, Proux C, Fournous G, Bruttin A, Brüssow H. 2003 Prophage genomics. *Microbiol. Mol. Biol. Rev.* **67**, 238–276. (doi:10.1128/MMBR.67.2.238-276.2003)
75. Burrus V, Pavlovic G, Decaris B, Guédon G. 2002 Conjugative transposons: the tip of the iceberg. *Mol. Microbiol.* **46**, 601–610. (doi:10.1046/j.1365-2958.2002.03191.x)
76. Hall JPJ. 2021 Is the bacterial chromosome a mobile genetic element? *Nat. Commun.* **12**, 6400. (doi:10.1038/s41467-021-26758-y)
77. Sheppard RJ, Beddis AE, Barraclough TG. 2020 The role of hosts, plasmids and environment in determining plasmid transfer rates: a meta-analysis. *Plasmid* **108**, 102489. (doi:10.1016/j.plasmid.2020.102489)
78. Vos M, Hesselman MC, te Beek TA, van Passel MWJ, Eyre-Walker A. 2015 Rates of lateral gene transfer in prokaryotes: high but why? *Trends Microbiol.* **23**, 598–605. (doi:10.1016/j.tim.2015.07.006)
79. Frost LS, Leplae R, Summers AO, Toussaint A. 2005 Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732. (doi:10.1038/nrmicro1235)
80. Hao C, Dewar AE, West SA, Ghoul M. 2022 Gene transferability and sociality do not correlate with gene connectivity. Figshare. (doi:10.6084/m9.figshare.c.6302904)

Chapter 4: Environmental Variability Mediates the Evolution of Bacterial Growth Rate and Genomic Characteristics

Abstract

Growth rates in bacteria vary widely across species. From an analysis of 149,932 genomes spanning 171 bacterial species, we explored the influence of ecological (environmental variability, host dependence) and genomic (genome size, metabolic capacity and metabolic versatility) factors on the variations of bacterial growth rates. Theories suggested a positive link between environmental variability and growth rates, and anticipated trade-offs between growth rates and three genomic features. Our findings, however, revealed positive correlations between each of the three genomic factors and growth rates. Phylogenetic path analysis further showed that environmental variability has a direct positive correlation with both growth rates and genomic features, while direct correlations between growth rates and genomic characteristics were either non-existent or weak. This suggests that environmental variability might mediate the relationship between growth rates and genomic characteristics, potentially counteracting anticipated trade-offs among them.

Introduction

Bacterial growth rates can vary substantially across species. Fast-growing bacteria, such as *Escherichia coli*, have a doubling time of around 20 minutes under steady-state laboratory conditions (1). In contrast, slow-growing bacteria like *Helicobacter pylori* typically need 3 days to double on suitable media (2). In extreme circumstances, such as deep subseafloor environments, bacterial generation time might even extend beyond 100 years (3, 4). Both genomic characteristics and ecological factors can shape bacterial growth rate variations across species, yet we have limited knowledge about the extent of their effect (5–8).

Environmental variability is a crucial ecological factor influencing bacterial growth rates (9). In unpredictable environments, rapid resource utilization for reproduction is favoured, as resources sporadically become available and waiting too long might mean missing out on reproductive opportunities before scarcity ensues. In contrast, stable environments tend to promote efficiency in resource use, leading to slower growth since resources are consistently available. Thus, rapid, wasteful reproduction isn't advantageous in such conditions (10, 11). A recent study corroborated this, showing that generalist bacteria, which likely encounter more varied environments, typically reproduce faster than specialists in metagenomic samples (12).

In addition to environmental variability, the level of host dependence is another key ecological factor that reflects bacterial life-history strategies and could potentially influence their growth rates. This factor can impact growth in multiple ways. On one hand, bacteria that are host-dependent often have access to a stable and abundant supply of nutrients, which can facilitate faster growth (13). For instance, some intestinal bacteria such as *Escherichia coli* can grow rapidly in the human gut (14). On the other hand, the growth of host-associated bacteria might also be regulated by the host, leading to a controlled rate of proliferation (15). For example,

aphid symbiont *Buchnera aphidicola* was estimated to double every 7 to 13 days in its host (16).

Regarding genomic characteristics, it has been hypothesised that reduced genome size could allow a higher growth rate (17–19). This stemmed from the understanding that rRNA and protein synthesis are the predominant activities of growing bacteria. Consequently, selection for high growth rates could introduce evolutionary pressure favouring reallocation of phosphorus (P) and nitrogen (N) from DNA to RNA or proteins (5, 17, 20, 21). However, the evidence for this hypothesis in bacteria is mixed (22–24). Some experiments have found that genome reductions boost growth in certain contexts (25, 26), yet others suggested that the relationship was conditional, with smaller genomes sizes only aiding growth rate with extra nutrient provision (27). Additionally, comparative analyses haven't identified any consistent correlation between genome size and growth rate across bacterial genomes (28–30).

Another genomic characteristic potentially impacting bacterial growth rates is metabolic capacity or versatility, which describes the range of metabolic functions encoded in a bacterial genome. Theories proposed that a trade-off exists between the efficiency of nutrient utilization (yield) and the speed of growth (rate). This trade-off manifests at both the cellular and genomic levels, with energy being allocated differently between more ribosome production for rapid growth and more metabolic enzyme synthesis for efficient nutrient utilization (31, 32). Two contrasting metabolic strategies thus arise from this: fast-but-inefficient growth and slow-but-efficient growth (33–35). For instance, ammonia-oxidizing bacteria exhibited fast-but-inefficient growth by encoding fragmented metabolic pathways for nitrification (36). However, computational models offer a counterpoint, suggesting that species with greater metabolic

flexibility actually grow faster, thus challenging the traditional understanding of the rate-yield trade-off (7).

Why is the evidence on the impact of genomic characteristics and ecological factors on bacterial growth rates mixed? One likely reason is the interactive nature of these factors, which complicates their individual effects on bacterial growth. For instance, environmental variability has been linked to an increase in genome size (37–39). This suggested that as bacteria occupy more diverse habitats, evolutionary pressures could drive both genome size and growth rate upward, potentially counteracting any expected negative trade-off between the two. Furthermore, obligate endosymbionts like *Buchnera aphidicola* and *Mycobacterium leprae* often experience genome streamlining (40–43). If their hosts also restrict their growth, the anticipated negative correlation between growth rate and genome size could actually become positive, contradicting existing hypotheses. Given these complexities, a comprehensive examination of the interplay among these various factors is essential for understanding what shapes bacterial growth rates.

Finally, it's crucial to consider the potential confounding impact of variations in gene content within bacterial species. Individual genomes within the same bacterial species often show significant differences in their gene sets (44–46). These intra-specific gene content variations have been linked to habitat preferences or life history strategies of bacteria (Dewar, A.E., 2023), and potentially could indirectly influence bacterial growth rates, further complicating the relationships among the variables under study (47).

We conducted a comparative analysis to explore how both ecological factors and genomic characteristics influence the evolution of bacterial growth rates. Using data from 149,932 genomes spanning 171 bacterial species, we initially tested hypotheses for each pair of variables. We then evaluated the complex interactions among multiple variables affecting bacterial growth rates through phylogenetic path analysis (Figure 1). To account for the potential confounding effects of the variations in gene content within bacterial species, we also included it as a covariate in each model.

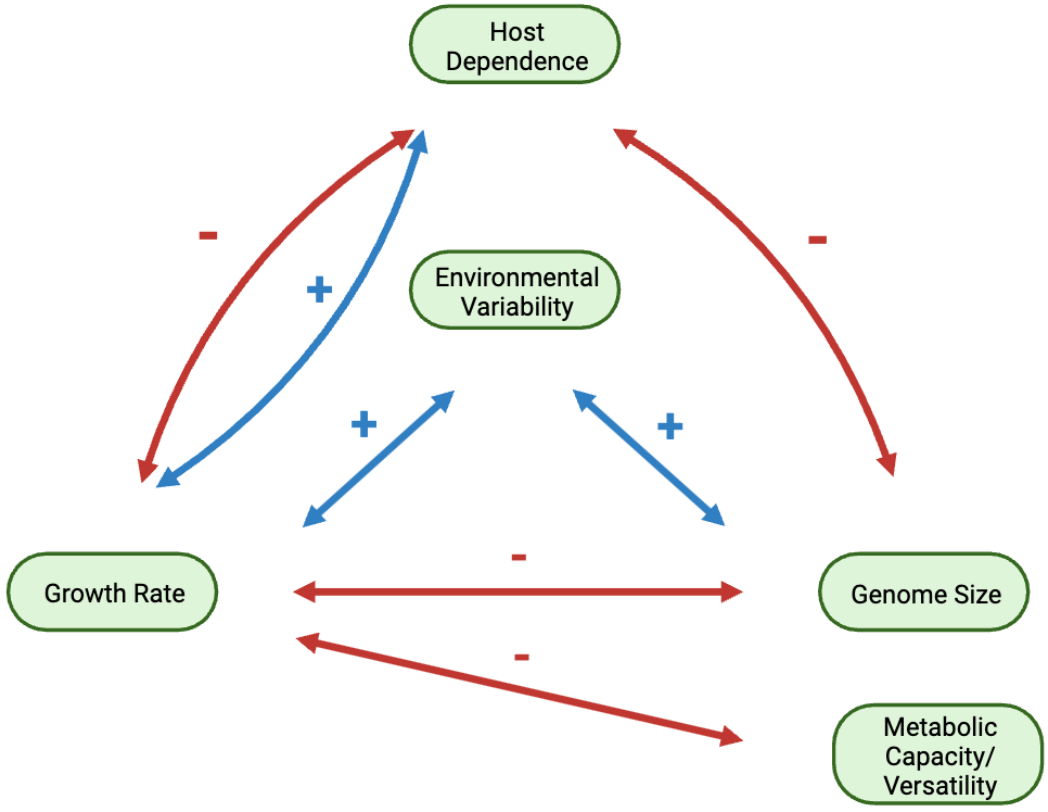


Figure 1. Conceptual diagram illustrating the hypothesized relationships among growth rate, ecological factors (environmental variability and host dependence), as well as genomic characteristics (genome size and metabolic capacity or versatility). Blue arrows represent positive correlations, while red arrows denote negative correlations. The direction of each correlation is based on hypotheses focusing solely on pairwise relationships.

Results

Dataset Compilation and Variable Definitions

We analysed genomic data from 149,932 bacterial genomes across 171 species to study how ecological factors (environmental variability, host dependence) and genomic characteristics (genome size, metabolic capacity, metabolic versatility) influence bacterial growth rates. The variables were determined as:

1) Growth Rate: Calculated as the negative logarithm of the predicted minimal doubling time (in hours) for each genome. The minimal doubling time was predicted based on codon usage patterns (48).
2) Environmental Variability: Defined as the number of habitat clusters where each species is found. These habitat clusters were created by grouping similar habitats based on the presence or absence of species. For example, human lung and human skin were grouped together as a single habitat cluster (Figure S1).
3) Host dependence: Species categorized as either host-dependent, free-living, or both, depending on their habitat preference (49).
4) Genome Size: Represented by the count of unique genes in each genome.
5) Metabolic Capacity: Denoted by the number of unique metabolic genes in each genome.
6) Metabolic Versatility: Reflected by the number of distinct metabolic pathways in each genome.

Additionally, to account for intra-specific variations in gene content, we included genome fluidity, a measure of such variations, as a covariate in our analyses. To allow cross-species comparisons, we averaged the growth rate and values of each genomic variable across all genomes within a species to create species-specific indices. The ecological and life history factors were already standardized at the species level.

Species found in more variable environments were likely to grow faster

We initially investigate the extent to which ecological factors (environmental variability and host dependence) could explain the variation in bacterial growth rates. We first analysed the role of each variable separately. We revealed a positive correlation between environmental variability and growth rate (MCMCglmm (48); pMCMC = 0.016, n = 142 species, R^2 of fixed effect = 0.047; Figure 2a; Table S1). However, this correlation plateaued when environmental variability exceeded a value of 6 (Figure 2a).

Our next finding suggested that host dependence did not significantly influence growth rates (MCMCglmm; pMCMC = 0.236, n = 170 species; Figure 2b; Table S1). Post-hoc analysis showed no significant difference between host-dependent species and those classified as "both" (MCMCglmm; pMCMC = 0.750, Table S1). Host-dependent species did differ from free-living species (MCMCglmm; pMCMC = 0.016; Table S1), but due to the small sample size of free-living species (n = 3), this result is inconclusive.

We also evaluated the influence of genome fluidity by setting it as a covariate in three different models: model 1, growth rate ~ environmental variability + host dependence + genome fluidity; model 2, growth rate ~ environmental variability + genome fluidity; model 3, growth rate ~ host dependence + genome fluidity. In all three models, only environmental variability was consistently correlated with growth rate, indicating that genome fluidity had no significant impact on the relationship between growth rate and environmental variability (Figure 2c; Table S1).

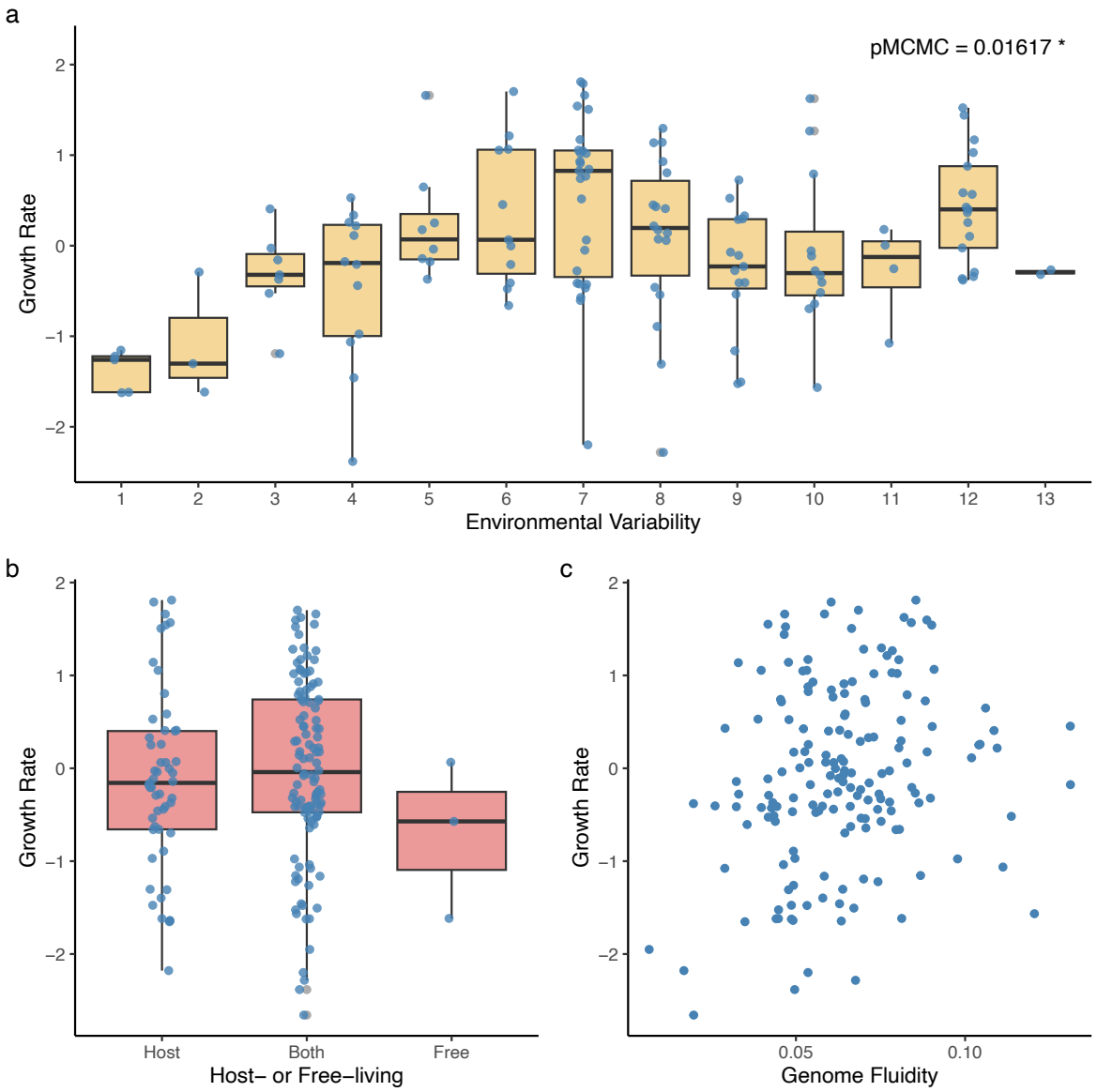


Figure 2. Relationship between growth rate and ecological factors. (a) Environmental variability vs. growth rate. (b) Host dependence vs. growth rate. (c) Genome fluidity vs. growth rate. Environmental variability was the only ecological variable that was positively correlated with growth rate across 142 species.

Genomic characteristics were positively correlated with growth rate

Next, we explored how genomic characteristics (genome size, metabolic capacity, and metabolic versatility) influence bacterial growth rates. When examined individually, all three genomic variables exhibited a positive correlation with growth rate (MCMCglmm; n = 170 species; genome size ~ growth rate: pMCMC = 0.003, R^2 of fixed effect = 0.133; Figure 3a; metabolic capacity ~ growth rate: pMCMC = 0.004, R^2 of fixed effect = = 0.089; Figure 3b; metabolic versatility ~ growth rate: pMCMC = 0.034, R^2 of fixed effect = 0.036; Figure 3c; Table S1).

To account for potential multicollinearity among these genomic feature variables, we first assessed their inter-relationships. Larger genomes were found to correlate with higher metabolic capacity and versatility (MCMCglmm; n = 171 species; pMCMC < 2e-04; Figure 3d-e; Table S1). Likewise, metabolic capacity and versatility were also positively correlated (MCMCglmm; pMCMC < 2e-04; Figure 3f; Table S1). To clarify whether these correlations influenced growth rate, we ran three multiple regression models: model 1, growth rate ~ genome size + metabolic capacity + metabolic versatility; model 2, growth rate ~ genome size + metabolic capacity; model 3, growth rate ~ genome size + metabolic versatility. In all models, only genome size remained consistently correlated with growth rate, suggesting the observed relationships between metabolic characteristics and growth rate were due to their multicollinearity with genome size (Table S1).

Lastly, we explored the role of genome fluidity as a covariate in the relationship between genome size and growth rate. We found that even when accounting for genome fluidity, genome size continued to positively correlate with growth rate, while genome fluidity itself showed no correlation with genome size (Figure S2, Table S1).

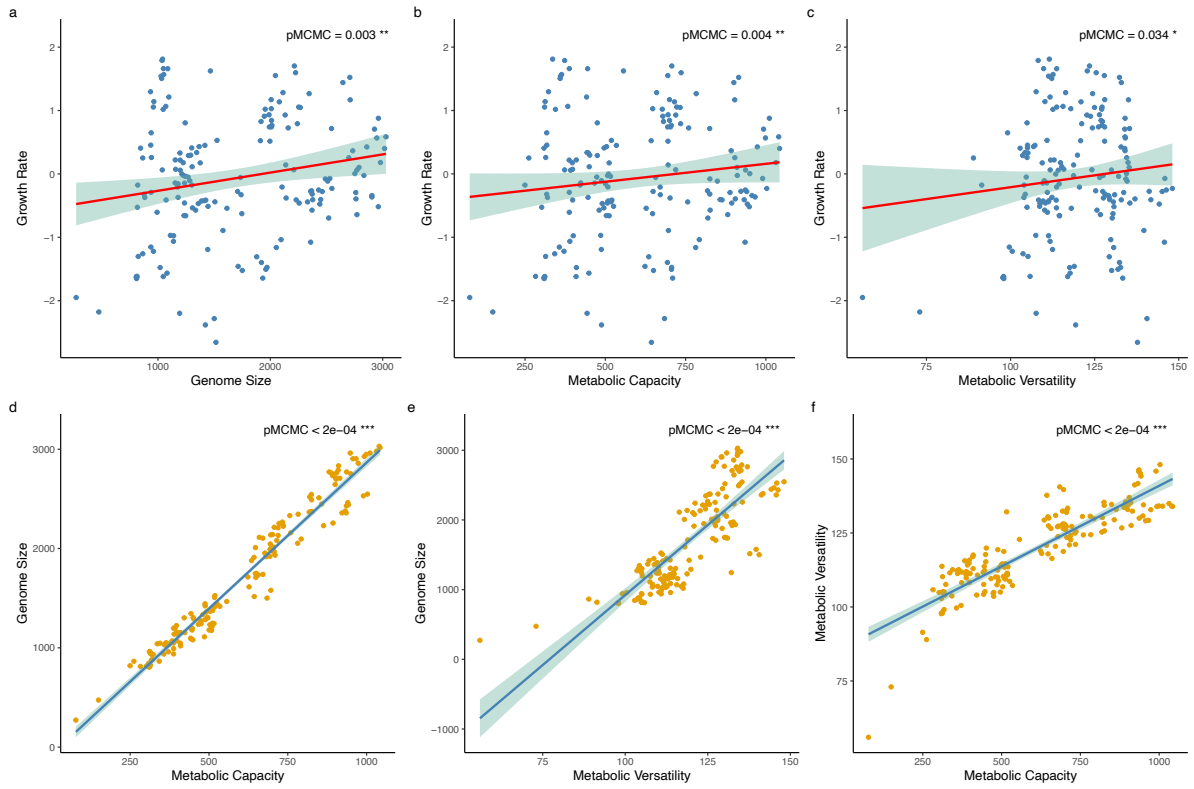


Figure 3. Relationship between growth rate and genomic characteristics. (a) Genome size vs. growth rate. (b) Metabolic capacity vs. growth rate. (c) Metabolic versatility vs. growth rate. (d) Genome size vs. metabolic capacity. (e) Genome size vs. metabolic versatility. (f) Metabolic versatility vs. metabolic capacity. All three genomic variables were found to be positively correlated with growth rate across 170 species when examined alone. However, only genome size was consistently correlated with growth rate after accounting for multicollinearity among them.

Environmental variability was positively correlated with all genomic characteristics

To examine the possibility that the observed relationships between bacterial growth rate and both ecological and genomic factors could be confounded by correlations between these ecological and genomic characteristics, we first conducted an analysis to identify if any such correlations exist.

We focused primarily on environmental variability, because we found host dependence did not affect bacterial growth. We found that when examined individually, species occupying a broader range of habitats tend to have larger genomes and greater metabolic capacity or versatility (MCMCglmm; n = 142 species; environmental variability ~ genome size: pMCMC < 2e-04, R^2 of fixed effect = 0.436; Figure 4a; environmental variability ~ metabolic capacity: pMCMC < 2e-04, R^2 of fixed effect = 0.441; Figure 4b; environmental variability ~ metabolic versatility: pMCMC < 2e-04, R^2 of fixed effect = 0.335; Figure 4c; Table S1). As for host dependence, we found that host dependence was not correlated with either metabolic capacity or versatility (Figure S3; Table S1). However, host dependence was negatively associated with genome size. Host-dependent species had smaller genomes compared to species categorized as "both" (Figure S3; Table S1).

When we controlled for genome fluidity as a covariate in these models, we still found that all genomic variables maintained a positive correlation with environmental variability. Meanwhile, genome fluidity itself showed no correlation with environmental variability across all models (Figure 4d, Table S1).

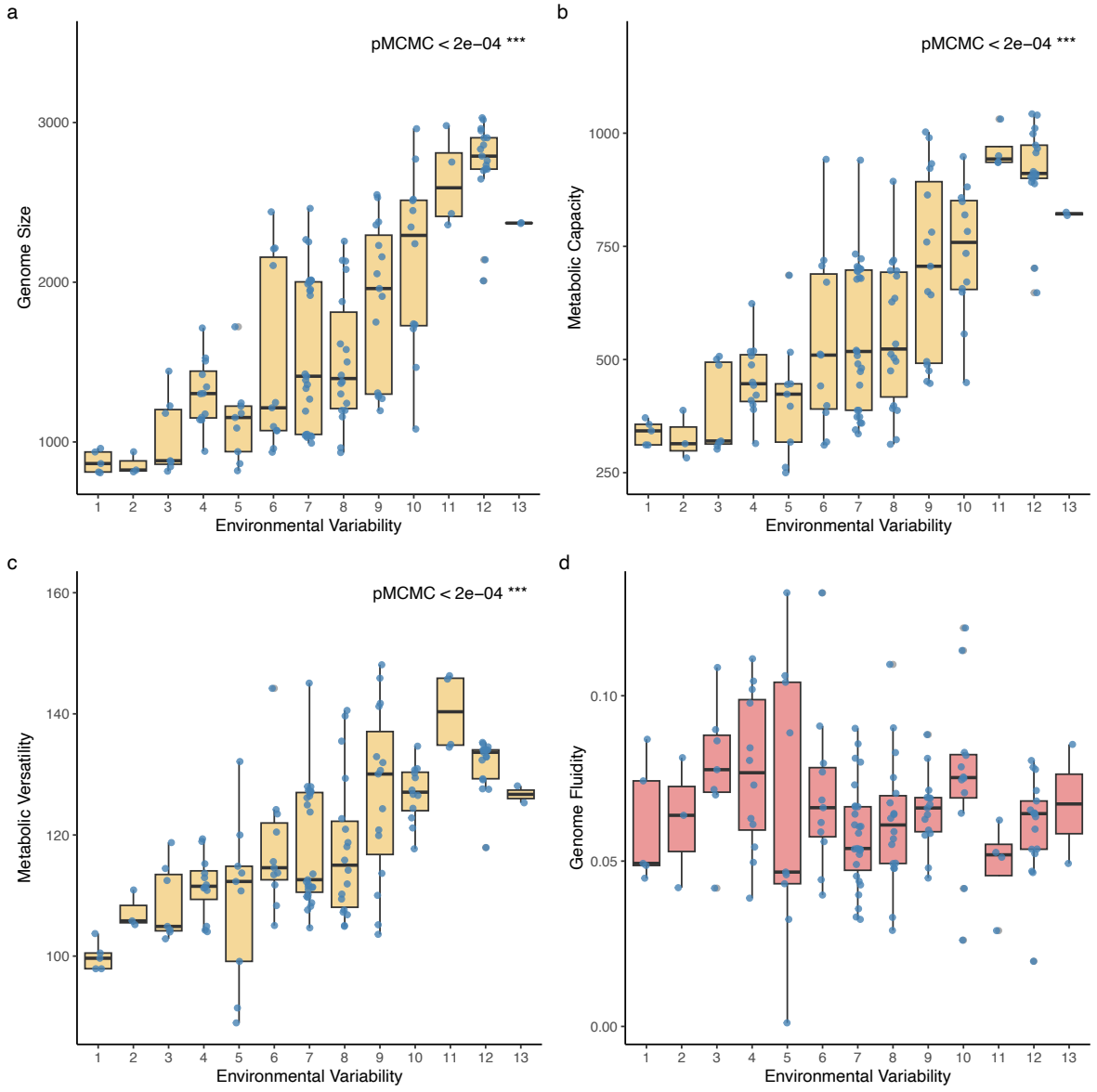


Figure 4. Relationship between environmental variability and genomic characteristics. (a) Genome size vs. environmental variability. (b) Metabolic capacity vs. environmental variability. (c) Metabolic versatility vs. environmental variability. (d) Genome fluidity vs. environmental variability. All three genomic variables were found to be positively correlated with environmental variability across 142 species when examined alone.

In summary, our findings presented that among ecological factors, only environmental variability had a positive correlation with bacterial growth rate. As for genomic characteristics, contrary to the hypotheses, we found no negative correlations between growth rate and genome size or metabolic properties. Environmental variability was also found to be positively correlated with all genomic characteristics. Lastly, host dependence was negatively correlated with genome size (Figure 5).

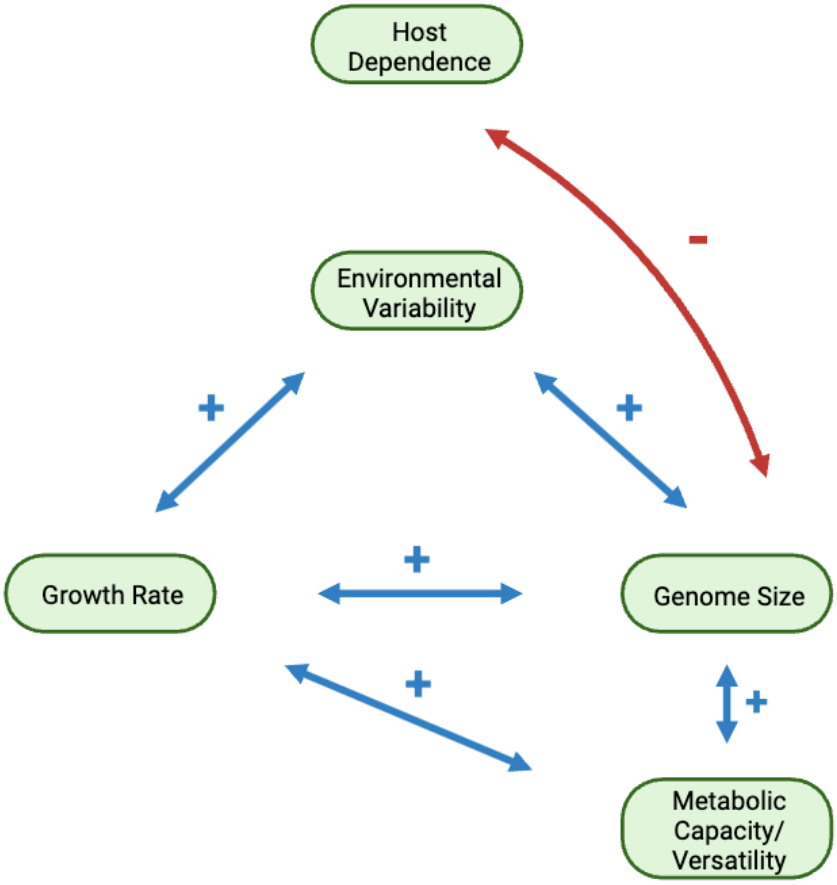


Figure 5. Revealed pairwise relationships between different variables differ from hypothesized predictions. Blue arrows link two variables to signify the detection of a positive correlation, whereas red arrow indicates a negative correlation. Notably, the results diverged from the expected outcomes: growth rate did not exhibit the anticipated negative correlation with genome size or metabolic properties.

Phylogenetic path analysis revealed causal models between growth rate, environmental variability, and genomic characteristics

Our prior findings implied that the observed relationships between bacterial growth rate and both environmental variability and genomic characteristics could be confounded by existing correlations between environmental variability and genomic characteristics. To comprehensively understand the interplay among growth rate, environmental variability, and genomic characteristics in evolutionary contexts, we used phylogenetic path analysis (PPA) to concurrently assess both direct and indirect effects among these variables while accounting for their phylogenetic relationships (50, 51). Path analysis is a form of structural equation modelling (SEM) that employs multivariate regression to test causal models (52, 53). This approach entailed: 1) formulating potential causal models; 2) implementing PPA and comparing model fits with Fisher's C, accepting models with P-values over 0.05; and 3) finalizing causal models based on averaged results of acceptable models, weighted by their relative evidence. We reported the final models here, while supplementary materials contain intermediate step outputs (Fig S4-7).

Previously, we discovered significant correlations between genome size and both metabolic capacity and versatility. Based on this, we initially explored the causal relationships among these three genomic variables. We identified significant strong direct causal links between genome size and metabolic capacity, as well as between metabolic capacity and versatility (Figure S4). These findings suggested that the correlations between genome size and metabolic versatility were mediated by metabolic capacity.

We next explored the causal relationships between growth rate, environmental variability, and genomic characteristics. Given the strong correlations among the three genomic characteristics,

we isolated one, like genome size, for each analysis. Subsequently, we verified if the resulting models were consistent across different genomic characteristics. This approach prevented the formulation of overly complex causal models with excessive variables, which typically complicated interpretation.

We discerned positive direct causal interactions between growth rate and environmental variability across models with varying genomic characteristics (Figure 6, a-c). Furthermore, we consistently observed positive direct causal links between environmental variability and genomic characteristics (Figure 6, a-c). Notably, we identified a significant yet weak positive direct causal link between growth rate and genome size (Figure 6a), but found no direct causal links between growth rate and either metabolic capacity or versatility. Our results suggested that with environmental variability as an intervening factor, the direct causal relationships between growth rate and genomic characteristics are either absent or weak. This implied that environmental variability could mediate the relationship between growth rates and genomic traits, possibly negating expected trade-offs between them.

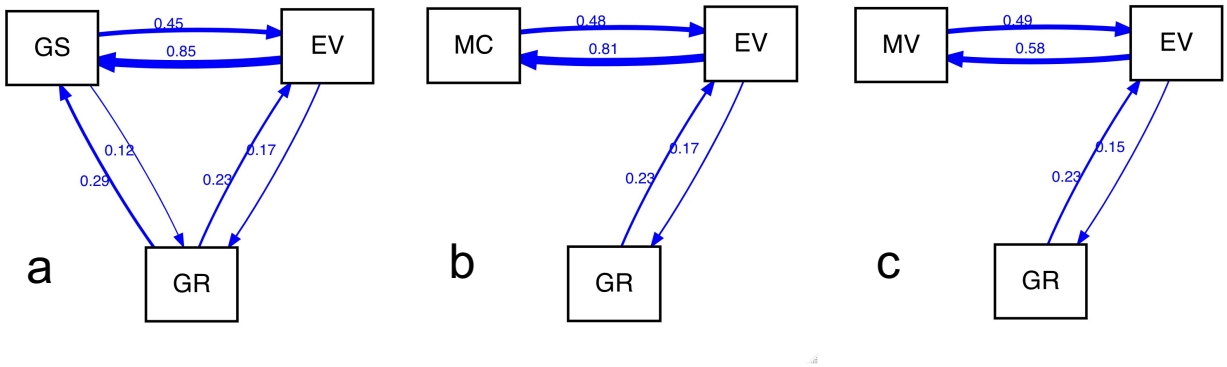


Figure 6. Results from path analyses displaying average best models for each variable group. (a - c) Genome level analyses illustrating the relationships between growth rate (GR), environmental variability (EV), genome size (GS), metabolic capacity (MC) and metabolic

versatility (MV). Blue arrows represent positive causations, with their width reflecting the magnitude of the standardized regression coefficients. Exact coefficient values were provided alongside the arrows.

Intra-specific correlations between growth rate and genomic characteristics varied across species

Our earlier analyses focused on inter-specific levels, but for a holistic view of patterns across evolutionary scales, we also assessed intra-specific (or strain-level) patterns. Using Pearson correlations, we studied the relationships between growth rate and genomic characteristics within different strains of each species. When investigated across species, the correlation coefficients between growth rate and genomic characteristics were not significantly different from zero, indicating that these correlations notably differed across species (MCMCglmm; [genome size ~ growth rate] ~ 0: pMCMC = 0.828; [metabolic capacity ~ growth rate] ~ 0: pMCMC = 0.991; [metabolic versatility ~ growth rate] ~ 0: pMCMC = 0.234; n = 141 species; Figure 7; Table S1). This observation suggested that in certain species, the anticipated trade-offs between growth rate and genomic characteristics were discernible at the strain level. However, in other species, these trade-offs either vanished or were in opposition to expectations. For instance, *Bifidobacterium pseudocatenulatum* displayed strong negative correlations between growth rate and genomic characteristics, whereas *Xylella fastidiosa* showed positive correlations (Figure 7, Table S1).

Finally, we examined whether environmental variability contributed to the variations of the intra-specific correlations between growth rate and genomic characteristics across different species. Our analysis found no significant effect of environmental variability on these correlations (MCMCglmm; [genome size ~ growth rate] ~ environmental variability: pMCMC

= 0.959; [metabolic capacity ~ growth rate] ~ environmental variability: pMCMC = 0.484; [metabolic versatility ~ growth rate] ~ environmental variability: pMCMC = 0.209; n = 141 species; Figure 7; Table S1).

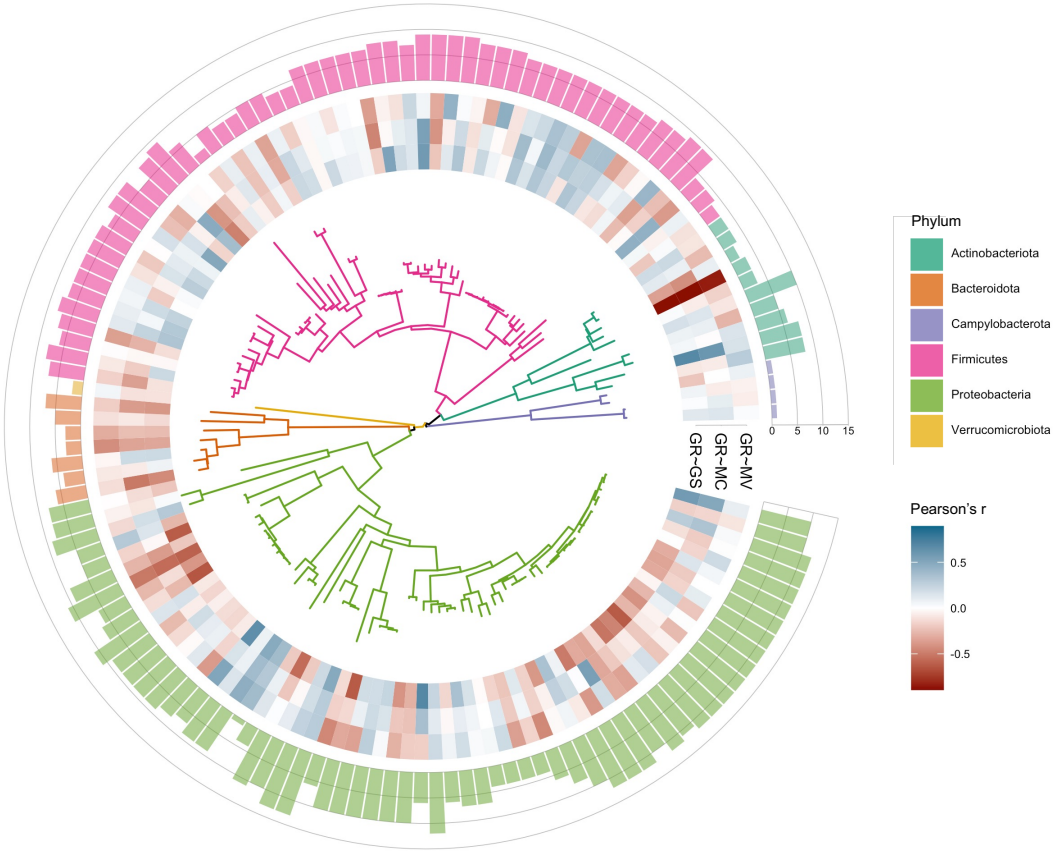


Figure 7. Intra-specific correlations between growth rate and genomic characteristics. The figure depicts a tree of 141 species, illustrating Pearson correlation coefficients via heatmap rings. GR ~ GS shows the correlation between growth rate and genome size; GR ~ MC represents the correlation between growth rate and metabolic capacity; and GR ~ MV indicates the relationship with metabolic versatility. Ring colours range from blue (high positive correlation) to red (high negative correlation) based on Pearson coefficients. Outer bar heights display environmental variability, with bar and tree clade colours denoting the species' taxonomic class (phylum).

Discussion

In this study, we investigated the influence of ecological factors and genomic characteristics on bacterial growth rates evolution. Using phylogenetic path analysis to account for the potential interplay of multiple variables, we identified positive direct correlations between growth rate and environmental variability, and between environmental variability and genomic characteristics (Figure 6). Contrary to prior hypotheses (17, 32), no trade-off was observed between growth rate and either genome size or metabolic properties (Figure 6).

We observed that species experiencing greater environmental variability tend to grow faster, as indicated in both our pairwise and path analyses (Figure 2a & 6). Earlier research found that generalist species grow faster than specialists, under the presumption that generalists inhabit more variable environments (12). While both findings support *r*- and *K*-selection theory predictions (9, 10), our study and the prior one differ in defining environmental variability. We emphasize the "abiotic environment," defining variability as the number of habitats a species can occupy; species with greater variability are termed generalists. Conversely, the prior study prioritized the "biotic environment," labelling generalists as species interacting with a wider range of other species (12). Despite these definitional differences, both approaches offered a perspective on the environmental variability concept (54).

Nonetheless, using *r*- and *K*-selection theory to understand our results presents challenges. First, the *r*-*K* paradigm was originally postulated to address the trade-offs between fitness in crowded versus uncrowded environments (9). Pianka's assertion that species in variable environments fall under *r*-selection, while those in constant environments fall under *K*-selection, did not directly stem from MacArthur and Wilson's original theory(10, 11). In scrutinizing our findings, it was evident that our environmental variability definition did not fully address density-

dependent regulation, including its associated factors such as resource availability and biotic interactions (9, 11, 55, 56). The earlier study, while considering biotic interactions, also overlooked density (12). Second, both the original theory by MacArthur and Wilson and Pianka's life-history distinctions have faced critique for oversimplifying natural selection (57). For instance, the *r/K* dichotomy may not capture the nuances of selection imposed by environments. Environments can vary in the degree to which they exert *r*- or *K*-selection pressures, and organisms might position themselves along a spectrum from pure *r*- to pure *K*-selection based on their exposure to these pressures (10, 11, 58). As a result, deviations from the *r-K* paradigm have been noted in both vertebrates and experimental *Escherichia coli* populations (59, 60). Our discovery, where growth rate did not consistently correlate with environmental variability, might be influenced by these complexities. Therefore, to understand discrepancies from the predictions of *r-K* paradigm, a deeper dive into species-specific life history traits is imperative.

Our subsequent analyses examined two potential trade-offs: between growth rate and genome size, and between growth rate and metabolism. Although the pairwise analysis hinted at weak positive correlations for these trade-offs, these correlations became negligible or remained weak in the path analyses (Figure 6). In both cases, we found no compelling evidence for these trade-offs. To elucidate the absence of anticipated trade-offs, we considered the potential mediating role of environmental variability, as both pairwise and path analysis validated the positive correlations between various genomic characteristics and environmental variability (Figure 4 & 6). Environmental variability imposed selective pressures that concurrently enhanced both genome size (or metabolism) and growth rate. This led to an indirect positive correlation between genome size (or metabolism) and growth rate. While direct negative correlations might exist between genomic characteristics and growth rate in the absence of

environmental variability as predicted, these correlations could be subtle and neutralized by the selective pressures introduced by environmental variability.

Another potential reason for the lack of trade-offs between growth rate and genomic features might be life history factors unrelated to environmental variability. These factors could also indirectly influence the relationship between genome size (or metabolism) and growth rate, leading to varying patterns across species. One such life history factor could be host dependence level. For example, *Mycobacterium leprae* possesses one of the smallest genomes among pathogenic bacteria yet exhibits a slower growth rate, with a doubling time of roughly 12-14 days (40, 41). This reduced growth rate might result from the species' obligate intracellular lifestyle, prioritizing survival within host tissues over swift proliferation (43). Yet, in our study, we did not find a significant correlation between host dependence and growth rate, or between host dependence and metabolic features. The only notable correlation was a negative one between genome size and host dependence. Based on these findings, it is uncertain whether host dependence has a role. A possible explanation for these outcomes is our species selection criteria: we chose species with at least 100 high-quality genomes from GTDB (61). This could mean our species set is skewed towards human pathogens, all of which are host-associated. In upcoming research, we aim to delve into the impact of host dependence by utilizing a more exhaustive database with a substantial number of free-living species. We also plan to explore the influence of other life history traits specific to host-associated species. For instance, the nature of the host relationship (whether it is obligate or facultative), the location within the host (intracellular or extracellular), and the impact on the host (whether it is pathogenic, mutualistic, or both) might all exert varying selective pressures on both genomic traits and bacterial growth rates (Dewar, A.E., 2023).

Indeed, the observed intra-specific correlations between genomic characteristics and growth rate varying among species suggested that there is more to the story than just a broad cross-species correlation (Figure 7). The variations might imply a more nuanced relationship at the strain level. While in some species the expected trade-offs might be evident at the strain level, in others they could be non-existent or even contrast with what one would predict. This underscored the possibility that distinct life history traits, inherent to each species, could influence the intra-specific relationship between genomic characteristics and growth rate. Environmental variability, as has been examined, doesn't appear to be one of these determinants (Figure 7). Therefore, it is crucial to expand the scope of investigation and consider other potential influencing factors. Host dependence, as well as other host-associated life history traits, could be a good starting point. By integrating more comprehensive data in subsequent analyses, we can delve deeper into understanding the nuances of these relationships and potentially uncover the underlying factors that dictate them.

Lastly, except for environmental variability and host dependence, other ecological and life history factors can potentially influence both genomic characteristics and bacterial growth rate, thereby shaping their relationship. For instance, biotic interactions, such as competition for resources, often drive bacteria towards optimizing their growth rate. However, a larger genome size, which may carry a broader array of functional genes, can give bacteria a competitive advantage when diverse resources are available (32, 34, 62). On the other hand, the Black Queen Hypothesis suggested that metabolic cross-feeding can lead to co-evolutionary genome reduction among species, and altering their growth dynamics (63). Moreover, other ecological factors, such as temperature (64), and life history traits like bacterial motility (65) and metabolic efficiency (66, 67), may play roles in affecting both growth and genomic characteristics, adding layers of complexity to the relationship.

To conclude, our findings offered potential explanations for the absence of evidence for two anticipated trade-offs: between growth rate and genome size, and between growth rate and metabolism. Our data indicated that environmental variability and species-specific life history traits can alter the relationship between growth rate and genomic characteristics, deviating from trade-off predictions. Further studies should investigate the impact of other ecological factors and life history traits on bacterial growth rate evolution and seek improved genomic markers to encapsulate these factors.

Methods

Genome collection and pangenome reconstruction

a) Species selection and genome collection

We retrieved a species list and their corresponding genomes from the Genome Taxonomy Database (GTDB (61); release 207) on 1st October 2022. To ensure high-quality of genomes, we applied a Minimum Information about a Metagenome-Assembled Genome (MIMAG) criterion (68), considering genomes with completeness ≥ 95% and contamination ≤ 5% as high-quality genomes. We focused on species that had at least 100 high-quality genomes to enable pangenome reconstruction in subsequent steps. This process resulted in a final species list comprising 171 species. We then used ncbi-genome-download scripts (version 0.3.1) to download the available amino acid sequences for all genomes (files named as "*_protein.faa.gz," where the asterisk represents an arbitrary string) from either GenBank (69) or RefSeq (70) database using the assembly entries specified in GTDB. As a result, we obtained a total of 149,932 genomes with their respective amino acid sequences for further analysis.

b) Pangenome reconstruction

Pangenome represents all the genes found in a species across all its genomes (46). A typical pangenome clustering pipeline involves three major steps: (1) gene prediction and annotation; (2) homologous genes identification; (3) determining core and accessory genes based on gene family presence/absence profiles (71, 72). For this study, we followed the protocol of KEGG Orthology (KO) database (73) to predict genes and identify orthologs for all genomes. KofamScan (74) was used to assign KO identifiers to protein sequences, grouping genes with the same KO identifiers into gene families. The presence/absence profile of each gene family across all genomes of every species was thus represented by the presence/absence of each KO.

Genome fluidity estimation

Genome fluidity quantifies the intra-specific gene content variability. It was determined by calculating the ratio of unique gene families to the union of gene families in pairs of genomes, averaged over randomly selected genome pairs from all genomes of a given species (75). This metric ranges between 0 and 1, where 0 indicates that all genomes of a species share the same gene families, and 1 indicates the absence of any shared gene families among the genomes.

For species with a large number of genomes such as *Escherichia coli* (25981 genomes in the dataset), calculating pairwise ratios for all selected genome pairs was computationally intensive. To address this, we applied Central Limit Theorem (CLT) to estimate the mean pairwise ratio (i.e., genome fluidity) for all the species' genomes. This estimation involved three steps: (1) randomly sampling 30 genomes 20,000 times, (2) calculating the mean pairwise ratio for each sample, and (3) using the mean value of all the sample means as the estimation of genome fluidity. According to the CLT, the average of sample means will converge to the population mean when the sample size is sufficiently large. As an illustrative example, we can examine the distribution of sample means for *Escherichia coli* (Fig. S8).

Estimation of environmental variability

We directly utilized pre-established habitat clusters to categorize the habitat preferences of each species in our dataset (see in Chapter 3). To evaluate the environmental variability of each species, we examined the number of habitat clusters in which each species could be found. This measurement of species environmental variability allowed us to understand the range of ecological niches that each species can inhabit. The distribution of environmental variability among our species can be found in Figure S9.

Estimation of genomic characteristics

We defined metabolic KOs as KOs that can be mapped to pathways in "Metabolism" category of KEGG PATHWAY Database (73). Within the "Metabolism" category, there are 189 pathways classified into 13 different classes. We retrieved associations data between KO (as represented by KO identifiers starting with "ko") to pathway (as represented by pathway identifiers starting with "map") from KEGG API (https://rest.kegg.jp/link/ko/pathway). We used Kofamscan for annotating KO identifiers to all genes in our dataset (74). Genes that could be assigned to at least one metabolic KO were considered as metabolic genes. To maintain consistency with our previous definition of pangenome, we directly used KOs to represent gene content, thus characterizing the metabolic gene content of the species under investigation.

Genome size was defined as the average number of unique KOs per genome per species. To explore the metabolic potential and functional diversity of the species in our dataset, we introduced two metabolic properties: metabolic capacity and metabolic versatility. Metabolic capacity was defined as the average number of unique metabolic KOs per genome per species, representing the species' ability to carry out metabolic reactions and processes. On the other hand, metabolic versatility was defined as the average number of unique metabolic pathways per genome per species, reflecting the species' potential to utilize a broad spectrum of different compounds as energy sources and nutrients.

Annotation of host dependence levels

We sourced information on host dependence from Madin's dataset (49). Based on this dataset, we categorized species exclusively found in host-associated environments as "host-dependent", those only present in non-host-associated environments as "free-living", and species found in both types of environments as having an "both" level of host dependence.

Growth rate estimation

The estimation of maximal microbial growth rates using genome-only data was pioneered by Vieira-Silva and Rocha (30). Their analysis revealed a strong correlation between high codon usage bias (CUB) in genes coding for ribosomal proteins and other highly expressed genes with high maximal growth rates. Building upon this work, a subsequent study expanded the investigation by assessing different dimensions of codon usage patterns to predict maximal growth rates from genomic data (48). In our study, we obtained the predicted maximal growth rates of genomes in our dataset from a comprehensive database EGGO (48), established by the aforementioned researchers. The maximal growth rates were represented by minimal doubling time in hours, with lower values indicating higher growth rates. To enhance understanding, we took the negative logarithm of the value. The higher the resulting value, the greater the growth rate. To characterize species-level maximal growth rates, we calculated the mean of the predicted growth rates for all genomes belonging to each species, resulting in a dataset of 170 species with their respective species-level maximal growth rates.

Phylogeny and statistics

a) Phylogeny

The phylogenetic reference tree of bacterial species was downloaded from GTDB (release 207, https://data.gtdb.ecogenomic.org/releases/release207/207.0/bac120_r207.tree). To focus on the 171 species relevant to our study, we pruned the phylogenetic tree using the "*keep.tip*" function in the R package "*ape*" (76). The dendrogram of the pruned tree, displaying the evolutionary relationships among the selected species, can be found in the supplementary material (Fig S10).

b) MCMCglmm

We used R package "*MCMCglmm*" to examine all pairwise correlations between variables in our study (77). MCMCglmm allowed us to fit generalized linear mixed-effects models (GLMMs) using a Markov chain Monte Carlo approach with a Bayesian statistical framework. To account for potential phylogenetic nonindependence among the species in our dataset, we incorporated the phylogeny as a random effect in the model (77). To ensure that the phylogenetic tree used in the MCMCglmm fitting had an ultrametric structure, we applied the "*force.ultrametric*" function in the R package "*phytools*" (with an option *"method = extend"*) to transfer our tree into an ultrametric form (78). To assess the goodness of fit for each model, we reported the *pMCMC* value (referred to as "p-value") and the R^2 of the fixed effect.

c) Phylogenetic path analysis

In our study, we utilized the R package "phylopath" to conduct phylogenetic path analysis (PPA) to explore the causal relationships between multiple variables (51). Path analysis is a form of structural equation modelling (SEM) that employs multivariate regression to test causal models (52, 53). Phylogenetic path analysis (PPA) extends this approach by utilizing phylogenetic regression methods, such as phylogenetic generalized least-squares (PGLS) models, to perform path analysis while accounting for phylogenetic nonindependence (51, 79). In our analysis, we used Pagel's lambda model by default to simulate the evolution of variables.

We used Fisher's C to evaluate the global goodness-of-fit of PPA models. A model is considered acceptable if its P-value of Fisher's C is greater than 0.05 (51, 52). Additionally, we used a modified version of the AIC, known as the C statistic Information Criterion (CIC), along with corresponding weights (w) calculated based on their likelihood to perform model selection.

Higher weights (w) indicate better model performance, aiding us in selecting the most suitable path for a given set of variables (50).

References

1. G. Sezonov, D. Joseleau-Petit, R. D'Ari, Escherichia coli Physiology in Luria-Bertani Broth. *Journal of Bacteriology* **189**, 8746–8749 (2007).

2. T. L. Testerman, D. J. McGee, H. L. T. Mobley, Helicobacter pylori Growth and Urease Detection in the Chemically Defined Medium Ham's F-12 Nutrient Mixture. *Journal of Clinical Microbiology* **39**, 3842–3850 (2001).

3. P. Starnawski, *et al.*, Microbial community assembly and evolution in subseafloor sediment. *Proceedings of the National Academy of Sciences* **114**, 2940–2945 (2017).

4. E. Trembath-Reichert, *et al.*, Methyl-compound use and slow growth characterize microbial life in 2-km-deep subseafloor coal and shale beds. *Proceedings of the National Academy of Sciences* **114**, E9206–E9215 (2017).

5. L. Dethlefsen, T. M. Schmidt, Performance of the Translational Apparatus Varies with the Ecological Strategies of Bacteria. *Journal of Bacteriology* **189**, 3237–3245 (2007).

6. B. S. Stevenson, T. M. Schmidt, Life History Implications of rRNA Gene Copy Number in Escherichia coli. *Applied and Environmental Microbiology* **70**, 6670–6677 (2004).

7. S. Freilich, *et al.*, Metabolic-network-driven analysis of bacterial ecological strategies. *Genome Biology* **10**, R61 (2009).

8. D. A. Ratkowsky, J. Olley, T. A. McMeekin, A. Ball, Relationship between temperature and growth rate of bacterial cultures. *Journal of Bacteriology* **149**, 1–5 (1982).

9. R. H. Macarthur, E. O. Wilson, *The Theory of Island Biogeography*, REV-Revised (Princeton University Press, 1967) (August 14, 2023).

10. E. R. Pianka, On r- and K-Selection. *The American Naturalist* **104**, 592–597 (1970).

11. D. Reznick, M. J. Bryant, F. Bashey, r- AND K-SELECTION REVISITED: THE ROLE OF POPULATION REGULATION IN LIFE-HISTORY EVOLUTION. *Ecology* **83**, 1509–1520 (2002).

12. F. A. B. von Meijenfeldt, P. Hogeweg, B. E. Dutilh, A social niche breadth score reveals niche range strategies of generalists and specialists. *Nat Ecol Evol* **7**, 768–781 (2023).

13. M. A. Brockhurst, A. Buckling, D. Racey, A. Gardner, Resource supply and the evolution of public-goods cooperation in bacteria. *BMC Biology* **6**, 20 (2008).

14. T. Korem, *et al.*, Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **349**, 1101–1106 (2015).

15. H. L. Cash, C. V. Whitham, C. L. Behrendt, L. V. Hooper, Symbiotic Bacteria Direct Expression of an Intestinal Bactericidal Lectin. *Science* **313**, 1126–1130 (2006).

16. B. Gibson, D. J. Wilson, E. Feil, A. Eyre-Walker, The distribution of bacterial doubling times in the wild. *Proceedings of the Royal Society B: Biological Sciences* **285**, 20180789 (2018).

17. D. O. Hessen, P. D. Jeyasingh, M. Neiman, L. J. Weider, Genome streamlining and the elemental costs of growth. *Trends in Ecology & Evolution* **25**, 75–80 (2010).

18. M. Lynch, Streamlining and Simplification of Microbial Genome Architecture. *Annual Review of Microbiology* **60**, 327–349 (2006).

19. S. J. Giovannoni, J. Cameron Thrash, B. Temperton, Implications of streamlining theory for microbial ecology. *ISME J* **8**, 1553–1565 (2014).

20. J. j. Elser, *et al.*, Biological stoichiometry from genes to ecosystems. *Ecology Letters* **3**, 540–550 (2000).

21. J. J. Elser, *et al.*, Growth rate–stoichiometry couplings in diverse biota. *Ecology Letters* **6**, 936–943 (2003).

22. T. R. Gregory, *The Evolution of the Genome* (Elsevier, 2011).

23. M. D. Bennett, I. J. Leitch, "CHAPTER 2 - Genome Size Evolution in Plants" in *The Evolution of the Genome*, T. R. Gregory, Ed. (Academic Press, 2005), pp. 89–162.

24. M. D. Bennett, Variation in Genomic Form in Plants and Its Ecological Implications. *New Phytologist* **106**, 177–200 (1987).

25. M.-C. Lee, C. J. Marx, Repeated, Selection-Driven Genome Reduction of Accessory Genes in Experimental Populations. *PLOS Genetics* **8**, e1002651 (2012).

26. Y. I. Wolf, E. V. Koonin, Genome reduction as the dominant mode of evolution. *BioEssays* **35**, 829–837 (2013).

27. J. Li, *et al.*, Predictive genomic traits for bacterial growth in culture versus actual growth in soil. *ISME J* **13**, 2162–2172 (2019).

28. A. Mira, H. Ochman, N. A. Moran, Deletional bias and the evolution of bacterial genomes. *Trends in Genetics* **17**, 589–596 (2001).

29. M. Touchon, E. P. C. Rocha, Causes of Insertion Sequences Abundance in Prokaryotic Genomes. *Molecular Biology and Evolution* **24**, 969–981 (2007).

30. S. Vieira-Silva, E. P. C. Rocha, The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLOS Genetics* **6**, e1000808 (2010).

31. T. Pfeiffer, S. Schuster, S. Bonhoeffer, Cooperation and Competition in the Evolution of ATP-Producing Pathways. *Science* **292**, 504–507 (2001).

32. D. Molenaar, R. van Berlo, D. de Ridder, B. Teusink, Shifts in growth strategies reflect tradeoffs in cellular economics. *Molecular Systems Biology* **5**, 323 (2009).

33. M. Basan, *et al.*, Overflow metabolism in Escherichia coli results from efficient proteome allocation. *Nature* **528**, 99–104 (2015).

34. M. F. Polz, O. X. Cordero, Bacterial evolution: Genomics of metabolic trade-offs. *Nat Microbiol* **1**, 1–2 (2016).

35. B. R. K. Roller, S. F. Stoddard, T. M. Schmidt, Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nat Microbiol* **1**, 1–7 (2016).

36. E. Costa, J. Pérez, J.-U. Kreft, Why is metabolic labour divided in nitrification? *Trends in Microbiology* **14**, 213–219 (2006).

37. C. Pál, B. Papp, M. J. Lercher, Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* **37**, 1372–1375 (2005).

38. J. P. McCutcheon, N. A. Moran, Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* **10**, 13–26 (2012).

39. P. Bentkowski, C. Van Oosterhout, T. Mock, A Model of Genome Size Evolution for Prokaryotes in Stable and Fluctuating Environments. *Genome Biology and Evolution* **7**, 2344–2351 (2015).

40. S. T. Cole, *et al.*, Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011 (2001).

41. admin, Cultivation and Viability Determination of Mycobacterium leprae. *International Textbook of Leprosy* (2016) (August 15, 2023).

42. S. Shigenobu, H. Watanabe, M. Hattori, Y. Sakaki, H. Ishikawa, Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. *Nature* **407**, 81–86 (2000).

43. P. Singh, *et al.*, Insight into the evolution and origin of leprosy bacilli from the genome sequence of Mycobacterium lepromatosis. *Proceedings of the National Academy of Sciences* **112**, 4459–4464 (2015).

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

45. J. O. McInerney, A. McNally, M. J. O'Connell, Why prokaryotes have pangenomes. *Nat Microbiol* **2**, 1–5 (2017).

46. M. A. Brockhurst, *et al.*, The Ecology and Evolution of Pangenomes. *Current Biology* **29**, R1094–R1103 (2019).

47. O. M. Maistrenko, *et al.*, Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *The ISME Journal* **14**, 1247–1259 (2020).

48. J. L. Weissman, S. Hou, J. A. Fuhrman, Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proceedings of the National Academy of Sciences* **118**, e2016810118 (2021).

49. J. S. Madin, *et al.*, A synthesis of bacterial and archaeal phenotypic trait data. *Sci Data* **7**, 170 (2020).

50. A. Gonzalez-Voyer, A. von Hardenberg, "An Introduction to Phylogenetic Path Analysis" in *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*, L. Z. Garamszegi, Ed. (Springer, 2014), pp. 201–229.

51. W. van der Bijl, phylopath: Easy phylogenetic path analysis in R. *PeerJ* **6**, e4718 (2018).

52. B. Shipley, A New Inferential Test for Path Models Based on Directed Acyclic Graphs. *Structural Equation Modeling: A Multidisciplinary Journal* **7**, 206–218 (2000).

53. B. Shipley, Confirmatory path analysis in a generalized multilevel context. *Ecology* **90**, 363–368 (2009).

54. J. P. Sexton, J. Montiel, J. E. Shay, M. R. Stephens, R. A. Slatyer, Evolution of Ecological Niche Breadth. *Annual Review of Ecology, Evolution, and Systematics* **48**, 183–206 (2017).

55. J. H. Andrews, R. F. Harris, "r- and K-Selection and Microbial Ecology" in *Advances in Microbial Ecology*, Advances in Microbial Ecology., K. C. Marshall, Ed. (Springer US, 1986), pp. 99–147.

56. M. E. Hibbing, C. Fuqua, M. R. Parsek, S. B. Peterson, Bacterial competition: surviving and thriving in the microbial jungle. *Nat Rev Microbiol* **8**, 15–25 (2010).

57. S. C. Stearns, S. C. Stearns, *The Evolution of Life Histories* (Oxford University Press, 1992).

58. , *Evolutionary Ecology* (Eric R. Pianka, 2011).

59. A. M. Bronikowski, EXPERIMENTAL EVIDENCE FOR THE ADAPTIVE EVOLUTION OF GROWTH RATE IN THE GARTER SNAKE THAMNOPHIS ELEGANS. *Evolution* **54**, 1760–1767 (2000).

60. L. S. Luckinbill, r and K Selection in Experimental Populations of Escherichia coli. *Science* **202**, 1201–1203 (1978).

61. D. H. Parks, *et al.*, GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* **50**, D785–D794 (2022).

62. M. F. Polz, E. J. Alm, W. P. Hanage, Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics* **29**, 170–175 (2013).

63. J. J. Morris, R. E. Lenski, E. R. Zinser, The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. *mBio* **3**, e00036-12 (2012).

64. S. Lax, C. I. Abreu, J. Gore, Higher temperatures generically favour slower-growing bacterial species in multispecies communities. *Nat Ecol Evol* **4**, 560–567 (2020).

65. S. Gude, *et al.*, Bacterial coexistence driven by motility and spatial competition. *Nature* **578**, 588–592 (2020).

66. T. M. Hoehler, B. B. Jørgensen, Microbial life under extreme energy limitation. *Nat Rev Microbiol* **11**, 83–94 (2013).

67. B. R. Roller, T. M. Schmidt, The physiology and ecological implications of efficient growth. *ISME J* **9**, 1481–1487 (2015).

68. R. M. Bowers, *et al.*, Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**, 725–731 (2017).

69. D. A. Benson, *et al.*, GenBank. *Nucleic Acids Research* **41**, D36–D42 (2013).

70. N. A. O'Leary, *et al.*, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733–D745 (2016).

71. A. J. Page, *et al.*, Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

72. G. Tonkin-Hill, *et al.*, Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020).

73. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457–D462 (2016).

74. T. Aramaki, *et al.*, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

75. A. O. Kislyuk, B. Haegeman, N. H. Bergman, J. S. Weitz, Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* **12**, 32 (2011).

76. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

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

78. L. J. Revell, phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**, 217–223 (2012).

79. C. M. Mason, E. W. Goolsby, D. P. Humphreys, L. A. Donovan, Phylogenetic structural equation modelling reveals no need for an 'origin' of the leaf economics spectrum. *Ecology Letters* **19**, 54–61 (2016).

Chapter 5: Discussion

In the thesis, I have studied the role of cooperation in bacterial ecological and evolutionary dynamics through a comparative genomics lens. My thesis began with an investigation into the influence of cooperation on the evolution of bacterial niche breadth (Chapter 2). Subsequently, I investigated the interplay between horizontal gene transfer and bacterial cooperation, focusing particularly on the interactions between gene connectivity and sociality in shaping gene transferability (Chapter 3). Broadening the perspective, I investigated how ecological factors and genomic characteristics influence the evolution of bacterial growth rates (Chapter 4). While individual chapters offer detailed discussions, this section aims to encapsulate the major findings, draw insights, and underscore limitations not previously touched upon. Finally, I will discuss potential avenues for future exploration, focusing on strategies to optimize the use of genomic datasets in unravelling the mysteries of bacterial ecology and evolution.

5.1 Bacterial Cooperation and Niche Breadth Evolution

In this chapter, we observed a positive correlation between the proportion of cooperative genes and bacterial niche breadth (Chapter 2, Figure 1). This correlation appeared to be causal: a reduced proportion of cooperative genes promotes niche contraction (Chapter 2, Figure 4). The question then arises, why does a lower proportion of cooperative genes result in specialization rather than a higher proportion promoting generalization? We suggest that although the gain of cooperative genes might offer benefits enabling specialists to expand their niches, the associated costs might result in strong selective pressures for their rapid loss. Thus, the frequent turnover of cooperative genes might be too swift to influence the long-term transition from specialists to generalists. To test this idea, we analysed the rates of cooperative gene gains and losses in both short-term and long-term evolution. We found that while cooperative genes are quickly gained and lost in the short-term, this is not the case over longer evolutionary scales,

supporting our hypothesis. Furthermore, the loss rates of these genes exceed their gain rates, further suggesting that the frequent loss of cooperative genes favours niche contraction (Chapter 2, Figure 5).

Our results and interpretations stem from a broad pattern across all cooperative genes we identified, encompassing biofilm formation, quorum-sensing, secretion systems, siderophore production, and antibiotic degradation. However, patterns of specific gene types may diverge from this general trend due to their unique properties or functions. Take biofilm formation genes as an example. Although we observed a positive correlation between the proportion of these genes and bacterial niche breadth, a higher proportion actually impeded niche expansion. Additionally, biofilm formation genes displayed lower gain and loss rates compared to private genes, and unlike other cooperative genes, their short-term evolutionary rates were not notably higher.

This insight into biofilm formation genes from our results is enlightening. First, the fact that a higher proportion of biofilm formation genes impedes niche expansion indicates these genes might confer advantages in specialized environments over diverse niches. The reason could be the biophysical properties of biofilms. In the majority of biofilms, microorganisms constitute less than 10% of the dry mass, while over 90% is made up of the extracellular matrix, primarily composed of extracellular polymeric substances (EPS) that structure the biofilm[1]. While EPS can protect biofilm bacteria from various ecological stressors, promoting local adaptation[2–5], it can also immobilize biofilm cells and limit bacterial dispersion to other niches which is essential for habitat generalization[1]. Even if bacterial species like *Pseudomonas aeruginosa* alternate between biofilm and planktonic modes, such 'dispersal' lifestyles are species-specific and triggered by genetic and environmental cues like nutrient availability[6–8]. Consequently, it's

essential to compare species with diverse biofilm lifestyles and evaluate the influence of various environmental contexts in understanding the role of biofilm formation genes in niche breadth evolution.

Second, the limited gains and losses observed in biofilm formation genes suggest their high conservation, pointing to their critical role in bacterial functions. As a result, generalist descendants of specialist ancestors might have a reduced proportion of biofilm formation genes, not due to these genes being lost or gained, but because as generalists expand their genomes, the relative proportion of these genes decreases. To examine if this is correct, the understanding of the essentiality of biofilm formation genes is crucial. While biofilms are undeniably pivotal in bacterial growth and protection against ecological threats such as antibiotics and host immune responses[1], direct evidence of their essentiality remains elusive. Only the indirectly inferred essentiality of biofilms is underscored by their widespread presence globally. Biofilms are estimated to prevail in nearly all terrestrial habitats, with the exception of oceans, comprising about 80% of bacterial and archaeal cells[9]. This pervasive presence implies biofilms are central to nearly all biogeochemical processes and represent the predominant mode of active life for bacteria and archaea[10]. To understand the impact of the essentiality of biofilm formation genes on their gain/loss dynamics and subsequent effects on bacterial niche breadth evolution, future research could evaluate the essentiality by analysing the evolutionary rates of these gene sequences or by assessing their centrality in gene networks. Typically, genes exhibiting slower evolutionary rates and greater centrality tend to be essential[11,12].

Third, while bacteria in biofilms often exhibit close cooperative interactions, they might also engage in antagonistic relationships. The type of these interactions within biofilms are likely determined by the spatial arrangement of various clones, strains, and species. When biofilms

segregate into clonal clusters, a cell's neighbours tend to be of the same clone, favouring the secretion of universally beneficial proteins. In contrast, a diverse mix of species within biofilms can promote antagonistic behaviour[13,14]. If biofilms predominantly foster interspecific competition over intraspecific cooperation, their impact on bacterial niche breadth becomes multifaceted, potentially influenced by factors such as species abundance. For instance, a theory posits that in the presence of interspecific competition, specialization is favoured when species are locally rare, and generalization is favoured when species are locally abundant[15]. Consequently, understanding biofilms' impact on niche breadth evolution may necessitate a detailed examination of population and community structures to determine whether biofilms predominantly encourage cooperation or competition.

Subsequently, we identified 357 orthologs that co-evolved with cooperative orthologs and examined the combined impact of cooperative genes and these co-evolved genes on bacterial niche breadth evolution. Our findings indicated that while cooperative genes promote niche expansion, the presence of co-evolved genes hampers this effect. Instead, co-evolved genes mainly serve to limit the niche breadth of existing generalists (Chapter 2, Figure 7-8).

Identifying genes that co-evolved with target genes is a challenging task. The way people define co-evolved genes can significantly shape the observed patterns associated with them. Our current findings might be subjected to this complexity. We characterized co-evolved ortholog groups (OGs) as those sharing similar phylogenetic distribution with cooperative OGs across the tree of life. Given that a non-cooperative OG can co-occur with multiple cooperative OGs, we set a criterion where non-cooperative OGs were deemed co-evolved if they significantly co-occurred with at least 270 cooperative OGs. A potential pitfall here is if the genes we identified as co-evolved are simply housekeeping genes[16]. These genes, fundamental

for cellular functions, would naturally be retained by most species, explaining their frequent co-occurrence with diverse cooperative genes. In support of this hypothesis, we found that although only 357 orthologs were tagged as co-evolved, they represent a substantial portion of genomes across species. The proportion of co-evolved genes in representative genomes of species vary between 0.11 and 0.71. In contrast, the cooperative genes' proportion ranged only from 0 to 0.05. We also found these genes do not have functional associations but share habitat preferences with cooperative genes (Chapter 2, Figure 6). If indeed they are housekeeping genes, their prevalence across various habitats, despite not having direct functional associations with cooperative genes, might result in shared habitat preferences with cooperative genes.

To address the potential shortcomings of our initial definition, we can implement a refined approach in future analyses. In this revised method, we would determine a co-evolved ortholog for each individual cooperative ortholog. Specifically, for a given cooperative ortholog, we would identify a non-cooperative ortholog with the most closely matching phylogenetic distribution pattern and designate it as the co-evolved counterpart for that specific cooperative ortholog. This approach ensures that the co-evolved gene of one cooperative gene is less likely to be the co-evolved gene of another. By adopting this definition, we aim to circumvent the potential bias of categorizing housekeeping genes as co-evolved. We anticipate that with this definition, there will be stronger functional associations between the newly defined co-evolved genes and their corresponding cooperative genes.

To summarise, in chapter 2, I pioneered the identification of the causative mechanism of cooperation in the evolution of bacterial niche breadth. The causal inference approach deployed here holds promise for wider applications across various topics within microbial ecology and evolution. Furthermore, our endeavour to evaluate the combined impact of cooperative genes

and their co-evolved counterparts presents a valuable perspective to truly comprehend genetic consequences in bacteria. Conventionally, research has primarily focused on specific gene types, examining their individual effects. Yet, genes don't operate in isolation; they function in gene networks, and selective pressures seldom target singular genes[12]. Recognizing the collective effects of interlinked genes is pivotal for a comprehensive understanding of myriad ecological and evolutionary processes. While our current methodology can benefit from refinements, the exploration of multi-gene coevolution and its implications undoubtedly represents a promising research avenue.

5.2 Gene transferability and sociality do not correlate with gene connectivity

In this chapter, I studied the interplay between gene connectivity and sociality, and their collective impact on gene transferability. The major findings of this study are: (i) as predicted by the complexity hypothesis, plasmid genes had consistently lower connectivity compared to chromosome genes (Chapter 3, Figure 2); (ii) contrary to the prediction of the complexity hypothesis, there was no correlation between plasmid mobility and gene connectivity (Chapter 3, Figure 3); and (iii) genes encoding extracellular proteins and genes encoding intracellular proteins did not differ in relative gene connectivity between chromosomes and plasmids (Chapter 3, Figure 4).

In this chapter, we have chosen to utilize genes encoding for extracellular proteins as a proxy of cooperative genes. Using extracellular protein-coding genes as a proxy has its clear limitations. First, not all genes that code for extracellular proteins are necessarily involved in bacterial cooperation. Second, this proxy might overlook several cooperative traits that don't fit this narrow definition. Third, such a definition doesn't offer any insights into the benefits or costs associated with carrying these genes or the consequences of producing their products.

The third limitation might be an inherent limitation with all bioinformatics tools attempting to pinpoint cooperative genes solely based on genomic data. Recent advancements have presented methods that address the shortcomings of this approach[17]. In chapter 2, I expanded on the definition of cooperative genes by considering their various types. Applying this enhanced definition can also shed light on the patterns discussed in the current chapter. If different types of cooperative genes play distinct roles in associating with gene transferability and connectivity, it could offer new perspectives on the functional variations of these genes, and could unearth novel inspiration in microbial cooperation. Later, I will also discuss some ideas on refining our current methodologies to better identify and understand cooperative genes.

Chapter 3 also revealed a crucial finding: there's no direct correlation between plasmid mobility and gene connectivity. This challenged the conventional view of plasmids as merely vectors for horizontal transfer, especially given recent findings that bacterial chromosomes can also mobilize, sometimes even faster than plasmids[18]. It's evident that plasmids and chromosomes have evolved differently, with distinct evolutionary purposes. Therefore, they differ not only in transfer rates but also in how selection shapes their traits[19,20]. In this chapter, we suggested that the stability of genes on plasmids might be more influential in determining gene connectivity than plasmid transferability alone. Given that plasmids often act as burden to their bacterial hosts, this could lead to their loss unless they offer a significant advantage to the cell[21]. Future research should delve deeper into the role of gene stability in affecting gene connectivity, possibly offering novel insights into the complexity hypothesis.

Overall, this chapter has expanded our understanding of how horizontal gene transfer influences bacterial cooperation and offered a fresh perspective at the complexity hypothesis.

It underscored the need for future research to examine characteristics of plasmids beyond their horizontal gene transfer potential.

5.3 Driving Factors Behind Bacterial Growth Rate Evolution

In Chapter 4, I expanded the scope of the thesis to explore beyond bacterial cooperation, focusing on how environmental factors and genomic features influence bacterial growth rates across various species. Existing hypotheses posit a direct positive relationship between environmental variability and growth rates and anticipate a trade-off between growth rates and genomic features such as genome size and metabolic versatility. Our research confirmed that species in variable environments generally exhibit faster growth rates (Chapter 4, Figure 2). Contrary to expectations, we observed a positive association between genomic features and growth rates (Chapter 4, Figure 3). Further phylogenetic path analysis provided insights into this discrepancy, suggesting that environmental variability acts as a third variable affecting both growth rates and genomic features, thereby countering the presumed trade-offs between them (Chapter 4, Figure 6).

This chapter highlights the value of utilizing phylogenetic path analysis to distinguish between direct and indirect effects, a task often challenging in traditional regression models[22]. While standard regressions in our study identified significant correlations between genomic features and growth rates, path analysis suggested these correlations were minimal or even absent when examined directly. This indicates that the perceived correlations from traditional models might arise from the indirect influence of environmental variability on both growth rates and genomic features. From a mathematical perspective, if environmental variability positively affects both growth rates and genomic characteristics in direct manners, an indirect positive correlation between the latter two is inevitable. But do genuine trade-offs exist between genomic features

and growth rates? To address this, we examined the relationship within species across various strains, accounting for environmental variability. Our findings revealed that, indeed, some species exhibit trade-offs between genomic characteristics and growth rates, while others do not, or even show the reverse trend (Chapter 4, Figure 7). This suggested that the occurrence of these trade-offs depends on the specific species, requiring a closer examination of their unique traits and life history strategies.

Path analysis offers a valuable method to distinguish between direct and indirect effects, enhancing our comprehension of causal relationships among variables. However, this technique has inherent limitations. Firstly, just because two variables are directly correlated in a path model doesn't necessarily mean that one directly causes the other. As a comparison, Granger Causality, mentioned in Chapter 2, establishes a temporal precedence between variables, adding another dimension to understanding causal relationships. Secondly, path analysis predominantly deals with observable variables, it struggles with unobservable or latent variables. These are variables that are hard to directly observe with accuracy but can be indirectly inferred through multiple observable ones[23–25]. For example, in our study, we used the number of habitats a species inhabits as a representation of environmental variability. But does this metric capture the entirety of environmental variability and its influence on bacterial growth rate evolution? Factors such as spatial heterogeneity, nutrient supply richness, and frequency of environmental fluctuations could also be significant components of environmental variability that influence bacterial growth rates[26–28]. To encapsulate a broad and abstract variable like 'environmental variability' from numerous quantifiable ones and study its causal relationships with other variables, it might be more suitable to employ other methods like structural equation modelling (SEM)[25]. Bayesian modelling could also be integrated into structural model to enhance causal inference, but I won't go into details here[29].

In summary, this chapter offers insights into the lack of anticipated trade-offs between growth rate and genomic characteristics, including genome size and metabolic capability. Despite our advancements, there's a need for more rigorous research to solidify these results, as mentioned in Chapter 4.

5.4 Future Directions

In this section, I outline two pivotal avenues for advancing our understanding of bacterial cooperation and the broader context of bacterial ecology and evolution: (i) Refining methods for inferring cooperative genes from genomic data; (ii) Identifying and investigating under-explored areas within bacterial biogeography.

5.4.1 Refining methods for inferring cooperative genes from genomic data

In this thesis, I employed two distinct methods to pinpoint cooperative genes using genomic information. Chapter 2 used functional annotation tools rooted in KEGG Orthology (KO)[30], whereas in Chapter 3, cooperative genes were identified as those encoding extracellular proteins, a classification achieved with the PSORTb tool[31]. A combined application of these identification strategies was also carried out in another research effort focused on validating the kin selection theory in gut microbiota[17].

The mentioned methods are not without pitfalls in inferring cooperative genes. The earlier section of this chapter outlined challenges with utilizing genes for extracellular proteins as cooperative gene proxies. Regarding the functional annotation approaches, there are also limitations. Firstly, this approach depends on a curated list of cooperative KO terms, derived from searching all KO terms for keywords related to known bacterial cooperative behaviours. Consider the siderophore production as an example. To identify genes involved, one would

first establish ‘siderophore KO terms’ based on text matches. Afterwards, any gene annotated with these terms would be classified as a siderophore production gene. However, the method’s efficacy heavily depends on understanding cooperative behaviours and accurately defining keywords. For instance, different species might use different names for their siderophores, like ‘enterobactin’ in *Escherichia coli*[32] and ‘bacillibactin’ in *Bacillus*[33]. Therefore, solely using ‘siderophore KO terms’ could lead to missing genes relevant to these compounds. Secondly, the accuracy of gene function annotation tools depends on the quality and quantity of reference data in databases. Tools perform optimally for well-researched organisms with ample related data. However, for less studied organisms with fewer data available, these tools may offer less reliable annotations[34,35].

We have devised a novel bioinformatics pipeline, SOCfinder, to identify cooperative genes in bacterial genomes (Belcher, L. J., et al. 2023). A key advantage of SOCfinder is its integration of various methods, facilitating a more comprehensive search for cooperative genes. It comprises three modules: Module 1 conducts BLAST searches against databases of genes with determined subcellular locations; Module 2 employs functional annotations via KEGG Orthology; Module 3 leverages antiSMASH[36] to pinpoint genes encoding synthesis pathways of cooperative secondary metabolites. While SOCfinder represents a marked advancement from prior approaches, its efficacy could potentially be heightened by incorporating techniques that other researchers utilize to discern genes of interest from genomic datasets. I will elucidate a few of these techniques.

A technique designed for the automatic detection of antiviral system genes offers a promising enhancement to existing methods[37]. This approach involves several primary steps:

(1) Enumerate all recognized defence systems. Similarly, we could enumerate all known cooperative behaviours.

(2) Establish Hidden Markov Model (HMM) profiles for every protein in each defence system to allow for homology searches[38]. Analogously, we could generate HMM profiles for proteins identified as cooperative from lab experiments. For instance, in *Pseudomonas aeruginosa*, the elastase *lasB* gene could be considered cooperative based on experimental findings[39–42].

(3) Each defence system may involve various interconnected proteins. Therefore, they designed decision rules using MacsyFinder syntax to aid in pinpointing defence systems by retaining only the HMM hits consistent with the genetic architecture of the targeted system. MacSyFinder is a tool dedicated to detecting macromolecular genetic architecture. The decision rules typically encompass a list of mandatory, accessory, or excluded proteins essential for identifying a specific defence system[43,44]. In the context of cooperation, which also involves multiple proteins, we can review the literatures to discern all proteins associated with a specific cooperative behaviour. Subsequently, we need to determine if each protein is mandatory, accessory, or excluded for the manifestation of this particular behaviour. We then use these decision rules to facilitate the search for proteins integral to a specific cooperative behaviour.

(4) Start with an initial compilation of HMM profiles and affiliated decision rules, assess the quality of the searches against existing datasets, and iteratively refine to optimize the HMM profiles and affiliated decision rules. In terms of cooperation, The SOCfinder database can be a valuable resource to enhance the performance of these HMM profiles and decision rules.

This method is now widely used, demonstrated by its successful application in detecting CRISPR-Cas systems[45], secretion systems[46], and bacterial capsules[47] in various studies. Consequently, it's endorsed within the scientific community as a reliable bioinformatics approach for identifying genes with specific functions, including genes for cooperation. Notably, the integration of machine learning (ML) might offer room for significant pipeline enhancement[48,49]. For instance, ML facilitates rigorous model validation, ensuring the HMM profiles and decision rules generalize well to unseen data, thereby mitigating overfitting. In addition, by training with the SOCfinder database, ML models can benefit from a vast knowledge of cooperation, enhancing their prediction and identification capabilities. Given sufficient data, ML can predict potential cooperative genes, extending beyond the established HMM profiles and decision rules, by using insights from previously identified cooperative genes.

5.4.2 *Identifying and investigating under-explored areas within bacterial biogeography*

In Chapter 1, I introduced a framework for examining eco-evo processes underpinning bacterial biogeography. According to this framework, bacterial diversity emerges from four main processes, speciation, selection, dispersal, and drift[50,51]. In Chapter 2, I studied bacterial cooperation, exploring how selection, driven by cooperation, shape bacterial biogeographic distributions. In this section, I will highlight some lesser-studied aspects of bacterial biogeography, with an emphasis on the role of drift in shaping these patterns.

Drift, inherently emphasizing stochasticity, can alter bacterial biogeography and species compositions, leading to changes in genotype frequencies[50]. Consider the distance-decay relationship as an example, which describes the decline in taxonomic similarity with increasing geographic distance[52,53]. Drift, resulting from stochastic variations in taxa births, deaths, and

migrations, can diversify microbial compositions spatially, thereby reinforcing this relationship[51]. However, discerning the impact of drift on global bacterial diversity distribution poses challenges. Drift's effects are most detectable at community or population scales, often identified by comparing observed microbial abundance to predictions from neutral community models (NCMs)[54,55] or through population genetic tools[56].

At community and population levels, the influence of genetic drift intertwined with biotic interaction-driven selection on bacterial biogeography is well-understood. First, genetic drift is associated with spatial segregation. In biofilms, some cell lineages often experience strong spatial bottlenecks when they get randomly isolated at the growing fronts, leading to population subdivision[13]. This has been observed where intensified genetic drift during range expansions significantly modifies the gene pool of organisms like *E. coli* and *S. cerevisiae*[57]. Second, genetic drift can stabilize cooperation during niche expansions. At expansion fronts, genetic drift creates regions of low genetic and species diversity due to repeated founder effects, promoting the use of cooperative enzymes by a single genotype. Experiments in expanding metapopulations of *S. cerevisiae* have demonstrated this, revealing a drift-induced spatial structure that promotes cooperation[58,59]. Third, while spatial structure from genetic drift can enhance cooperation, it can also hinder mutualistic relationships, as mutualists thrive in multi-species coexistence regions. Experiment has illustrated that genetic drift can counteract mutualism during niche expansions[60].

Based on the current finding, I propose that more research is essential to decipher the role of drift, especially its interaction with biotic interaction-driven selection such as cooperation, in bacterial biogeography across different scales, particularly at macroecological or global scales. Estimating the influence of drift on global bacterial distribution remains a daunting task due to

the immense diversity and intricacy of microbial habitats, compounded by the interaction of numerous biogeographical factors over varying timescales. Thus, innovative methodologies, especially how to make use of ample genomic or metagenomic data, are imperative to address these challenges.

Lastly, other facets of bacterial biogeography, including the roles of speciation and diversification[61], as well as the relationship between bacterial biogeography and disease[62,63], also represent promising future research avenues. In the future, I will explore more on these aspects.

References

1. Flemming, H.-C. & Wingender, J. The biofilm matrix. *Nat Rev Microbiol* **8**, 623–633 (2010).
2. Costerton, J. W., Stewart, P. S. & Greenberg, E. P. Bacterial Biofilms: A Common Cause of Persistent Infections. *Science* **284**, 1318–1322 (1999).
3. Stoodley, P., Sauer, K., Davies, D. G. & Costerton, J. W. Biofilms as Complex Differentiated Communities. *Annual Review of Microbiology* **56**, 187–209 (2002).
4. Flemming, H.-C. *et al.* Biofilms: an emergent form of bacterial life. *Nat Rev Microbiol* **14**, 563–575 (2016).
5. Koo, H., Allan, R. N., Howlin, R. P., Stoodley, P. & Hall-Stoodley, L. Targeting microbial biofilms: current and prospective therapeutic strategies. *Nat Rev Microbiol* **15**, 740–755 (2017).
6. Rumbaugh, K. P. & Sauer, K. Biofilm dispersion. *Nat Rev Microbiol* **18**, 571–586 (2020).
7. Sauer, K. *et al.* Characterization of Nutrient-Induced Dispersion in Pseudomonas aeruginosa PAO1 Biofilm. *Journal of Bacteriology* **186**, 7312–7326 (2004).

8. Sauer, K. *et al.* The biofilm life cycle: expanding the conceptual model of biofilm formation. *Nat Rev Microbiol* **20**, 608–620 (2022).
9. Flemming, H.-C. & Wuertz, S. Bacteria and archaea on Earth and their abundance in biofilms. *Nat Rev Microbiol* **17**, 247–260 (2019).
10. Bar-On, Y. M. & Milo, R. Towards a quantitative view of the global ubiquity of biofilms. *Nat Rev Microbiol* **17**, 199–200 (2019).
11. Zhang, J. & Yang, J.-R. Determinants of the rate of protein sequence evolution. *Nat Rev Genet* **16**, 409–420 (2015).
12. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101–113 (2004).
13. Nadell, C. D., Drescher, K. & Foster, K. R. Spatial structure, cooperation and competition in biofilms. *Nat Rev Microbiol* **14**, 589–600 (2016).
14. 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).
15. Sargent, R. D. & Otto, S. P. The Role of Local Species Abundance in the Evolution of Pollinator Attraction in Flowering Plants. *The American Naturalist* **167**, 67–80 (2006).
16. Vandecasteele, S. J., Peetermans, W. E., Merckx, R. & Van Eldere, J. Quantification of Expression of Staphylococcus epidermidis Housekeeping Genes with Taqman Quantitative PCR during In Vitro Growth and under Different Conditions. *Journal of Bacteriology* **183**, 7094–7101 (2001).
17. Simonet, C. & McNally, L. Kin selection explains the evolution of cooperation in the gut microbiota. *Proceedings of the National Academy of Sciences* **118**, e2016046118 (2021).
18. Humphrey, S. *et al.* Bacterial chromosomal mobility via lateral transduction exceeds that of classical mobile genetic elements. *Nat Commun* **12**, 6509 (2021).

19. 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).
20. Hall, J. P. J. Is the bacterial chromosome a mobile genetic element? *Nat Commun* **12**, 6400 (2021).
21. 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).
22. Bijl, W. van der. phylopath: Easy phylogenetic path analysis in R. *PeerJ* **6**, e4718 (2018).
23. Grace, J. B., Anderson, T. M., Olff, H. & Scheiner, S. M. On the specification of structural equation models for ecological systems. *Ecological Monographs* **80**, 67–87 (2010).
24. Eisenhauer, N., Bowker, M. A., Grace, J. B. & Powell, J. R. From patterns to causal understanding: Structural equation modeling (SEM) in soil ecology. *Pedobiologia* **58**, 65–72 (2015).
25. Fan, Y. *et al.* Applications of structural equation modeling (SEM) in ecological studies: an updated review. *Ecol Process* **5**, 19 (2016).
26. Pickett, S. T. A. & Cadenasso, M. L. Landscape Ecology: Spatial Heterogeneity in Ecological Systems. *Science* **269**, 331–334 (1995).
27. Acar, M., Mettetal, J. T. & van Oudenaarden, A. Stochastic switching as a survival strategy in fluctuating environments. *Nat Genet* **40**, 471–475 (2008).
28. Kussell, E. & Leibler, S. Phenotypic Diversity, Population Growth, and Information in Fluctuating Environments. *Science* **309**, 2075–2078 (2005).
29. Lee, S.-Y. *Structural Equation Modeling: A Bayesian Approach*. (John Wiley & Sons, 2007).

30. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457–D462 (2016).
31. 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).
32. Raymond, K. N., Dertz, E. A. & Kim, S. S. Enterobactin: An archetype for microbial iron transport. *Proceedings of the National Academy of Sciences* **100**, 3584–3588 (2003).
33. Dertz, E. A., Xu, J., Stintzi, A. & Raymond, K. N. Bacillibactin-Mediated Iron Transport in Bacillus subtilis. *J. Am. Chem. Soc.* **128**, 22–23 (2006).
34. Jones, C. E., Brown, A. L. & Baumann, U. Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics* **8**, 170 (2007).
35. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLOS Computational Biology* **5**, e1000605 (2009).
36. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **49**, W29–W35 (2021).
37. Tesson, F. *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun* **13**, 2561 (2022).
38. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**, W29–W37 (2011).
39. Chen, R., Déziel, E., Groleau, M.-C., Schaefer, A. L. & Greenberg, E. P. Social cheating in a Pseudomonas aeruginosa quorum-sensing variant. *Proceedings of the National Academy of Sciences* **116**, 7021–7026 (2019).

40. Sandoz, K. M., Mitzimberg, S. M. & Schuster, M. Social cheating in Pseudomonas aeruginosa quorum sensing. *Proceedings of the National Academy of Sciences* **104**, 15876–15881 (2007).
41. 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).
42. Özkaya, Ö., Balbontín, R., Gordo, I. & Xavier, K. B. Cheating on Cheaters Stabilizes Cooperation in Pseudomonas aeruginosa. *Current Biology* **28**, 2070-2080.e6 (2018).
43. Abby, S. S., Néron, B., Ménager, H., Touchon, M. & Rocha, E. P. C. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLOS ONE* **9**, e110726 (2014).
44. Néron, B. *et al.* MacSyFinder v2: Improved modelling and search engine to identify molecular systems in genomes. *Peer Community Journal* **3**, (2023).
45. Couvin, D. *et al.* CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research* **46**, W246–W251 (2018).
46. Abby, S. S., Denise, R. & Rocha, E. P. Identification of protein secretion systems in bacterial genomes using MacSyFinder version 2. 2023.01.06.522999 Preprint at https://doi.org/10.1101/2023.01.06.522999 (2023).
47. Rendueles, O., Garcia-Garcerà, M., Néron, B., Touchon, M. & Rocha, E. P. C. Abundance and co-occurrence of extracellular capsules increase environmental breadth: Implications for the emergence of pathogens. *PLOS Pathogens* **13**, e1006525 (2017).
48. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat Rev Genet* **16**, 321–332 (2015).
49. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol* **23**, 40–55 (2022).

50. Vellend, M. Conceptual Synthesis in Community Ecology. *The Quarterly Review of Biology* **85**, 183–206 (2010).

51. Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C. & Martiny, J. B. H. Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Microbiol* **10**, 497–506 (2012).

52. Green, J. L. *et al.* Spatial scaling of microbial eukaryote diversity. *Nature* **432**, 747–750 (2004).

53. Horner-Devine, M. C., Lage, M., Hughes, J. B. & Bohannan, B. J. M. A taxa–area relationship for bacteria. *Nature* **432**, 750–753 (2004).

54. Hubbell, S. P. The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32). in *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)* (Princeton University Press, 2011). doi:10.1515/9781400837526.

55. Sloan, W. T. *et al.* Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environmental Microbiology* **8**, 732–740 (2006).

56. Whitaker, R. J., Grogan, D. W. & Taylor, J. W. Geographic Barriers Isolate Endemic Populations of Hyperthermophilic Archaea. *Science* **301**, 976–978 (2003).

57. Hallatschek, O., Hersen, P., Ramanathan, S. & Nelson, D. R. Genetic drift at expanding frontiers promotes gene segregation. *Proceedings of the National Academy of Sciences* **104**, 19926–19930 (2007).

58. Datta, M. S., Korolev, K. S., Cvijovic, I., Dudley, C. & Gore, J. Range expansion promotes cooperation in an experimental microbial metapopulation. *Proceedings of the National Academy of Sciences* **110**, 7354–7359 (2013).

59. Van Dyken, J. D., Müller, M. J. I., Mack, K. M. L. & Desai, M. M. Spatial Population Expansion Promotes the Evolution of Cooperation in an Experimental Prisoner's Dilemma. *Current Biology* **23**, 919–923 (2013).

60. Müller, M. J. I., Neugeboren, B. I., Nelson, D. R. & Murray, A. W. Genetic drift opposes mutualism during spatial population expansion. *Proceedings of the National Academy of Sciences* **111**, 1037–1042 (2014).
61. Sriswasdi, S., Yang, C. & Iwasaki, W. Generalist species drive microbial dispersion and evolution. *Nat Commun* **8**, 1162 (2017).
62. Stacy, A., McNally, L., Darch, S. E., Brown, S. P. & Whiteley, M. The biogeography of polymicrobial infection. *Nat Rev Microbiol* **14**, 93–105 (2016).
63. Azimi, S., Lewin, G. R. & Whiteley, M. The biogeography of infection revisited. *Nat Rev Microbiol* **20**, 579–592 (2022).

Supplementary materials

In this section, I provide supplementary figures and tables for each chapter. Supplementary materials for Chapter 2 and Chapter 4 can be found below. Supplementary materials for Chapter 3 can be downloaded from https://rs.figshare.com/collections/Supplementary_material_from_Gene_transferability_and_sociality_do_not_correlate_with_gene_connectivity_/6302904

Chapter 2 Supplementary Information

Supplementary figures

Figure S1. Distribution of the proportion of cooperative genes across 25,785 species.

Figure S2. Ancestral state reconstruction of niche breadths for 24,912 ancestral species.

Figure S3. Causal inference: whether higher cooperative gene proportions influencing generalization across different cooperative gene types.

Figure S4. Causal inference: whether lower cooperative gene proportions influencing specialization across different cooperative gene types.

Figure S5. Causal inference: whether generalization/specialization influencing the increase/decrease of cooperative genes across various cooperative gene types.

Figure S6. Phylogenetic tree of 171 species used in estimating short-term gene gains/losses.

Figure S7. Distribution of the proportion of genes co-evolved with cooperative genes across 25,785 species.

Figure S8. Heatmap of Silhouette index for the result of hierarchical clustering using method as “ward.D2”.

Figure S9. The distribution of the number of cooperative KOs with which each non-cooperative KO co-occurred.

Supplementary tables

Table S1. MCMCglmm analyses results.

Table S2. Information on 114 ProkAtlas habitats and their corresponding habitat clusters.

Table S3. PGLS-ANOVA analyses results.

Table S4. Causal inference: whether higher cooperative gene proportions influencing generalization across different cooperative gene types.

Table S5. Causal inference: whether lower cooperative gene proportions influencing specialization across different cooperative gene types.

Table S6. Causal inference: whether generalization/specialization influencing the increase/decrease of cooperative genes across various cooperative gene types.

Table S7. List of cooperative KOs.

Table S8. List of KOs that co-evolved with cooperative KOs.

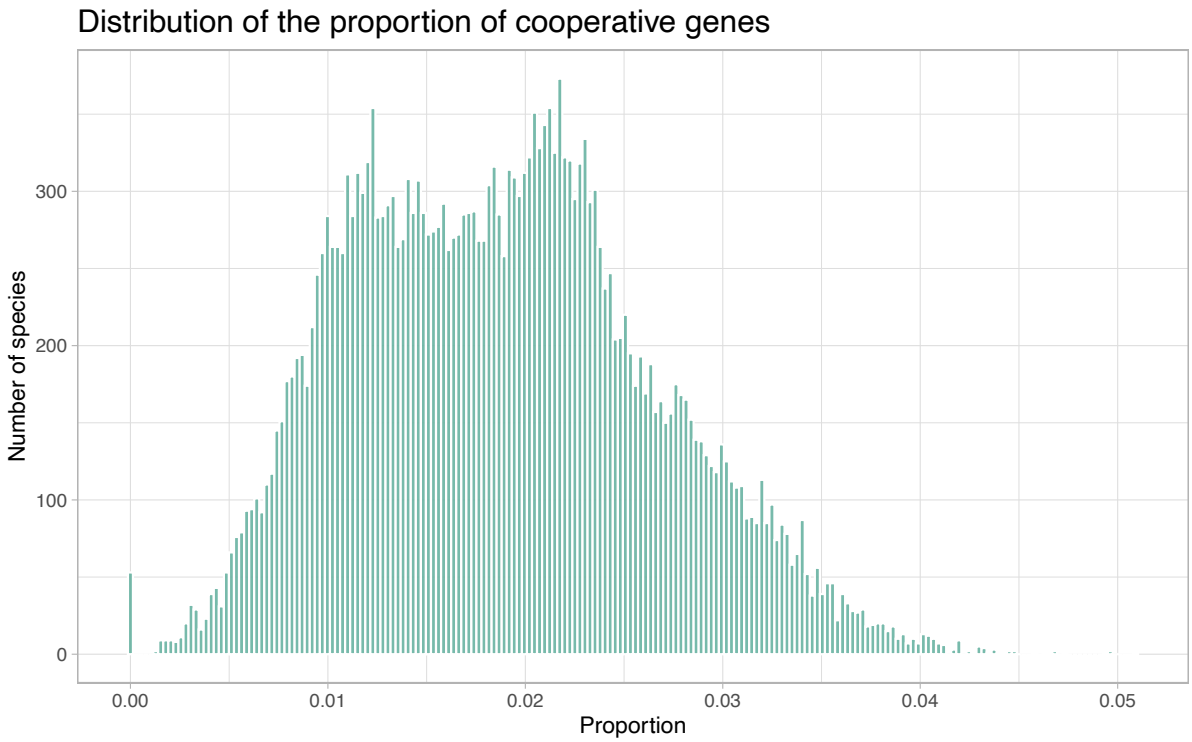


Figure S1. Distribution of the proportion of cooperative genes across 25,785 species. Proportions of cooperative genes in each species' representative genome indicate the extent of cooperative gene carriage for each species, ranging from 0 to 0.05. Notably, some species lack cooperative genes.

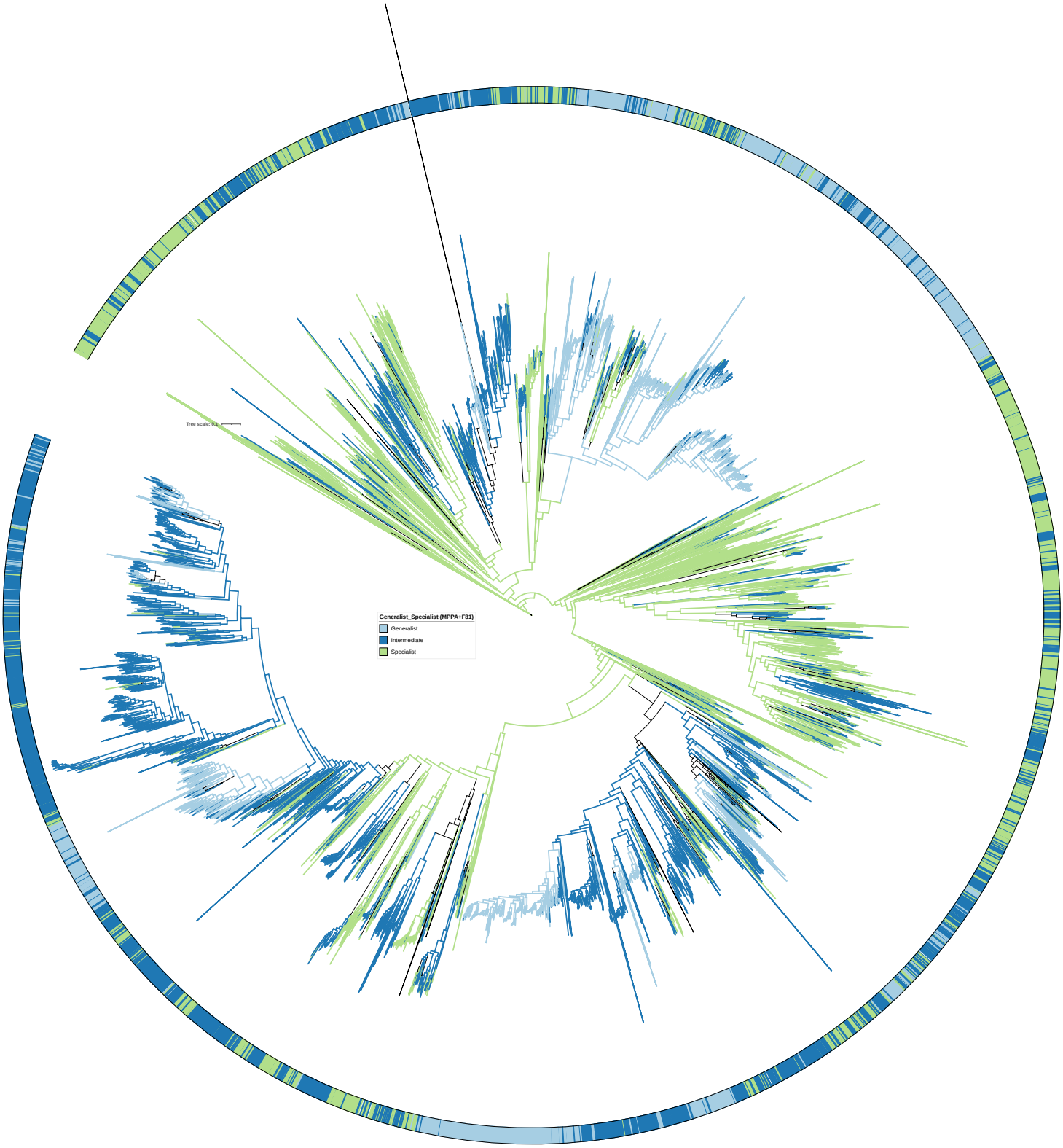


Figure S2. Ancestral state reconstruction of niche breadths for 24,912 ancestral species. The outer circle displays the niche breadths of extant (tip) species, while the inner tree illustrates those of ancestral (node) species. Green denotes specialists, dark blue denotes intermediate species, and light blue represents generalists.

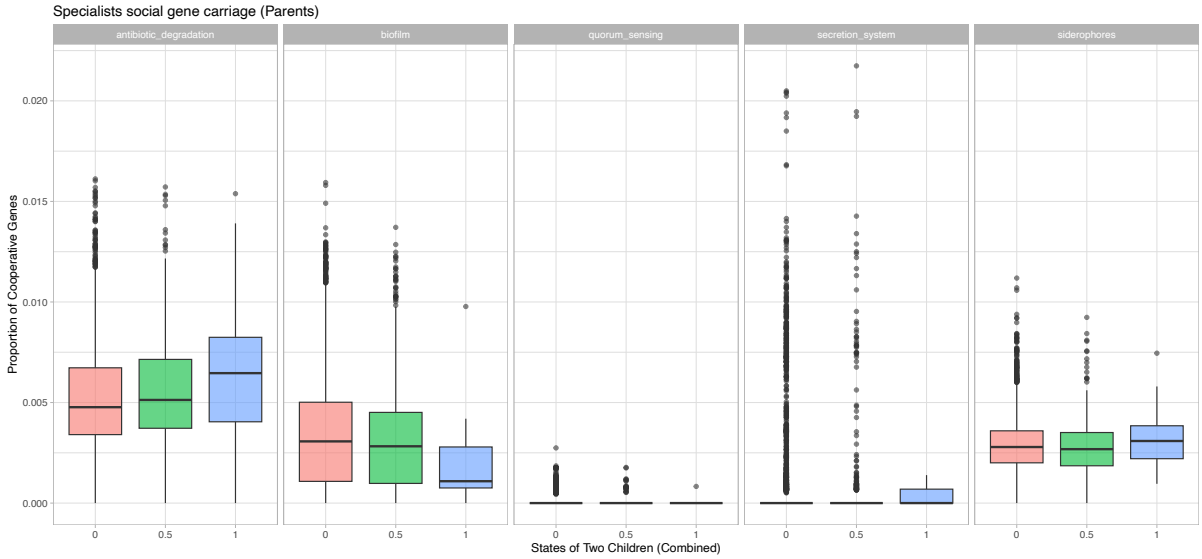


Figure S3. Causal inference: whether higher cooperative gene proportions influencing generalization across different cooperative gene types. Carrying higher proportion of genes for antibiotic degradation was found to promote generalization, whereas a higher proportion of biofilm formation genes actually hindered, not facilitated generalization (Table S4). All other types of genes did not display any causation in this direction.

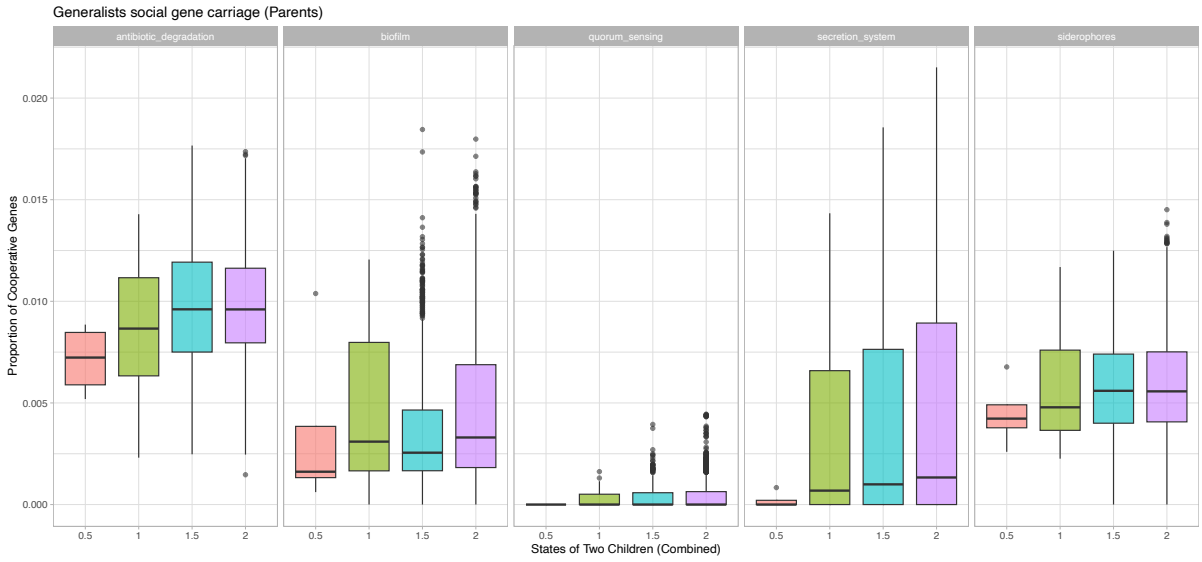


Figure S4. Causal inference: whether lower cooperative gene proportions influencing specialization across different cooperative gene types. We found carrying lower proportions of antibiotic degradation genes, biofilm formation genes, and secretion system genes facilitates specialization (Table S5).

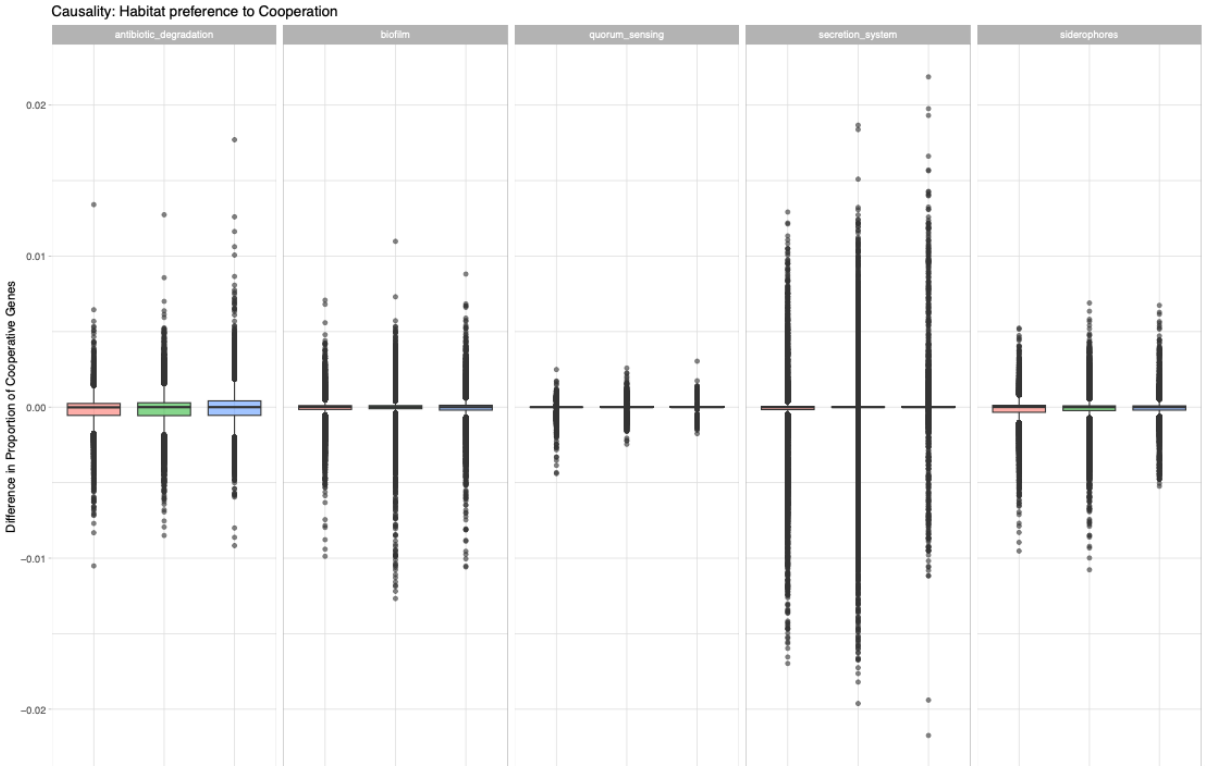


Figure S5. Causal inference: whether generalization/specialization influencing the increase/decrease of cooperative genes across various cooperative gene types. We observed that generalization/specialization influenced the rise or decline of all types of cooperative genes, as determined by the Kruskal-Wallis test (Table S6). However, potential sample size inflation (n = 24590 ancestral species) could result in these significant results. To address this, we assessed the effect sizes using "eta-squared." Eta-squared values can be interpreted as: 0.01 - 0.06 indicating a small effect, 0.06 - 0.14 denoting a moderate effect, and values greater than 0.14 suggesting a large effect. Effect sizes for all types of cooperative genes were small (Table S6).

Figure S6. Phylogenetic tree of 171 species used in estimating short-term gene gains/losses, based on tree from the GTDB. Species were represented with the NCBI accession numbers of their representative genomes. Accession numbers starting with "RS_" represent species from RefSeq database, and "GB_" represent species from GenBank database.

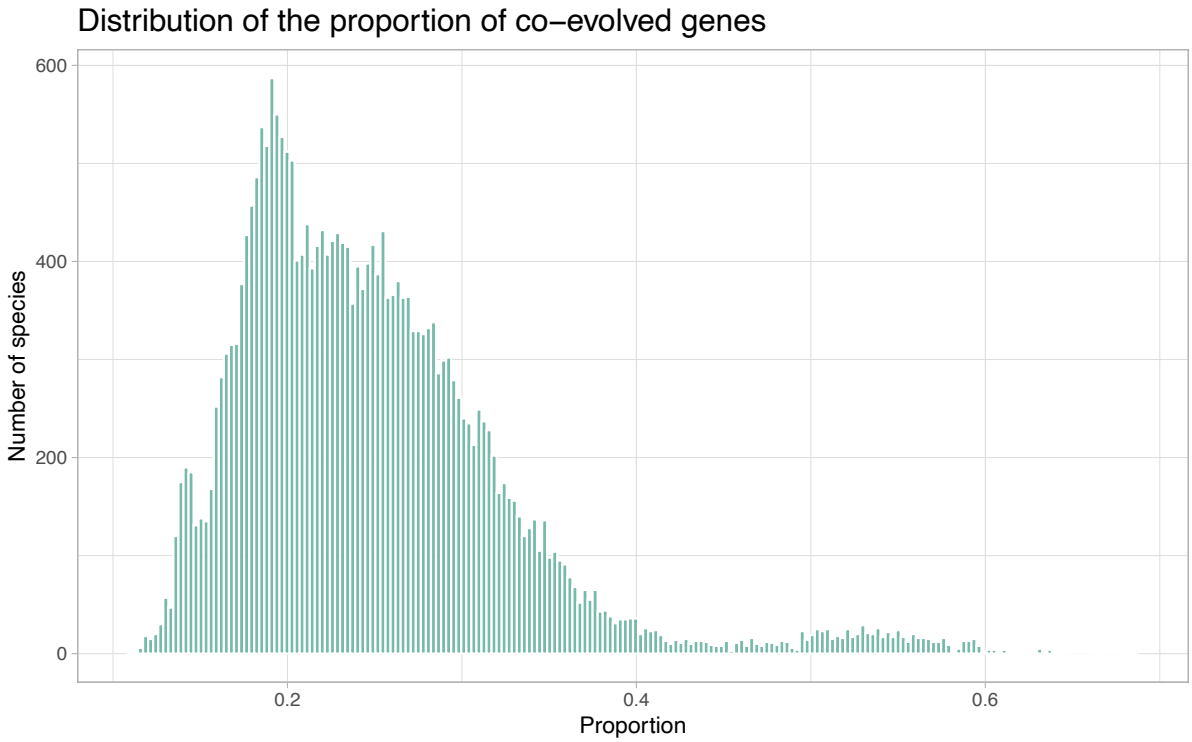


Figure S7. Distribution of the proportion of genes co-evolved with cooperative genes across 25,785 species. Proportions of co-evolved genes in each species' representative genome indicate the extent of co-evolved gene carriage for each species, ranging from 0.11 to 0.71.

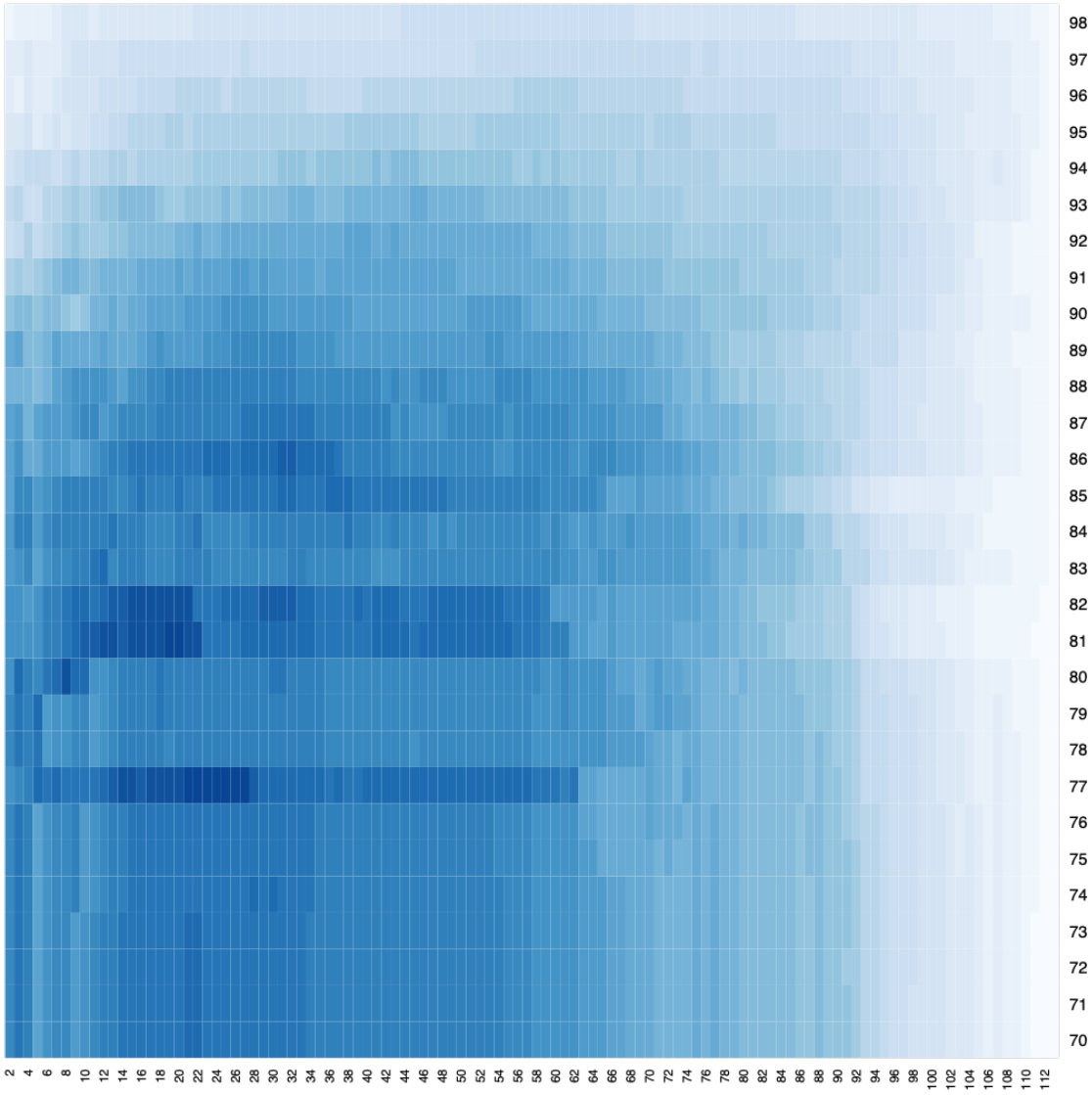


Figure S8. Heatmap of Silhouette index. This heatmap depicts the Silhouette index results for hierarchical clustering via the "ward.D2" method, a metric measuring cluster quality where higher values indicate better clustering. The X-axis represents the number of clusters, varying from 2 to 113, while the Y-axis reflects the sequence identity threshold, spanning 70% to 98%. Colour intensity corresponds to the Silhouette index value, with darker shades signifying higher values. Optimal clustering was observed with a 77% threshold and 26 habitat clusters using the "ward.D2" method.

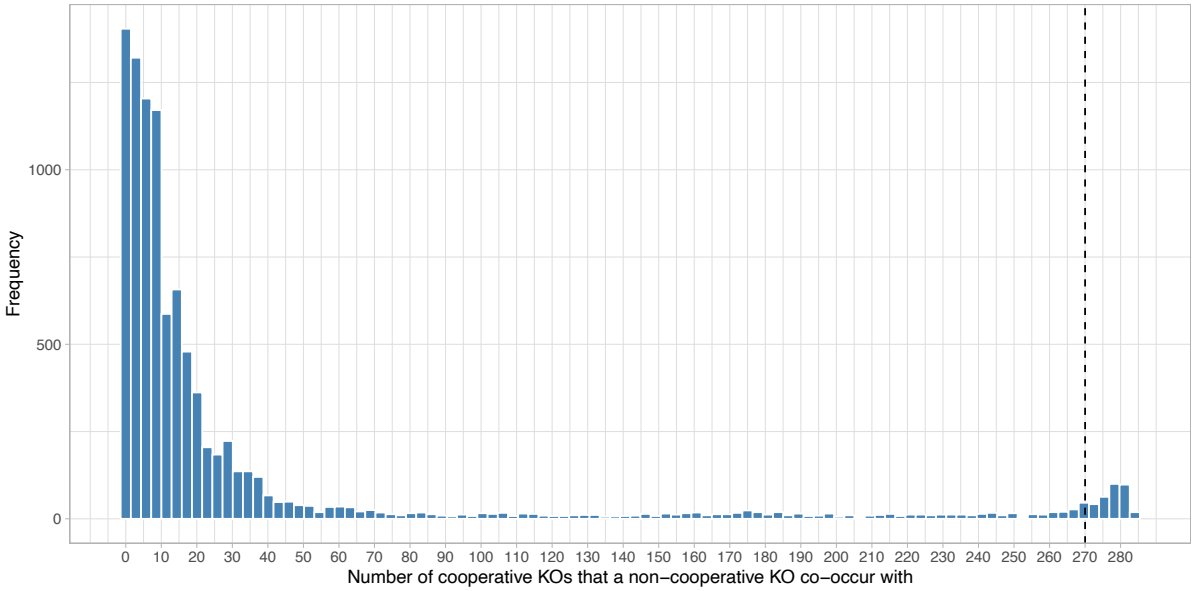


Figure S9. The distribution of the number of cooperative KOs with which each non-cooperative KO co-occurred. We identified an enrichment threshold at 270 cooperative KOs, suggesting that number of non-cooperative KOs become increased when these KOs can co-occur with at least 270 cooperative KOs. Consequently, we defined KOs as having co-evolved with cooperative KOs if they were significantly co-occurred with at least 270 cooperative KOs.

Table S1. MCMCglmm analyses results.

Model description	Term	Posterior mean	l-95% CI	u-95% CI	pMCMC	Sig.	Sample size
Causality: Whether having a higher proportion of cooperative genes facilitated generalization	0.5 vs. 0	-1.414e-04	-3.446e-04	5.262e-05	0.153	ns	3,832 species
	1 vs. 0	-6.329e-04	-1.597e-03	4.307e-04	0.229	ns	3,832 species
Causality: Whether having a lower proportion of cooperative genes facilitated specialization	1 vs. 0.5	0.003676	0.001304	0.005939	0.002979	**	9,173 species
	1.5 vs. 1	1.005e-04	-4.511e-04	6.539e-04	0.71106	ns	9,173 species
	2 vs. 1.5	0.0004175	0.0002779	0.0005659	< 2e-04	***	9,173 species
Causality: Whether generalization promoted an increase in cooperative genes or whether specialization facilitated a decrease in such genes	Generalists Vs. Intermediate	6.062e-05	-8.026e-05	1.833e-04	0.361	ns	24,590 species
	Generalists Vs. Specialists	9.183e-05	-1.219e-04	2.894e-04	0.374	ns	24,590 species
Extant species: Proportion of co-evolved genes ~ Niche Breadth	Intermediate Vs. Generalists	0.009965	0.008756	0.011171	< 2e-04	***	25,785 species
	Specialists Vs. Generalists	0.020331	0.018534	0.021894	< 2e-04	***	25,785 species
Extant species: Environmental variability ~ Proportion of co-evolved genes	Proportion of co-evolved genes	-10.750	-11.447	-10.021	< 2e-04	***	25,785 species
Extant species: Environmental variability ~ Proportion of co-evolved genes * Proportion of cooperative genes (all proportional data were scaled)	Proportion of co-evolved genes	-0.81012	-0.87308	-0.74596	< 2e-04	***	25,785 species
	Proportion of	0.11777	0.08138	0.15293	< 2e-04	***	25,785 species

	cooperative genes						
	Interaction	-0.06331	-0.09540	-0.03023	< 2e-04	***	25,785 species
Causality: Whether having a higher proportion of co-evolved genes facilitated generalization	Proportion of co-evolved genes	-0.1441	-3.3534	2.8905	0.929	ns	3,832 species
Causality: Children niche breadth ~ Ancestral (specialists) co-evolved gene proportions * Ancestral cooperative gene proportions (all proportional data were scaled)	Proportion of co-evolved genes	-0.04539	-0.28065	0.17307	0.701	ns	3,832 species
	Proportion of cooperative genes	-0.06144	-0.21058	0.08971	0.396	ns	3,832 species
	Interaction	0.02290	-0.12429	0.16318	0.743	ns	3,832 species
Causality: Whether having a lower proportion of co-evolved genes facilitated specialization	Proportion of co-evolved genes	-10.530	-12.577	-8.549	< 2e-04	***	9,173 species
Causality: Children niche breadth ~ Ancestral (generalists) co-evolved gene proportions * Ancestral cooperative gene proportions (all proportional data were scaled)	Proportion of co-evolved genes	-0.55691	-0.70036	-0.42170	< 2e-04	***	9,173 species
	Proportion of cooperative genes	0.09859	0.01596	0.17229	0.0123	*	9,173 species
	Interaction	-0.11714	-0.17206	-0.05487	< 2e-04	***	9,173 species
Causality: Whether generalization promoted an increase in co-evolved genes or whether specialization facilitated a decrease in such genes	Proportion of co-evolved genes	-1.8970	-5.3206	0.2803	0.154	ns	24,590 species

Causality: Ancestral niche breadth ~ Children co-evolved gene proportions * Children cooperative gene proportions (all proportional data were scaled)	Proportion of co-evolved genes	-0.017888	-0.034895	0.003727	0.152	ns	24,590 species
	Proportion of cooperative genes	0.005164	-0.012220	0.022488	0.621	ns	24,590 species
	Interaction	0.006508	-0.003640	0.013975	0.151	ns	24,590 species

Table S2. Information on 114 ProkAtlas habitats and their corresponding habitat clusters. Since the size of the table is too large, I uploaded the table into GitHub. https://github.com/haochh/Cooperation_Niche_Breadth/blob/main/habitat_to_cluster.csv

Table S3. PGLS-ANOVA analyses results.

Model description	Term	F-statistic	P-value	Sig	Sample size
Extant species: proportion of cooperative genes ~ Niche breadth	Cooperative genes	78.731	< 0.001	***	25,785 species
Extant species: proportion of different types of cooperative genes ~ Niche breadth	Antibiotic degradation	126.198	< 0.001	***	25,785 species
	Biofilm formation	7.526	< 0.001	***	25,785 species
	Quorum sensing	4.291	0.014	*	25,785 species
	Siderophores	21.657	< 0.001	***	25,785 species
	Secretion system	8.573	< 0.001	***	25,785 species
Extant species: proportion of co-evolved genes ~ Niche breadth	Co-evolved genes	276.562	< 0.001	***	25,785 species

Table S4. Causal inference: whether higher cooperative gene proportions influencing generalization across different cooperative gene types. We used Kruskal-Wallis test to compare cooperative gene proportions between different groups.

Social behaviour	Species number	Statistic	df	p-value	Sig
Antibiotic degradation	3832	10.1	2	0.00648	**
Biofilm	3832	8.58	2	0.0137	*
Quorum sensing	3832	0.718	2	0.698	ns
Secretion system	3832	1.41	2	0.495	ns
Siderophores	3832	3.33	2	0.189	ns

Table S5. Causal inference: whether lower cooperative gene proportions influencing specialization across different cooperative gene types. We used Kruskal-Wallis test to compare cooperative gene proportions between different groups.

Social behaviour	Species number	Statistic	df	p-value	Sig
Antibiotic degradation	9173	13.4	3	0.00379	**
Biofilm	9173	37.5	3	< 0.001	***
Quorum sensing	9173	9.06	3	0.0285	ns
Secretion system	9173	32.9	3	< 0.001	***
Siderophores	9173	4.68	3	0.197	ns

Table S6. Causal inference: whether generalization/specialization influencing the increase/decrease of cooperative genes across various cooperative gene types. We used Kruskal-Wallis test to compare cooperative gene proportions between different groups. Additionally, we used "eta-squared" to assess the effect sizes. Eta-squared values can be interpreted as: 0.01 - 0.06 indicating a small effect, 0.06 - 0.14 denoting a moderate effect, and values greater than 0.14 suggesting a large effect.

Social behaviour	Species number	Statistic	df	p-value	Sig	Eta^2
Antibiotic degradation	24590	56.9	3	< 0.001	***	0.00112
Biofilm	24590	41.0	3	< 0.001	***	0.00079
Quorum sensing	24590	227.0	3	< 0.001	***	0.00458
Secretion system	24590	12.6	3	< 0.001	***	0.00022
Siderophores	24590	44.9	3	< 0.001	***	0.00087

Table S7. List of cooperative KOs. I uploaded the table into GitHub. This table contains information of cooperative KOs, the keywords I used to retrieve these KOs, and their corresponding cooperative behaviours.
https://github.com/haochh/Cooperation_Niche_Breadth/blob/main/Cooperative_KOs.csv

Table S8. List of KOs that co-evolved with cooperative KOs. I uploaded the table into GitHub. This table contains information of KOs that co-evolved with cooperative KOs, and their corresponding KEGG pathway annotations.
https://github.com/haochh/Cooperation_Niche_Breadth/blob/main/co_evolved_ko_path_info.csv

Chapter 4 Supplementary Information

Supplementary figures

Figure S1. Classification of habitat clusters.

Figure S2. Relationship between genome fluidity and genome size.

Figure S3. Relationship between host dependence and genomic characteristics.

Figure S4. Path analysis displaying relationship between three genomic characteristics.

Figure S5-7. Results of intermediate steps of path analysis in main text.

Figure S8. Applying Central Limit Theorem (CLT) to estimate genome fluidity of *Escherichia coli.*

Figure S9. Distribution of environmental variability for 171 species.

Figure S10. Phylogenetic tree of 171 species in our dataset.

Supplementary tables

Table S1. MCMCglmm analyses results.

Cluster Dendrogram

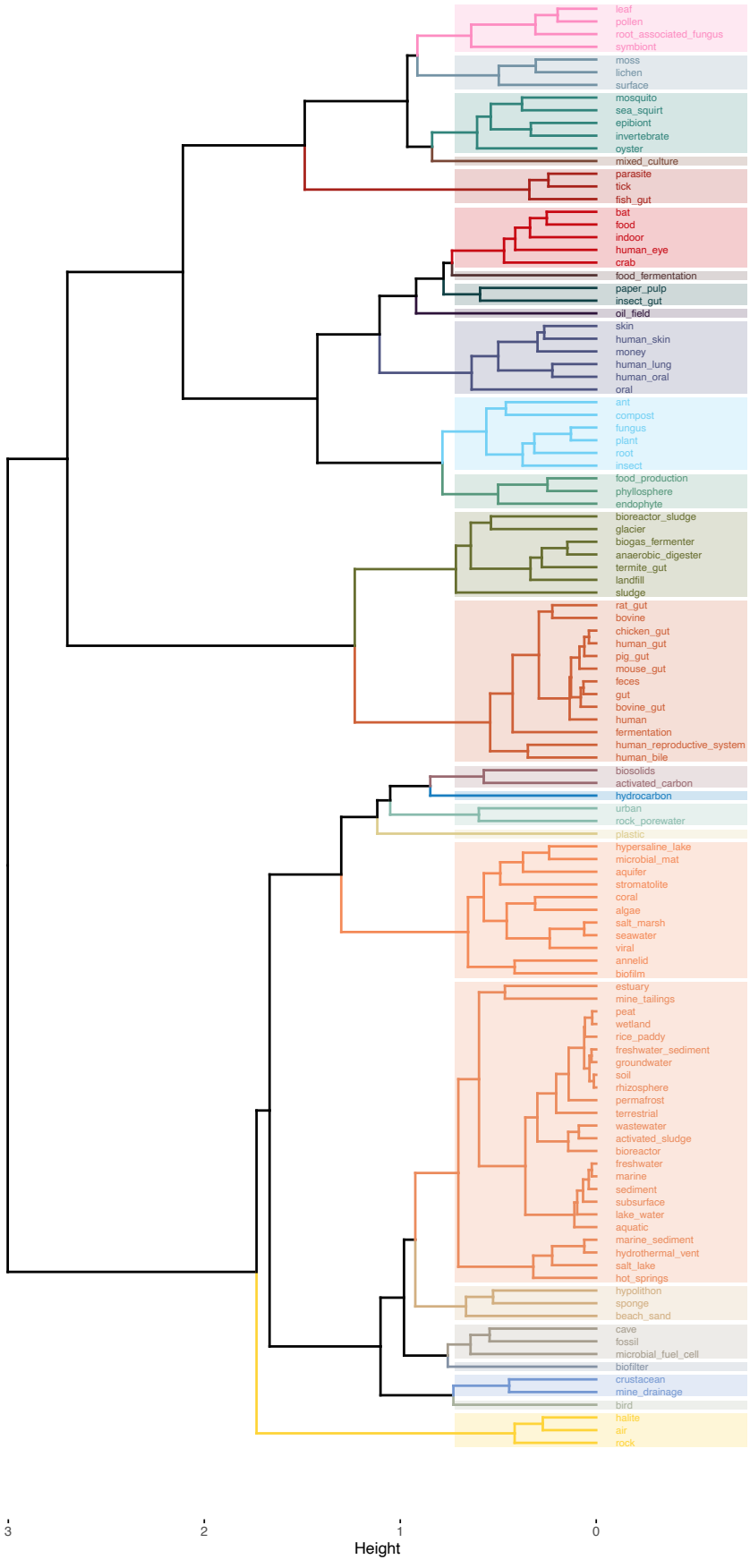


Figure S1. Classification of habitat clusters. We directly utilized pre-established habitat clusters to categorize the habitat preferences of each species in our dataset (see in Chapter 3). The 114 original habitats from the ProkAtlas database were clustered into 26 habitat clusters.

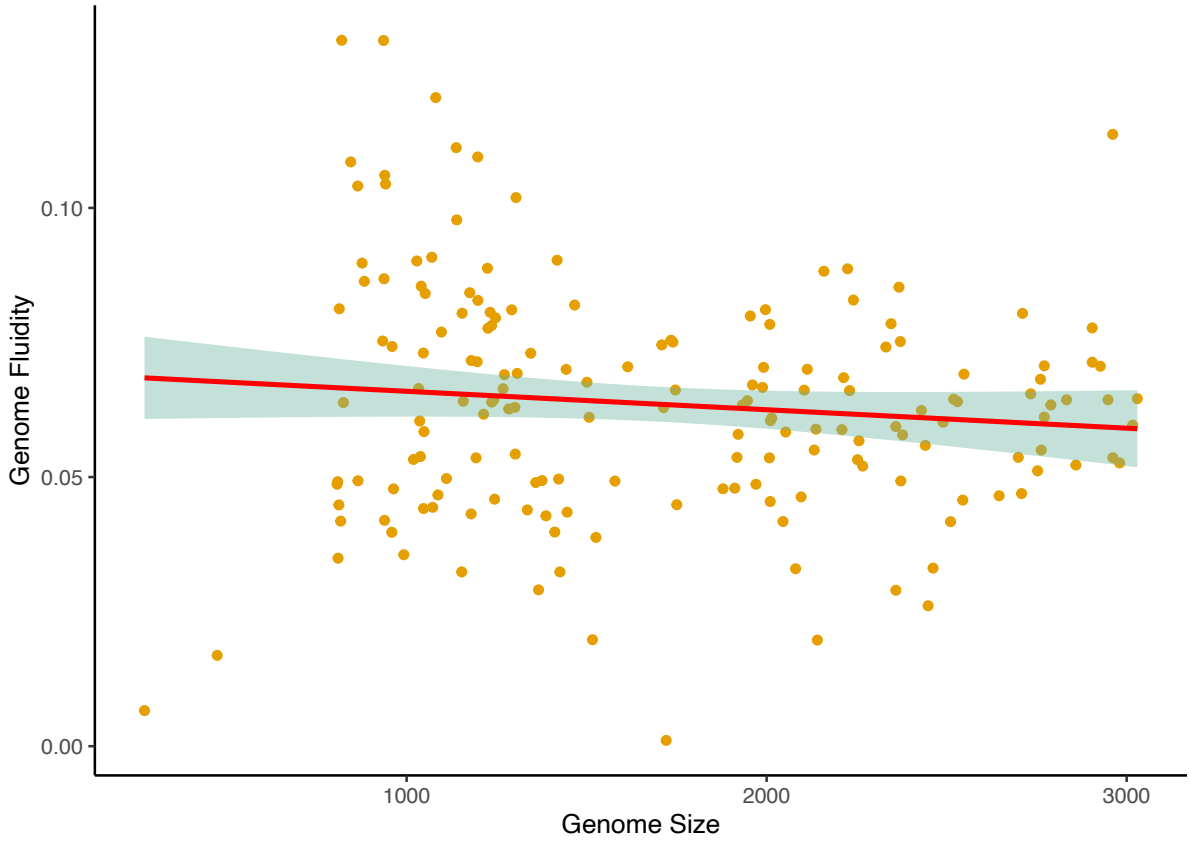


Figure S2. Relationship between genome fluidity and genome size. No significant correlation between genome fluidity and genome size was detected across 171 species in our dataset.

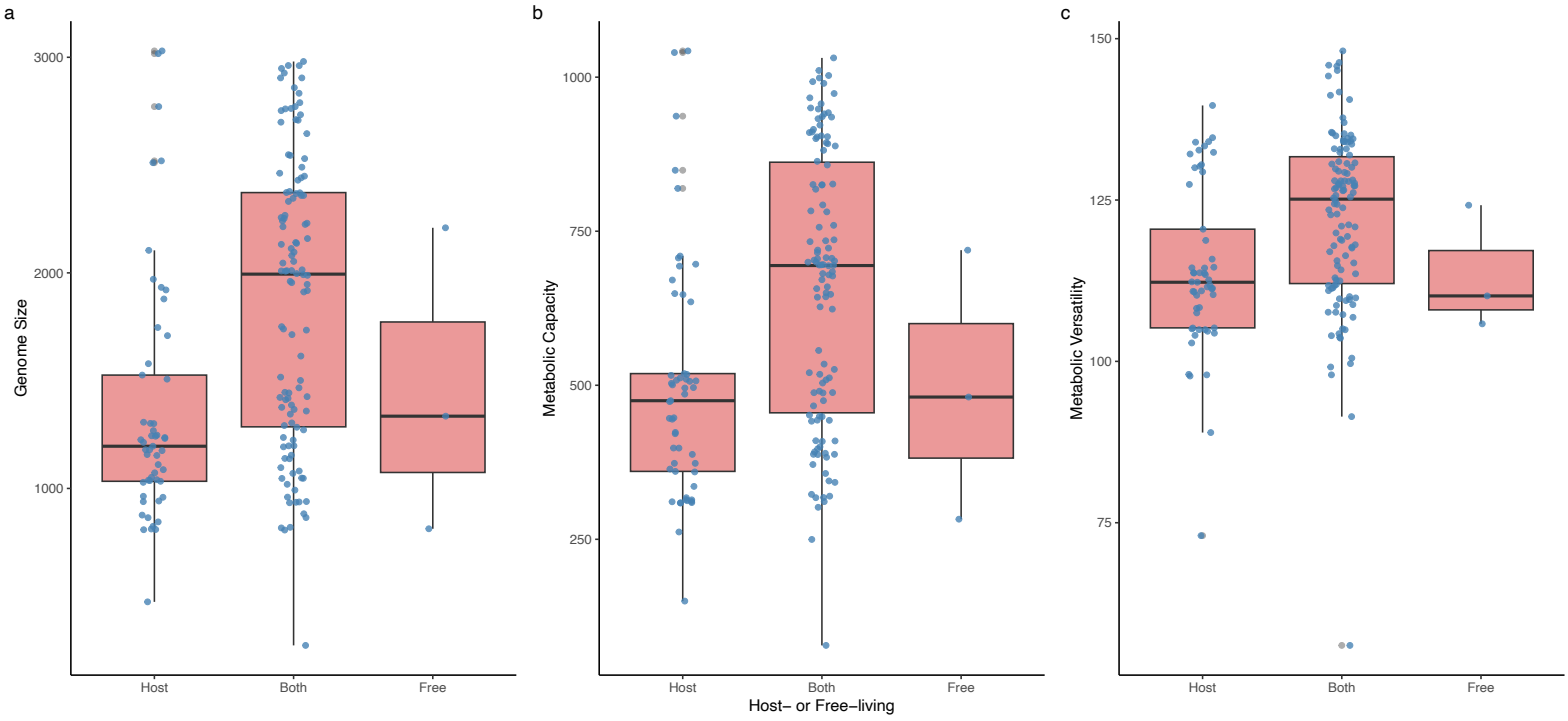


Figure S3. Relationship between host dependence and genomic characteristics. (a) Genome size vs. host dependence. (b) Metabolic capacity vs. host dependence. (c) Metabolic versatility vs. host dependence. Only genome size was found to be negatively associated with host dependence across 170 species. Host-dependent species had smaller genomes compared to species categorized as “both”. Free-living species (n = 3) were too few in our dataset.

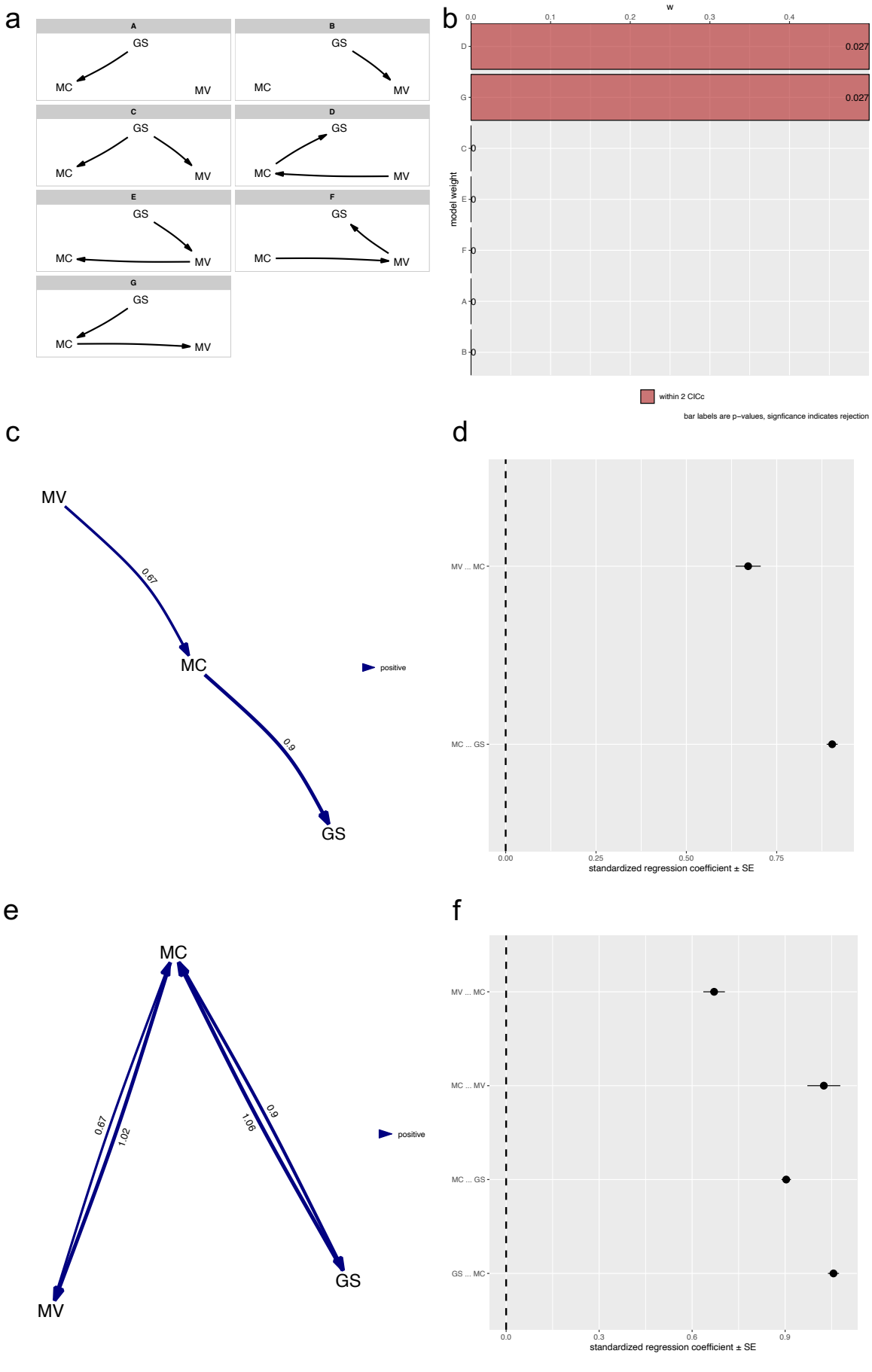

a
A
GS
MC
MV
B
GS
MC
MV
C
GS
MC
MV
D
GS
MC
MV
E
GS
MC
MV
F
GS
MC
MV
G
GS
MC
MV
b
w
0.0
0.1
0.2
0.3
0.4
D
0.027
G
0.027
C
0
E
0
F
0
A
0
B
0
model weight
within 2 CICc
bar labels are p−values, signficance indicates rejection
c
MV
0.67
MC
0.9
GS
positive
d
MV ... MC
MC ... GS
0.00
0.25
0.50
0.75
standardized regression coefficient ± SE
e
MC
0.67
1.02
0.9
1.06
GS
MV
positive
f
MV ... MC
MC ... MV
MC ... GS
GS ... MC
0.0
0.3
0.6
0.9
standardized regression coefficient ± SE

Figure S4. Path analysis displaying relationship between three genomic characteristics. Genome level analyses illustrating the relationships between genome size (GS), metabolic capacity (MC) and metabolic versatility (MV). (a) Initial set-ups of potential causal models; (b) Comparing model support for different causal models. Models are ordered by their value of w, a measure of model support, and numbers on the bars correspond to overall p-values of the model. Models with P-values over 0.05 were acceptable (red coloured). (c) Illustration of one of the supported models, metabolic versatility directly causes metabolic capacity, and metabolic capacity directly causes genome size. Blue arrows represent positive correlations. (d) Correlation coefficients and confidence intervals for each pair of variables in the model illustrated in figure S4c. (e) Correlation coefficients and confidence intervals for each pair of variables in the average best model. (f) Average of the best models, weighted by their relative evidence. The correlations among three genomic characteristics become bidirectional when averaged.

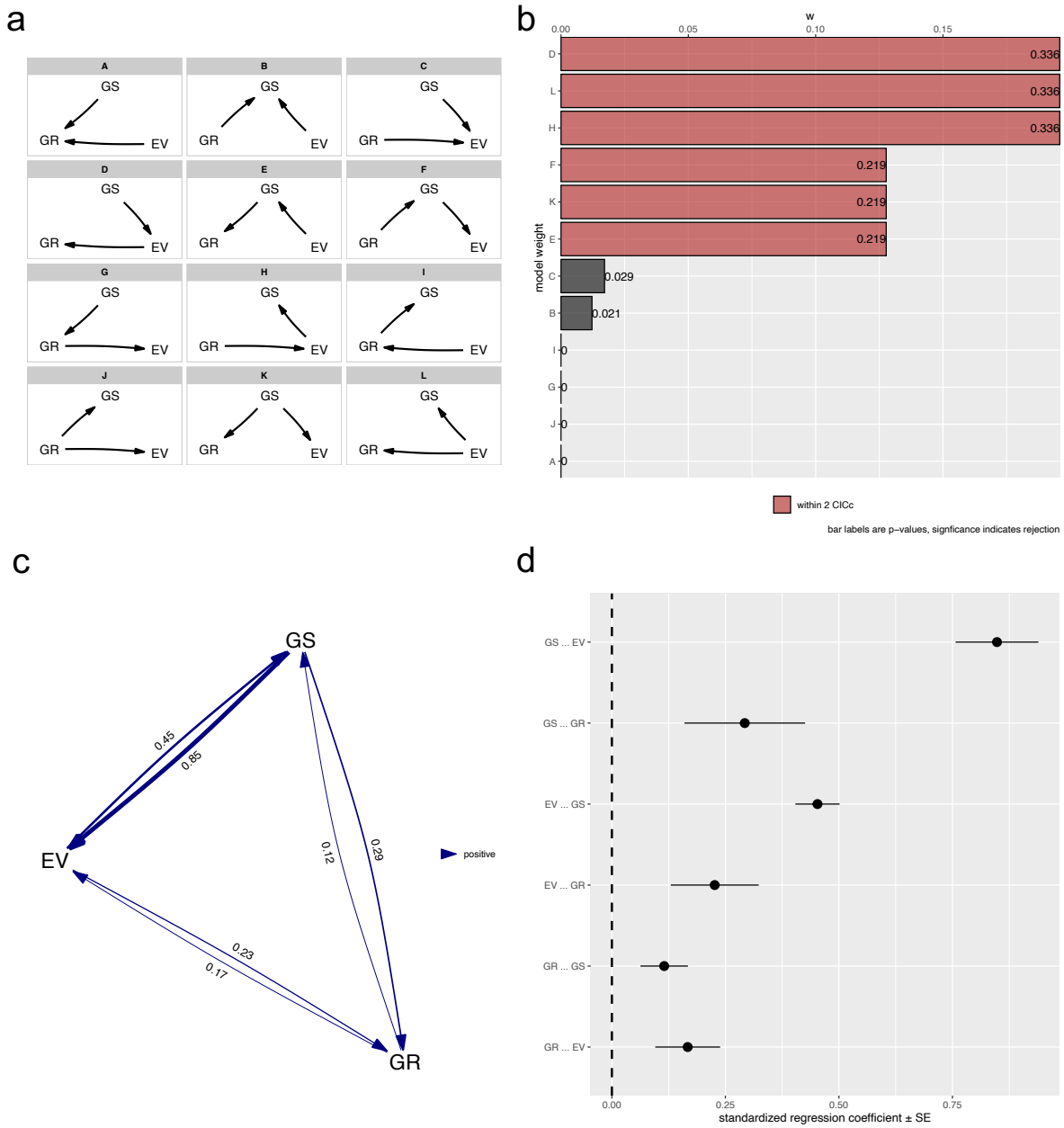


Figure S5. Path analysis displaying relationship between genome size (GS), growth rate (GR), and environmental variability (EV). (a) Initial set-ups of potential causal models; (b) Comparing model support for different causal models. (c) Average of the best models. (d) Correlation coefficients and confidence intervals for each pair of variables in the average best model.

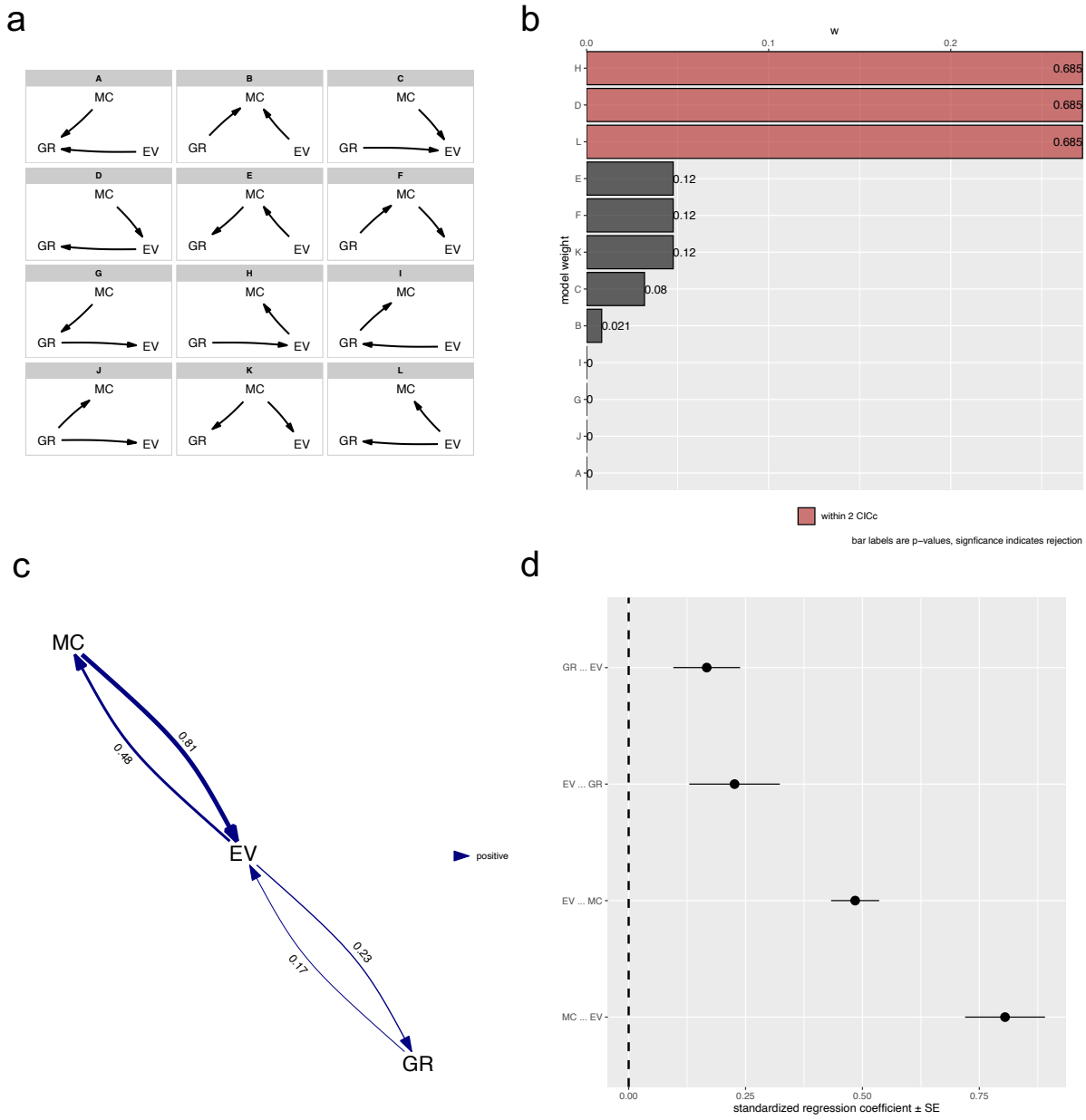


Figure S6. Path analysis displaying relationship between metabolic capacity (MC), growth rate (GR), and environmental variability (EV). (a) Initial set-ups of potential causal models; (b) Comparing model support for different causal models. (c) Average of the best models. (d) Correlation coefficients and confidence intervals for each pair of variables in the average best model.

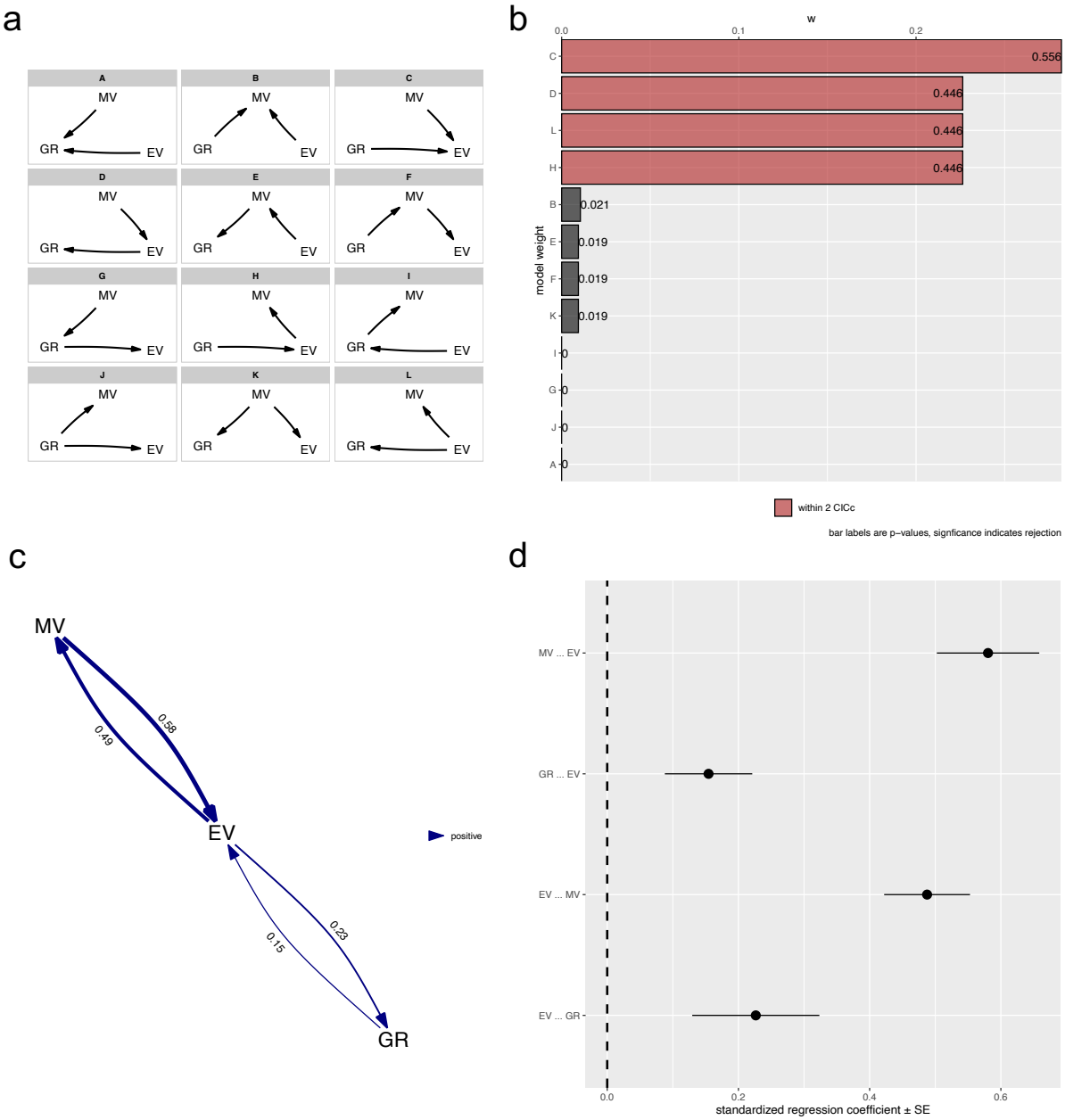


Figure S7. Path analysis displaying relationship between metabolic versatility (MV), growth rate (GR), and environmental variability (EV). (a) Initial set-ups of potential causal models; (b) Comparing model support for different causal models. (c) Average of the best models. (d) Correlation coefficients and confidence intervals for each pair of variables in the average best model.

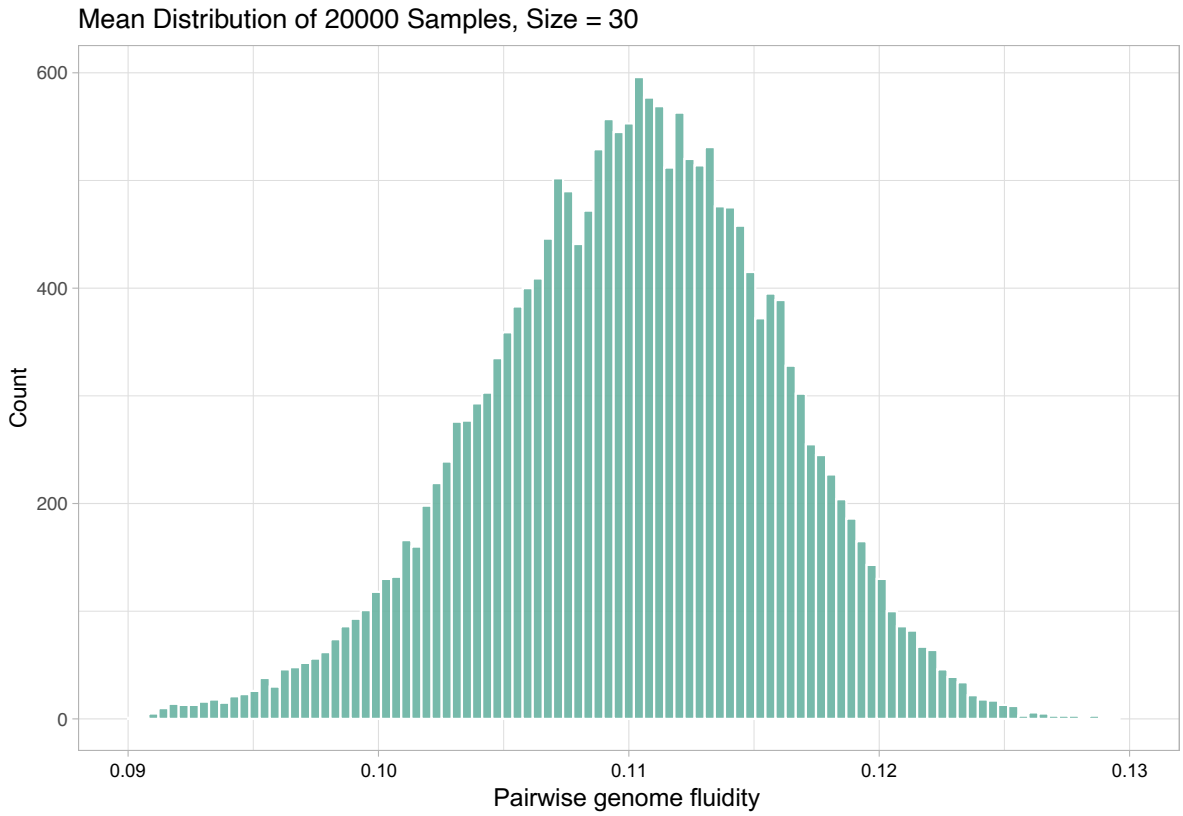


Figure S8. An illustration of applying Central Limit Theorem (CLT) to estimate genome fluidity of species with huge number of genomes, such as *Escherichia coli.* This estimation involved three steps: (1) randomly sampling 30 genomes 20,000 times, (2) calculating the mean pairwise ratio for each sample, and (3) using the mean value of all the sample means as the estimation of genome fluidity for *Escherichia coli*.

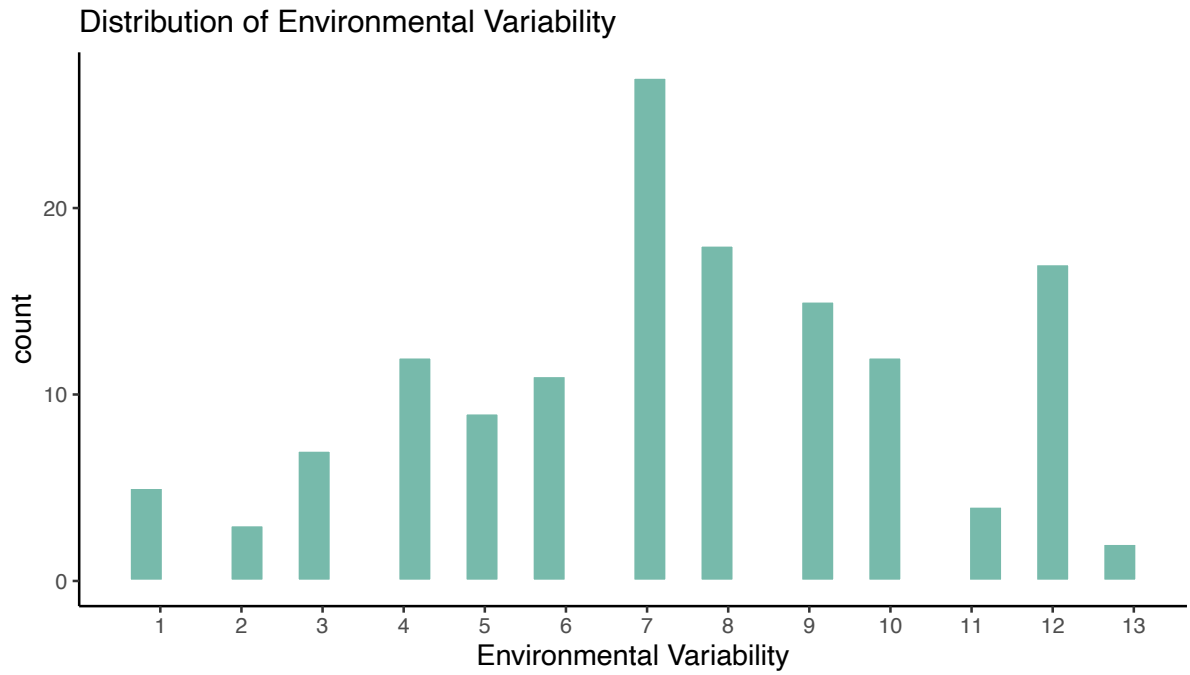


Figure S9. Distribution of environmental variability for 171 species in our dataset. Environmental variability was defined as the number of habitat clusters in which each species could be found.

Figure S10. Phylogenetic tree of 171 species in our dataset, based on tree from the GTDB. Species were represented with the NCBI accession numbers of their representative genomes. Accession numbers starting with "RS_" represent species from RefSeq database, and "GB_" represent species from GenBank database.

Table S1. MCMCglmm analyses results.

	Model description	Term	Posterior mean	l-95% CI	u-95% CI	pMCMC	Sig.	Sample size	R2 (if calculated)
1a	Growth rate ~ Environmental variability	Environmental variability	0.067	0.009	0.125	0.016	*	142 species	0.047
1b	Growth rate ~ Host dependence	Host dependence	-0.265	-0.728	0.191	0.236	ns	170 species	
1c	Growth rate ~ Host dependence (host-dependent vs. both)	Both	-0.040	-0.296	0.190	0.750	ns	170 species	
1d	Growth rate ~ Host dependence (host-dependent vs. free-living)	Free-living	-1.025	-1.848	-0.241	0.016	*	170 species	
1e	Growth rate ~ Environmental variability + Host dependence + Genome fluidity	Environmental variability	0.074	0.013	0.130	0.015	*	141 species	
		Host dependence	-0.418	-0.915	0.060	0.090	ns	141 species	
		Genome fluidity	2.231	-3.668	8.332	0.470	ns	141 species	
1f	Growth rate ~ Environmental variability + Genome fluidity	Environmental variability	0.067	0.008	0.123	0.023	*	141 species	
		Genome fluidity	1.681	-4.334	7.925	0.589	ns	141 species	
1g	Growth rate ~ Host dependence + Genome fluidity	Host dependence	-0.318	-0.8333	0.160	0.208	ns	141 species	
		Genome fluidity	2.565	-3.750	8.575	0.422	ns	141 species	
2a	Growth rate ~ Genome size	Genome size	0.000527	0.000204	0.000853	0.003	**	170 species	0.133
2b	Growth rate ~ Metabolic capacity	Metabolic capacity	0.00128	0.00041	0.00225	0.004	**	170 species	0.089
2c	Growth rate ~ Metabolic versatility	Metabolic versatility	0.01281	0.00169	0.02584	0.034	*	170 species	0.036
2d	Metabolic capacity ~ Genome size	Genome size	0.347	0.335	0.361	< 2e-04	***	171 species	
2e	Metabolic versatility ~ Genome size	Genome size	0.019	0.017	0.021	< 2e-04	***	171 species	
2f	Metabolic versatility ~ Metabolic capacity	Metabolic capacity	0.056	0.050	0.062	< 2e-04	***	171 species	

2g	Growth rate ~ Genome size + Metabolic capacity + Metabolic versatility	Metabolic capacity	-0.0036	-0.0083	0.0011	0.135	ns	171 species	
		Metabolic versatility	0.0010	-0.0193	0.0244	0.932	ns	171 species	
		Genome size	0.0018	0.0003	0.0033	0.020	*	171 species	
2h	Growth rate ~ Genome size + Metabolic capacity	Metabolic capacity	-0.00350	-0.00799	0.00049	0.106	ns	171 species	
		Genome size	0.00175	0.00020	0.00327	0.029	*	171 species	
2i	Growth rate ~ Genome size + Metabolic versatility	Metabolic versatility	-0.00614	-0.02525	0.01389	0.532	ns	171 species	
		Genome size	0.00066	0.00013	0.00120	0.017	*	171 species	
2j	Growth rate ~ Genome Size + Genome Fluidity	Genome Fluidity	5.426	0.087	10.929	0.053	ns	170 species	
		Genome size	0.00050	0.00017	0.00081	0.003	**	170 species	
3a	Genome size ~ Environmental variability	Environmental variability	102.02	80.51	122.54	< 2e-04	***	142 species	
3b	Metabolic versatility ~ Environmental variability	Environmental variability	2.088	1.562	2.628	< 2e-04	***	142 species	
3c	Metabolic capacity ~ Environmental variability	Environmental variability	36.50	28.95	44.31	< 2e-04	***	142 species	
3d	Genome size ~ Host dependence (host-dependent vs. both)	Both	142.73	38.60	255.01	0.0106	*	142 species	
3e	Genome size ~ Host dependence (host-dependent vs. free-living)	Free-living	34.72	-340.21	395.36	0.849	ns	142 species	
3f	Metabolic capacity ~ Host dependence	Host dependence	75.752	-2.523	148.166	0.052	ns	142 species	
3g	Metabolic versatility ~ Host dependence	Host dependence	4.528	-1.178	10.010	0.123	ns	142 species	
3h	Genome fluidity ~ Environmental variability	Environmental variability	0.0003	-0.0013	0.0018	0.731	ns	142 species	

Evolution Letters, 2023, **7**(5), 315–330
https://doi.org/10.1093/evlett/qrad029
Advance access publication 18.July 2023
Letter

Signatures of kin selection in a natural population of the bacteria *Bacillus subtilis*

Laurence J. Belcher, Anna E. Dewar, Chunhui Hao, Melanie Ghoul, Stuart A. West

Department of Biology, University of Oxford, Oxford, United Kingdom

Corresponding author: Department of Biology, University of Oxford, Oxford OX1 3SZ, United Kingdom. Email: laurence.belcher@biology.ox.ac.uk
M.G. and S.A.W. are joint last authors.

Abstract

Laboratory experiments have suggested that bacteria perform a range of cooperative behaviors, which are favored because they are directed toward relatives (kin selection). However, there is a lack of evidence for cooperation and kin selection in natural bacterial populations. Molecular population genetics offers a promising method to study natural populations because the theory predicts that kin selection will lead to relaxed selection, which will result in increased polymorphism and divergence at cooperative genes. Examining a natural population of *Bacillus subtilis*, we found consistent evidence that putatively cooperative traits have higher polymorphism and greater divergence than putatively private traits expressed at the same rate. In addition, we were able to eliminate alternative explanations for these patterns and found more deleterious mutations in genes controlling putatively cooperative traits. Overall, our results suggest that cooperation is favored by kin selection, with an average relatedness of $r = .79$ between interacting individuals.

Keywords: cooperation, kin selection, public goods, population genetics, relatedness, inclusive fitness

Lay Summary

Bacteria produce and secrete a wide range of molecules. The benefits of these molecules can be shared by nearby bacterial cells. For example, secreted molecules that deactivate certain antibiotics provide protection to the whole group, including cells that do not produce the molecule themselves. Laboratory experiments have shown that this is a form of cooperation, which evolves because it benefits closely related cells that share the gene for cooperation (kin selection). However, it has been challenging to find evidence for cooperation and kin selection in natural bacterial populations. To address this, we have used a new way of studying cooperation in natural populations. Evolutionary theory tells us that cooperation leaves a distinct "footprint" in DNA sequence data. We examined the bacterial genomes of the soil-dwelling bacteria *Bacillus subtilis* taken from a natural population in Dundee, Scotland, and used this theory to look for evidence of cooperation in the DNA sequences. Our results suggest that these shared molecules are indeed cooperative traits that have been favored by kin selection in natural populations.

Introduction

Laboratory studies have suggested that bacteria cooperate in a diversity of ways (Strassmann et al., 2011; West et al., 2006). Individual cells produce and secrete molecules to collectively scavenge nutrients, fight antibiotics, and move through their environment (Dugatkin et al., 2005; Griffin et al., 2004; McNally et al., 2014). This cooperation is however vulnerable to cheating by nonproducers, which withhold their own cooperation while benefiting from that of others (Ghoul et al., 2014). A resolution to this vulnerability is kin selection, where cooperation is favored because the benefits of cooperation go to related cells that share the gene for cooperation (Hamilton, 1964). Laboratory experiments have also supported a role of kin selection, with experimental evolution, and by showing how clonal growth makes neighboring cells highly related, and limited diffusion keeps secreted molecules in the neighborhood (Diggle et al., 2007; Griffin et al., 2004; Kümmerli et al., 2009; Strassmann et al., 2000; Velicer et al., 2000).

In contrast, there is little evidence for cooperation and kin selection in natural populations of bacteria outside the lab, with the exception of *Pseudomonas aeruginosa* (Andersen et al., 2015; Butaite et al., 2017; Cordero et al., 2012; Ghoul et al., 2017). The extent to which bacteria cooperate and interact with close relatives is likely to be highly dependent on environmental conditions. It is hard to know whether the artificial environments and gene knockouts of lab experiments are representative of natural populations. Experiments have also shown that some traits can be cooperative in some environments but private in others (Jautzus et al., 2022). Across species, comparative studies have shown that cooperation is more common in species where relatedness is higher (Fisher et al., 2013; Simonet & McNally, 2021), but this does not help us determine whether specific traits are evolving as cooperative public goods. For this, we need a way to study bacteria in their natural environment.

A combination of bioinformatics and molecular population genetics offers a promising method to test for cooperation and

Received April 10, 2023; revisions received June 14, 2023; accepted July 7, 2023

kin selection in natural populations. Theory from population genetics predicts that kin selection leaves a distinct signature (footprint) of selection, which we can detect in genome sequence data (Figure 1) (Linksvayer & Wade, 2009, 2016; Van Dyken & Wade, 2010, 2012; Van Dyken et al., 2011). Genes that code for private traits provide a direct benefit to the individual expressing them. Genes that code for cooperative traits provide indirect benefits to other cells in the population. In a nonclonal population, the cells that benefit from cooperation might not carry the gene for cooperation, as relatedness $r < 1$. If the benefits of cooperation are going to cells that do not carry the gene, then this relaxes selection on cooperative traits relative to private traits. In a clonal population, the cells that benefit from cooperation will also carry the cooperative gene, as relatedness $r = 1$. This means that selection will not be relaxed on cooperative traits in clonal populations, but will be relaxed in nonclonal populations.

The relaxed selection when $r < 1$ results in an increased probability of fixation for deleterious mutations and a decreased probability of fixation for beneficial mutations (14–16). The consequence of this change in fixation probabilities, when $r < 1$, is that it would lead to increased polymorphism and divergence in cooperative genes relative to genes that have direct fitness effects (Figure 1). Consequently, by examining patterns of polymorphism and divergence, we can test for signatures of cooperation favored by kin selection. This method was first applied to the signal and response QS genes across several species of bacteria (Van Dyken & Wade, 2012) and has since been applied to the social amoeba *Dictyostelium discoideum* and the bacteria *P. aeruginosa*. The results from *D. discoideum* were mixed, although relatedness is close to $r = 1$ in this species, so we might not expect a significant signature of kin selection (de Oliveira et al., 2019; Gilbert et al., 2007; Noh et al., 2018; Ostrowski et al., 2015). The results from *P. aeruginosa* found evidence for cooperative traits favored by kin selection (Belcher et al., 2022).

However, these previous studies were unable to account for some potentially confounding factors. First, the genomes had been collected in a variety of geographic locations and/or over a long period time. For example, the *P. aeruginosa* genomes were sampled over several decades and from six different continents (Belcher et al., 2022). The underlying population genetic theory assumes that all genomes are sampled from a single population, at a single time point. Violating this assumption could have led to biased or spurious results (Hahn, 2018). Second, previous analyses were unable to directly test whether genes have conditional fitness effects. If a gene only has fitness effects in certain environments or certain generations, then this can also lead to relaxed selection, with increased polymorphism and divergence (Van Dyken & Wade, 2010). For example, the study on bacteria compared to genes that may vary substantially in sociality and conditional expression (van Dyken & Wade, 2012), and the study on *P. aeruginosa* lacked data on gene expression, and so made targeted comparisons between cooperative and private traits that were likely to be co-expressed at the same time (Belcher et al., 2022). If this assumption did not hold, then their patterns could alternatively be explained by conditional expression rather than kin selection for cooperation (Van Dyken & Wade, 2010).

We were able to address these problems by taking advantage of two recent data sets in the bacteria *Bacillus subtilis*. *Bacillus subtilis* is found in soil and the gastrointestinal tracts of several animals, including humans, and is used on an industrial scale by biotechnology companies (Logan & De Vos, 2015). A number of laboratory studies have suggested that *B. subtilis* is a highly cooperative species, which secretes a number of potentially cooperative enzymes (Arnaouteli et al., 2021; Dragoš et al., 2018; Kalamara et al., 2018; Veening et al., 2008). First, we used a natural population consisting of 31 environmental isolates collected as part of a citizen science project in Dundee, Scotland (Kalamara et al., 2021) (Supplementary Table 1). All these strains were collected from the wild around the same time and from similar niches. Second, we were able to directly test our assumptions about genes being co-expressed at the same time, by using two genome-wide studies of gene expression across several time points during biofilm formation (Futo et al., 2021; Pisithkul et al., 2019). This gene expression data allowed us to test whether

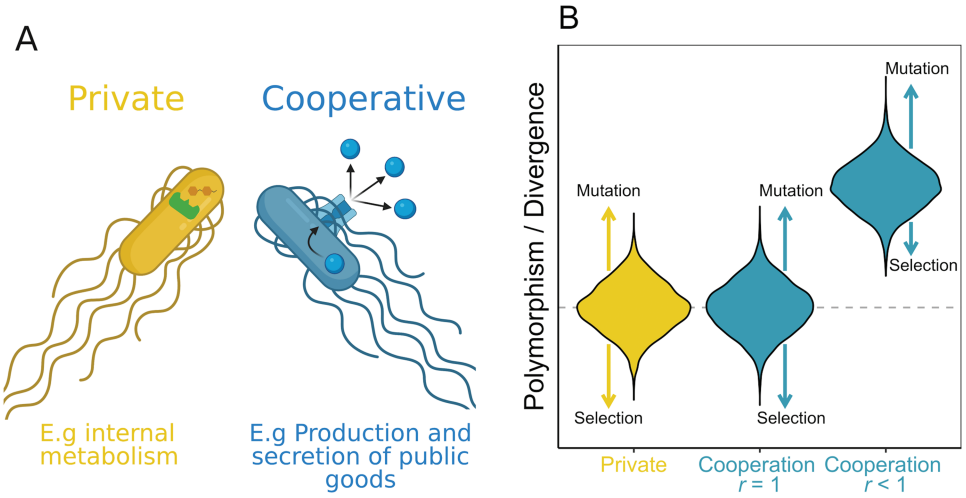


Figure 1. Population genetic theory for cooperative traits. (A) Representation of the categorization of traits as private (yellow) or cooperative (blue). Cooperative traits involve the production and secretion of molecules whose fitness benefits are shared with other nearby cells. Private traits are those whose fitness benefits are only felt by the individuals expressing the trait (B) Prediction for relative polymorphism and divergence for private (yellow) and cooperative (blue) genes. If relatedness, $r = 1$, then cooperative genes (middle blue violin) should have the same polymorphism and divergence as private genes (left yellow violin). In contrast, if $r < 1$, then cooperative genes (right blue violin) should show greater polymorphism and divergence than private genes. Figure based on Linksvayer and Wade (2009), van Dyken & Wade (2010) .

the groups of private and cooperative genes that we compare in our main analysis tend to be expressed together. An additional advantage of this population genetics approach is that it tests for signatures of kin selection over recent evolutionary time, and so provides an answer averaged across the different environments encountered in nature, rather than examining a single environment.

Materials and methods

Strains

We use the whole-genome sequences of 31 strains of *B. subtilis* (Kalamara et al., 2021). The strains are environmental isolates, collected from a citizen science project in Dundee where people brought soil samples from their garden. *Bacillus subtilis* is most commonly found in soil, but is also found living as a commensal in animal intestines (Tam et al., 2006) and in marine environments (Fan et al., 2011). While durable spores than can disperse through the air can allow long-distance migration (Roberts & Cohan, 1995), and strains do not phylogenetically cluster based on environment (Brito et al., 2018), several factors are in favor of using these strains as our natural population. First, these samples were all collected at the same time for the same project (Kalamara et al., 2021). Furthermore, we know that rates of migration between populations scales with geographic distance, most of the sequence diversity within the species is contained in local population (Roberts & Cohan, 1995), and there is evidence for fine-scale genetic structure in micro-scale populations (Duncan et al., 1994).

We downloaded raw sequence data for each strain from the European Nucleotide Archive (accession number PRJEB43128). The full list of strains can be found in Supplementary Table S1.

Genes regulated by quorum sensing

For our set of quorum sensing (QS)-controlled genes, we combine three published data sets: First, 88 genes controlled by ComXAP (Comella & Grossman, 2005); second, 114 genes controlled by *degU* (Kobayashi, 2007); third, 40 genes controlled by Spo0A (Molle et al., 2003). We did not use a fourth possible data set, of 166 late competence genes affected by the transcription factor ComK, which are indirectly regulated by ComA (Berka et al., 2002). This is partly because our undomesticated reference strain (NCIB 3610) carries a plasmid-encoded protein that interferes with the competence machinery (Konkol et al., 2013) (Supplementary S7). In addition, we wanted to focus on the QS systems known to produce public goods (Figure 3 of Kalamara et al., 2018).

Identifying social genes

We used an artisan approach to identify social genes, whereby we categorize genes as social based on laboratory studies, which have demonstrated that a trait is cooperative. The gold-standard test for a cooperative gene involves a wild-type strain that produces the traits and a mutant strain that does not. If a trait is cooperative, then the producer will outperform the nonproducer when each is grown clonally, but nonproducers will outperform producers in groups (West et al., 2006). As an example, we look at the first gene on the list, *bslA* (formerly known as *yuaB*), which is involved in biofilm formation, and specifically in making the biofilm hydrophobic to resist chemical attack (Arnaouteli et al., 2017). A nonproducer of *bslA* cannot form normal biofilms on its own, but can get into mature biofilms when in co-culture with producers (Ostrowski et al., 2011). Further work mixing producers and nonproducers at a range of starting ratios demonstrated that the biofilm can maintain function as long as >50% of cells are producers (Arnaouteli et al., 2017).

The full list of cooperative genes can be found in Supplementary Table 2.

For the robustness check of whether deleterious mutations are over- or under-represented in cooperative genes, we used the protein localization tool PSORTb 3.0 (Yu et al., 2010). We categorize cooperative genes as those which PSORTb predicts to be extracellular. We also follow previous studies in removing genes for which PSORTb cannot make a definitive prediction (Dewar et al., 2021).

Controlling for conditional expression

Conditional expression can lead to the same signatures of relaxed selection as kin selection for cooperation. We directly examined expression rates for the set of genes regulated by QS, using data from (Futo et al., 2021), who measured gene expression of >4,000 genes at 11 time points during biofilm formation.

For any pair of genes, we can calculate the correlation in gene expression across the 11 time points of the biofilm. For the 178 QS genes in our data set, there are 15,753 unique pairs of genes. The mean pairwise Spearman's correlation in gene expression is .302. To test whether this set of genes is more or less correlated than a randomly chosen set of genes, we use a bootstrap approach. We take a random set of 178 genes and calculate mean pairwise correlation in the same way as before. Then we repeat 10,000 times. We find that the correlation in expression of our QS-controlled genes is higher than in 99.7% of our bootstrap samples (Supplementary Figure 3), demonstrating that our candidate set of genes is appropriate for our analysis of signature of selection.

Other cooperative traits

We also examined five other types (groups) of traits, where we could compare genes for putatively private and cooperative traits, which are likely to be expressed at similar rates (Table 1; Figure 4). First, we used iron-scavenging via siderophores, which is a well-studied cooperative trait that is important for growth and survival of bacteria (Griffin et al., 2004; Kümmerli et al., 2015). Specifically, we looked at the *B. subtilis* siderophore bacillibactin (Miethke et al., 2006; Pi & Helmann, 2017). We classified the genes for biosynthesis bacillibactin as cooperative, and genes for the uptake and release of bound bacillibactin as private (Supplementary Table 3). We classified the genes for biosynthesis bacillibactin as cooperative and genes for uptake and release of bound bacillibactin as private (Supplementary Table 3).

Second, we looked at antibiotic resistance genes. There are many mechanisms of antibiotic resistance, some of which are cooperative. For example, the secretion of beta-lactamases is a cooperative trait as they detoxify the external environment, providing benefits to the local population (Amanatidou et al., 2019; Dugatkin et al., 2005; Frost et al., 2018; Wang et al., 2023). We also classified aminoglycoside resistance as cooperative, as the modification of the antibiotic detoxifies the local environment (Poole, 2005). For the private genes, we used the eight ABC transporters that are thought to be involved in multidrug resistance by pumping antibiotics outside the cell (Quentin et al., 1999; Torres et al., 2009) (Supplementary Table 4).

Third, we looked at the range of peptidase proteases produced by *B. subtilis*. The functions of these proteases are broad, covering processing, regulation, and feeding, but we can separate them into cooperative and private genes by looking at those that are secreted (i.e., are extracellular) and those that are not secreted (Harwood and Kikuchi, 2022). The secreted proteases

Table 1. Traits used for comparisons of cooperative versus private genes The full gene lists are in Supplementary Tables 2–7.

Trait	Private genes	Cooperative genes
1. Quorum sensing traits	Genes that only affect the fitness of the producing cell (*N* = 25)	Genes for public goods that potentially provide benefits to the local group of cells (*N* = 153)
2. Iron scavenging	Genes for uptake and use of iron via the *B. subtilis* siderophore bacillibactin (*N* = 5)	Genes for biosynthesis and secretion of bacillibactin (*N* = 5)
3. Antimicrobial resistance	Genes for ABC transporters involved in multidrug resistance. These genes transport the intact antibiotic outside the cell (*N* = 11)	Genes for enzymes that deactivate beta-lactam and aminoglycoside antibiotics, providing cooperative benefits to other cells (*N* = 3)
4. Proteases	Genes for intracellular proteases, which are likely to be involved in processing and regulation of proteins within the cell (*N* = 9)	Genes for extracellular proteases, which are likely to be involved in collective feeding and motility (*N* = 8)
5. Toxins	Genes for contact-dependent LXG toxins, which are delivered by a Type VII secretion system (*N* = 6)	Genes for the diffusible secreted toxin bacilysin, which is activate against a broad range of bacterial competitors (*N* = 8)
6. Antimicrobial activity	Genes for the secondary metabolite bacillaene, which defends against predation by *M. xanthus* (*N* = 13)	Genes for the secondary metabolite plipistatin, which attacks fungal competitors (*N* = 4)

are more likely to have cooperative fitness effects on other cells, through nutrition, interacting with host immune systems etc. (Supplementary Table 6).

Fourth, we looked at toxins. For the cooperative genes, we used bacilycin, which is a secreted antimicrobial peptide that is active against a range of bacteria (Ertekin et al., 2020). Because bacilycin can diffuse through the environment, it likely has cooperative fitness effects on others. *Bacillus subtilis* also has many toxins that are involved in contact-dependent inhibition, and therefore likely have private effects on fitness. For the private gene, we used six LXG toxins (Kobayashi, 2021), which are delivered by a Type VII secretion system. While these toxins can still have cooperative effects by removing competitors, they are by their nature less cooperative than secreted molecule such as bacilycin. Both of these sets of genes are under the control of the DegS–DegU system (Supplementary Table 5).

Fifth, we looked at antimicrobials. *B. subtilis* produce a series of antimicrobial molecules, which vary in which organisms they target, how they act, and how they are secreted. We can however distinguish between antimicrobials that have a more defensive role in traits such as predation avoidance and those that have a more offensive role in competition with other species. While both of these categories likely have some component of cooperative and private effects on fitness, the offensive ones will be relatively more cooperative. This is because defensive molecules have a stronger effect on the individuals producing them (Supplementary Table 7).

Identifying deleterious mutations

We used the variant annotation tool SnpEff (Cingolani et al., 2012) to look for SNPs that generate deleterious mutations in our data set. Specifically, we annotate two types of mutations: (a) premature stop codons and (b) frameshift mutations. This gives us a list of genes that have at least one deleterious mutation. To test whether a given set of genes are overrepresented for deleterious mutations, we use two percentages: (a) the % of genes in the whole genome that are in that set and (b) the % of genes with deleterious mutations that are in this set. We compare these values using a binomial test with the null hypothesis that the number of deleterious mutations in the gene set is equal to that expected by the frequency of the gene set. For any given gene set, we conduct a further test where we use the total number of deleterious mutations in that set of genes, rather than just the presence/absence of deleterious mutations for that gene.

Statistics and figures

We conducted all statistical analysis in R (R, 2020). For the main statistical analysis comparing molecular population genetic parameters of cooperative and private genes, we use one of two statistical tests, depending on the variable in question. Some of the molecular population genetic parameters we calculate are normally distributed, but others tend to be highly skewed. The skewed variables cannot be transformed into normally distributed, so we use different statistical tests. For variables that are normally distributed, we used an ANOVA to compare the three groups of genes. Because of unequal sample sizes, we used Welch ANOVA that does not require equal variance. We used the Games–Howell post hoc test, which is similar to Tukey's HSD, but designed for Welch ANOVA where we do not have to assume equal variance. For variables that are not normally distributed, we used the Kruskal–Wallis test, which compares medians. We then used the Dunn test for post hoc comparisons of groups.

All results figures were made using the ggplot2 package in R (Wickham, 2016) using color palettes from the packages wesanderson (github.com/karthik/wesanderson) and BirdBrewer (https://github.com/lauriebelch/BirdBrewer). Figure 4 illustrating the secondary comparisons was made using BioRender.

Bioinformatics

Raw reads for each of the 31 strains were downloaded from the European Bioinformatics Institute's European Nucleotide Archive (accession number PRJEB43128). We then used an SNP calling pipeline to find SNPs in each strain compared to the reference NCIB 3610 (accession NZ_CP020102.1).

Trimming and quality control

We used Trimmomatic to remove adapters remove low-quality reads, which we did by removing leading and trailing reads if the quality score was <3 or if average quality in a four-base sliding window was <20. We manually checked the output of this step using the reports produced by FastQC (Andrews, 2010).

Mapping

We used BWA (Li & Durbin, 2009) to map reads from each strain to the reference strain. We used SAMtools (Li et al., 2009) to convert the mapping files from BAM to SAM and used Picard tools (Broad Institute, 2019) to remove PCR duplicates.

Variant calling

We used BCFtools (Danecek et al., 2021) to call variants on all strains and produce a VCF file that can be read by R for population genetic analysis.

Filtering and quality control

We conducted further filtering to remove indels, and filter for mapping quality, read depth, and strain bias using the default setting of SAMtools vcfutils python script. We then removed all sites which hadn't been called in at least 80% of strains. We also used the coverageBed tool in BEDtools (Quinlan & Hall, 2010) to record the percentage of each gene length that had been mapped, in order to adjust per-site measures to the correct length. After filtering, we had a total of 256,769 SNPs among the 31 strains.

Outgroup

We used the phylogeny in Kalamara (2021) (the source of these strains) to identify *Bacillus subtilis subsp. spizizenii* str. W23 as an appropriate outgroup (accession NC_014479.1, raw sequencing data SRR2063059). We used the same variant calling pipeline as above to produce a second VCF which included the SNPs from the outgroup.

Population genetic measures

We used the PopGenome package from R (Pfeifer et al., 2014) to conduct the main molecular population genetic analysis. All parameters were scaled to the corresponding mapped gene length, and any gene with mapped length <50% of their full length or lacking polymorphism data was removed from the analysis, leaving 3,817 genes for the population genetics analysis. Using PopGenome, we calculated Nucleotide polymorphism, Tajima's *D*, Fu and Li's *D**, the Mcdonald–Kreitman *p*-value, Direction of Selection statistic, and neutrality index. We also calculated separate measures for synonymous and non-synonymous sites where appropriate.

To calculate divergence, we measured the rate of protein evolution K_a/K_s by comparing the reference strain to the outgroup. We did this by creating a pseudogenome of the outgroup by inserting the relevant SNPs into the reference sequence using the GATK suite of tools (McKenna et al., 2010). This pseudogenome could then be read by R, and we used the seqinR package (Charif & Lobry, 2007) to calculate divergence.

Results and discussion

We compared genes controlling traits that are hypothesized to be cooperative, with traits that are hypothesized to be private. We identified six different types of putatively cooperative behavior, where an appropriate comparison could be made with genes controlling private traits, which are likely to be expressed at similar rates (Table 1). We used the first type, QS, for the main analysis, and we summarized the other five types at the end of the Results and discussion section.

Quorum sensing

We started by examining genes induced by the ComQXPA QS signaling system (Arnaouteli et al., 2021; Kalamara et al., 2018). This system regulates gene expression in response to the density of a diffusible signal molecule. At high cell densities, the density of the signal molecule increases, causing ComA to activate, and the upregulation of a number of traits (Comella & Grossman, 2005; López & Kolter, 2010; Nakano et al., 1991; Špacapan et al., 2020). We categorized genes as cooperative or private based on a search of the literature on *B. subtilis* (Methods). For example, the 15 genes coding for the exopolysaccharide EPS are classed as cooperative. EPS is the main biofilm matrix component (Branda et al., 2001) and is required for biofilm formation (Branda et al., 2004). EPS is costly to produce, it provides benefits to nonproducers, and nonproducers can exploit producers (Van Gestel et al., 2014). This is a classic public good. Similarly, TasA, a protein fiber that is needed for biofilm structural integrity (Erskine et al., 2018; Romero et al., 2010), has also been shown in lab experiments to be a public good (Dragoš et al., 2018). Mutants lacking genes for either EPS or TasA can also complement each other, providing further evidence that the benefits of these genes are shared (Branda et al., 2006). Private genes include those coding for traits such as asparagine synthase (AsnB), which controls peptidoglycan hydrolysis for cell growth and cell-wall synthesis (Yoshida et al., 1999). We found that QS controls a mixture of private and cooperative traits in *B. subtilis*, categorizing $N = 25$ of our QS-controlled genes as cooperative, and $N = 153$ as private (Supplementary Table 2).

We started by focusing on QS because it offers a number of advantages for our purpose. First, the large size and nature of this network means that there are sufficient private and cooperative genes for a targeted analysis ($N = 153$ and $N = 25$, respectively). Second, shared control by the same signaling system means that private and cooperative genes controlled by the QS system are likely to be co-expressed at the same time (Azimi et al., 2020a; Rutherford & Bassler, 2012). Third, there are data on gene expression allowing us to directly look at co-expression rates (Futo et al., 2021). Fourth, the coregulation of genes acts as a control for mutations in noncoding regulatory and promoter regions that could affect the production of cooperative public goods.

QS: testing if genes are co-expressed at the same time

Differential gene expression can also influence the strength of selection and so needs to be accounted for. Theory tells us that the fraction of generations in which a trait is expressed can determine the extent to which selection is relaxed (Van Dyken & Wade, 2010). We therefore need to compare genes that are switched on or off in the same conditions. Shared control by the same signaling system means that private and cooperative genes controlled by the QS system are highly likely to be co-expressed (Azimi et al., 2020a; Rutherford & Bassler, 2012). Furthermore, many genes regulated by QS form operons that share a promoter and are transcribed and translated together (Kalamara et al., 2018). We were, however, also able to test this assumption directly by examining two data sets on gene expression (Futo et al., 2021; Pisithkul et al., 2019).

Futo et al. (2021) measured gene expression at 11 different points in the formation of a biofilm, following normal practice by normalizing their results to the median to convert expression levels to the same scale. This gives us an excellent data set, as we can use simple correlations between pairs of genes to see whether they are up- and downregulated at the same time. For our 178 private and cooperative genes, there are 15,753 unique pairs of genes. The average correlation in gene expression for a pair of these genes is .302 (Spearman's correlation). We then used a bootstrap approach to see if this correlation was greater or lower than for randomly chosen genes. We took a random set of 178 genes and calculate mean pairwise correlation in the same way and repeated this process 10,000 times. We found that the mean pairwise Spearman's correlation was .251 for the random (bootstrap) samples and that the correlation in

expression of our QS-controlled genes was higher than 99.7% of our bootstrap samples (Figure 2). Consequently, our candidate set of genes has expression rates that are correlated significantly higher than expected by chance, supporting our choice for their use in an analysis of signature of selection ($p < .004$). This is unsurprising, given that many of our QS genes are clustered together as operons, while the randomly sampled genes are unlikely to be. However, for our purposes we only need to test whether genes are co-expressed, we do not need to control for factors like operons that might explain *why* they are co-expressed.

Pisithkul et al. (2019) provided a different data set measuring gene expression in biofilms. Whereas Futo et al. measured expression over 2 months in a biofilm with a solid–air interface, Pisithkul et al. focused on the initial stages of biofilm growth, measuring expression over 24 hr in a biofilm with a liquid–air interface. Using this second data set, we again found that QS-controlled genes have expression rates that are correlated significantly higher than expected by chance ($N = 160$ genes, $p < .02$; Supplementary S11).

QS: polymorphism

Polymorphism (π) is measured as the average number of pairwise nucleotide differences per site in a gene. We found that genes for putatively cooperative traits had significantly greater polymorphism than genes for private traits (Figure 3; ANOVA $F_{2,61} = 11.82$, $p < .001$; Games–Howell test $p < .001$; $N = 25$ cooperative genes, $N = 153$ private genes). Cooperative genes also had significantly greater polymorphism when we only examined nonsynonymous sites (Kruskal–Wallis $\chi^2(2) = 10.7$, $p < .01$; Dunn test $p = .0240$. Figure 4A) or only synonymous sites (ANOVA $F_{2,61} = 7.30$, $p < .001$; Games–Howell test $p = .007$) (Figure 4B). The increase in polymorphism that we see in cooperative traits does not imply an increase in cheating, as it is just the consequence of the relaxation of selection.

The trend for greater synonymous polymorphism is possibly surprising as such sites should be under much weaker selection, and we would not necessarily expect to see an effect of kin selection. However, we also found this pattern in *P. aeruginosa* (Belcher et al., 2022). There is some evidence that differences in the use of preferred codons between cooperative and private genes could explain this pattern (Supplementary S1). Synonymous mutations can also have substantial fitness effects on social traits (Azimi et al., 2020b; Meir et al., 2020; Salverda et al., 2010; Zwart et al., 2018). Because we focus on the relative level of polymorphism, features of *B. subtilis* such as spore-formation cannot explain these patterns unless they only affect cooperative genes.

The ratio between nonsynonymous and synonymous polymorphism does not differ between cooperative and private genes (Kruskal–Wallis $\chi^2(2) = 6.97$, $p = .0306$; Dunn test $p = .183$. Supplementary Figure 1), which reflects the fact that cooperative genes have elevated diversity at both types of site, possibly due to the large fitness effects of many traits (both cooperative and private) that are QS controlled (Schuster et al., 2017; Whiteley et al., 2017) (Supplementary S2).

As an additional control, we were able to repeat all of these analyses comparing cooperative QS genes against a different group of private genes not controlled by QS ($N = 1,832$). This different set of $N = 1,832$ private genes, which we call "background genes" are those that are not controlled by QS and whose products are found in the cytoplasm, where they are least likely to have a cooperative function. In all cases, we found the same pattern, with cooperative QS genes showing elevated polymorphism compared to background genes (Supplementary S2).

QS: divergence

We found that cooperative genes had significantly greater divergence compared to private genes at both nonsynonymous and synonymous sites (nonsynonymous: Kruskal–Wallis $\chi^2(2) = 10.4$, $p = .006$; Dunn test $p = .00553$; Figure 5A; synonymous: ANOVA $F_{2,59} = 7.26$, $p < .01$; Games–Howell test $p = .011$; Figure 5B). We examined synonymous and nonsynonymous sites separately because we measure divergence through rates of protein evolution. A signature of selection can be found at both synonymous and nonsynonymous sites, implying that synonymous variation is not neutral (Belcher et al., 2022; de Oliveira et al., 2019). Because divergence is elevated at both types of sites (similar to polymorphism), there is no difference in the ratio of nonsynonymous to

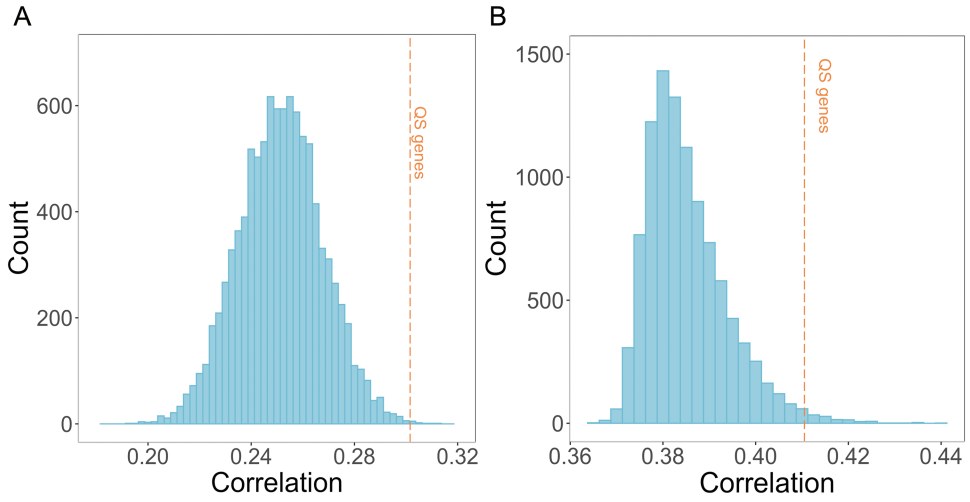


Figure 2. Average correlation in gene expression between genes during biofilm growth. (A) Correlation across 11 time points of biofilm formation for randomly sampled gene sets of the same size as our quorum sensing-controlled genes (N = 178). Data from Futo et al. (2021). (B) Correlation across eight time points of biofilm formation for randomly sampled gene sets of the same size as our quorum sensing-controlled genes ($N = 160$). Data from Pisithkul et al. (2019). In both panels, the orange line shows the average correlation for the quorum sensing-controlled genes, which is >99.7% of our random samples for (A) and >98.5% of random samples for (B).

synonymous divergence (Kruskal–Wallis $\chi^2(2) = 13.32$, $p = .00128$; Dunn test $p = .189$. Supplementary Figure 2), although both sets of QS genes have a higher ratio than the background genes (Supplementary S2). This could reflect stronger selection of both private and cooperative QS traits compared to background traits.

Alternative hypotheses

We were also able to eliminate alternative explanations for the patterns that we observed with both polymorphism and divergence (Figures 3–5). Greater polymorphism would also have been expected if cooperative genes were more likely to be experiencing balancing or frequency-dependent selection, while greater divergence would arise from positive/directional selection leading to fixation of adaptive mutations (Linksvayer & Wade, 2009, 2016). Alternatively, the pattern we observed could have been caused by other differences between putatively cooperative and private genes. We assessed these alternative hypotheses by testing other predictions that they make.

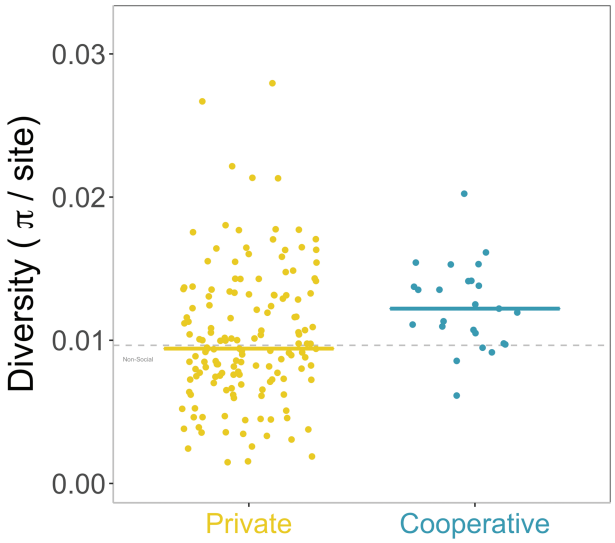


Figure 3. Nucleotide diversity per site for private (yellow) and cooperative (blue) genes controlled by quorum sensing. Each point is a gene, and the horizontal line shows the median for each group. The gray line shows the median for background private genes across the genome.

First, we found no evidence that cooperative genes were more likely to be under balancing selection, which would lead to a deficit of rare alleles in the population (Tajima's *D*, Fu and Li's *F*, Fu and Li's *D*, Supplementary S3). Second, we found no evidence that cooperative genes are more likely to be under positive selection, which would lead to an excess of nonsynonymous divergence compared to nonsynonymous polymorphism, as adaptive mutations would quickly spread and only be detected as divergence. We test for this with the commonly used Mcdonald–Kreitman test and downstream analysis (Neutrality Index and Direction of Selection statistic), and also a modified version of the McDonald–Krietman test that is more conservative and deals better with non-neutrality of synonymous mutations (Supplementary S4). Third, we found no evidence that the patterns observed are due to a lack of statistical power, or any differences in gene length or likelihood of horizontal gene transfer between cooperative and private genes, or a lack of statistical power (Supplementary S5). Fourth, we found no evidence that our results could be explained by noise due to variation in the recombination rate. The set of strains that we analyzed vary in genetic competence, the ability to take up DNA from the environment (Kalamara et al., 2021), which is a form of recombination in bacteria. This variation in competency could create noise in our population genetic measures that focus on SNPs, due to the variation in recombination. Considering the 31 strains we analyzed, 18 are genetically competent (Kalamara et al., 2021). We conducted an analysis of polymorphism and divergence using only the competent strains and found the same patterns as when we use all strains (Supplementary S7). Fifth, we found the same signature of selection when analyzing operons as independent data points, rather than genes (Supplementary S12). Sixth, we found the same signature of selection when removing the small number of essential genes that are regulated by QS in this species (Commichau et al., 2013; Supplementary S13). Seventh, we found that cooperative and private genes do not differ in their maximum expression level (Supplementary S14), which is another factor known to predict selection on genes (Urrutia & Hurst, 2001, 2003).

Finally, we found no evidence that the patterns we observed could be caused by division of labor (Cooper & West, 2018; Cooper et al., 2021; Liu et al., 2021; West & Cooper, 2016), which is a cooperative hypothesis not mutually exclusive with kin selection. In *B. subtilis* biofilms, some cells will produce the

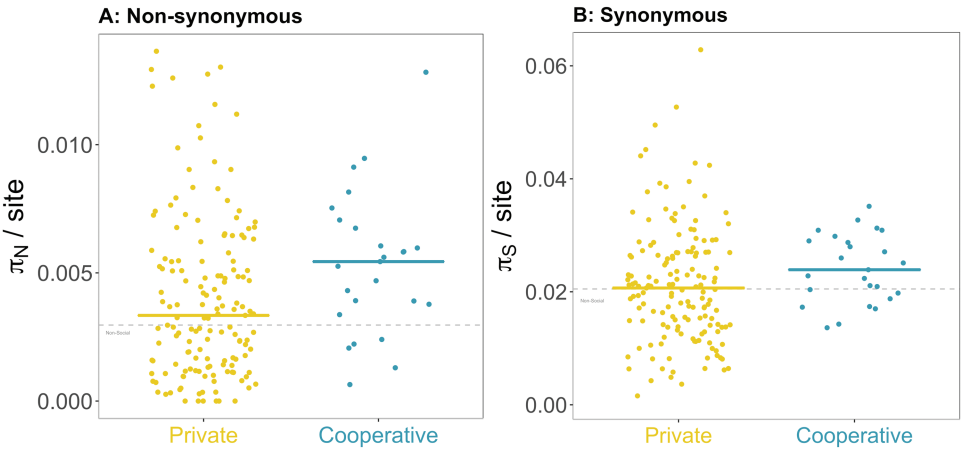


Figure 4. Nucleotide diversity at nonsynonymous (A) and synonymous (B) sites for private (yellow) and cooperative (blue) genes controlled by quorum sensing. Each point is a gene, and the horizontal line shows the median for each group. The gray line shows the median for background private genes across the genome.

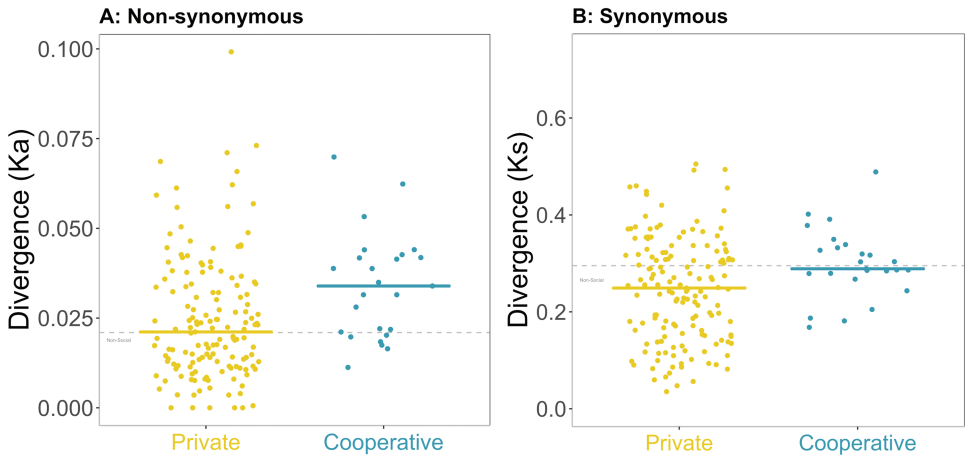


Figure 5. Divergence at nonsynonymous (A) and synonymous (B) sites for private (yellow) and cooperative (blue) genes controlled by quorum sensing. Divergence is measured by rates of protein evolution, for example, number of synonymous substitutions per synonymous site for (B). Each point is a gene, and the horizontal line shows the median for each group. The gray line shows the median for background private genes across the genome.

polysaccharide EPS, and some will produce TasA amyloid fibers (Chai et al., 2008). Mutants lacking one or the other cannot grow alone but can grow together (Dragoš et al., 2018). The extra level of conditional expression in these public goods (over that caused by being QS controlled) could leave a signature of relaxed selection that is not caused by kin selection. To investigate this possibility, we took advantage of the fact that this heterogeneity is ultimately caused by Spo0A. Spo0A is a bistable switch that is active in only a subset of cells (Chai et al., 2008; González-Pastor, 2011) and activates SinR antirepressors which control the operons for both EPS and TasA (Dragoš et al., 2018; Kobayashi, 2008; Molle et al., 2003). If the extra conditionality of division of labor was causing an effect, we would expect that genes under the control of Spo0A (N = 20) would have greater polymorphism than other QS-controlled genes (*N* = 157), but this is not the case (Supplementary S6). We note that the cooperative genes EPS and TasA that are known to have a division of labor both stand out as highly polymorphic within the genes controlled by Spo0A, providing support to our hypothesis that sociality causes the effect. Overall, we can rule out any confounding effect of the extra level of conditionality in some cooperative traits (Supplementary S15).

Other cooperative traits

Our above analyses have provided strong evidence for relaxed selection due to kin selection for cooperation, considering genes controlled by QS. We then tested if the same pattern was found in five other types (groups) of traits, where we could compare genes for putatively private and cooperative traits (Table 1; Figure 6). These genes are likely to be expressed at similar rates, which controls for the confounding effect of conditional expression. The extent to which the distinction between private and cooperative traits can be made varies across these other groups of traits. Consequently, we might not expect to see a signature of kin selection in every case, and so our aim is to see if there is a relatively consistent pattern.

First, *B. subtilis* produces and secretes a siderophore named bacillibactin, which binds to iron in the environment (Miethke et al., 2006; Pi & Helmann, 2017) (Figure 6B). The bound complex can be taken into the cell, but is also available to nonproducers, and is therefore a public good. We separated the genes involved in the bacillibactin pathway into cooperative and private components, with genes involved in biosynthesis and export classed as cooperative, and those involved in uptake and release of bound iron classed as private. This is our strongest comparison, as epistatic interactions mean that production and uptake are necessarily linked. Second, *B. subtilis* exhibits resistance to antimicrobials by either pumping intact antibiotics outside the cell (private) or producing enzymes such as beta-lactamases that detoxify the environment for the entire community (Bucher et al., 2019; Noguchi et al., 1993; Torres et al., 2009) (cooperative; Figure 6C). Third, *B. subtilis* produces proteases to break down proteins, with different proteases acting either inside the cell (private) or secreted to act outside the cell (Harwood & Kikuchi, 2022; Koo et al., 2017; Pohl et al., 2013) (cooperative; Figure 6D). Fourth, *B. subtilis* produces toxins that can either be contact dependent (relatively private) or diffusible throughout the community (cooperative public goods; Figure 4D). However, this comparison is relative, and possibly weak, as killing cells with contact-dependent toxins could also provide a cooperative benefit to other local cells, which experience reduced competition (McNally et al., 2017). Fifth, *B. subtilis* has a number of antimicrobial traits that are more defensive against predators without affecting the predators' growth (private) and those that are more offensive against competitors (cooperative) (Müller et al., 2014; Romero et al., 2007). This is also a weak comparison, as bacillaene (the defensive molecule) is secreted from cells, and so likely also has some cooperative component. However, the defensive traits provide a relatively more private benefit in providing personal protection, whereas the removal of direct competitors by the offensive traits provides a relatively more cooperative benefit. For all comparisons, we find that the set of genes has significantly correlated expression, using the same methodology and data as for the comparison with QS genes (Supplementary S9) (Futo et al., 2021).

The number of cooperative genes was too small to analyze each case separately, for example, the iron-scavenging comparison involves only 10 genes (5 private and 5 cooperative). Consequently, we examined the data in three ways. First, we grouped all cooperative genes together into one set (*N* = 52) and compared them to the grouped private genes across all comparisons (*N* = 194). Second, we grouped all of the cooperative genes from just the five new comparisons (i.e., not QS) (*N* = 27), and compared them to the private genes from these five comparisons (*N* = 41). Third, we consider each of the six categories of cooperative versus private genes (Table 1) as its own data point (*N* = 6).

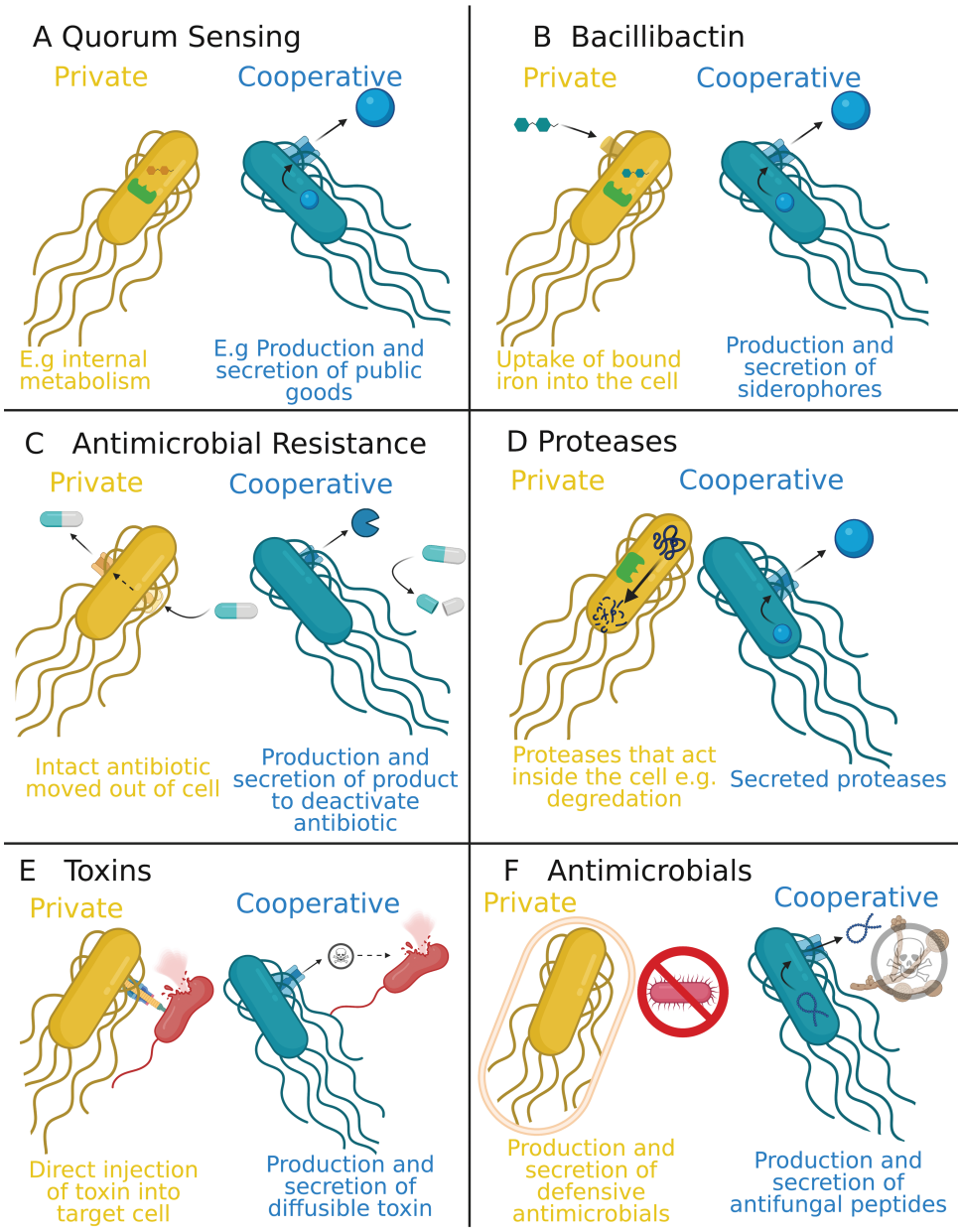


Figure 6. Diagram of how traits are categorized as either private (yellow) or cooperative (blue). (A) Categorization of quorum sensing-controlled traits for the main analysis, with private traits giving fitness benefits only to those expressing the gene, and cooperative traits giving fitness benefits that can potentially be shared with other cells. (B–F) Other cooperative traits: (B) bacillibactin (iron scavenging), (C) antimicrobial resistance, (D) proteases, (E) toxins, (F) antimicrobials. Note that for some of these comparisons it is the relative level of sociality that is different. Figure created with BioRender.com.

Other cooperative traits: polymorphism and divergence

We found that polymorphism was consistently higher in cooperative genes than in private genes (Figure 7). This pattern is consistent across the three different ways that we can analyze our data: All gene comparison (ANOVA $F_{2,127} = 10.59$, $p < .0001$; Games–Howell test $p \leq .001$); just the genes for the social traits other than QS (ANOVA $F_{2,47} = 5.94$, $p < .001$; Games–Howell test $p < .01$); and the six-category comparison (Wilcoxon signed-rank test $V = 21$, $p = 0031$).

Polymorphism at nonsynonymous sites is also significantly higher in cooperative genes than in all private genes. This pattern is consistent across the three different ways that we can analyze our data: All genes comparison (Kruskal–Wallis $\chi^2(2) = 19.71$, $p < 10^{-4}$, Dunn test $p < 10^{-4}$); just the genes for the social traits other than QS (Kruskal–Wallis $\chi^2(2) = 9.90$, $p < .01$, Dunn test $p < .01$); and the six-category comparison (Wilcoxon $V = 21$, $p = .031$) (Supplementary Figure 3).

Polymorphism at synonymous sites is also significantly higher in all cooperative genes than in all private genes when analyzing all genes (ANOVA $F_{2,130} = 4.21$, $p = .016$; Games–Howell test $p = .019$). However, this trend was not significant when examining just the genes for the social traits other than QS (ANOVA $F_{2,47} = 2.28$, $p = .113$; Games–Howell test $p = .10$) or when using categories as data points (synonymous polymorphism is marginally higher in private genes than cooperative genes for the iron-scavenging and AMR categories; Wilcoxon $V = 18$, $p = .156$). We would expect the pattern to be weaker with synonymous

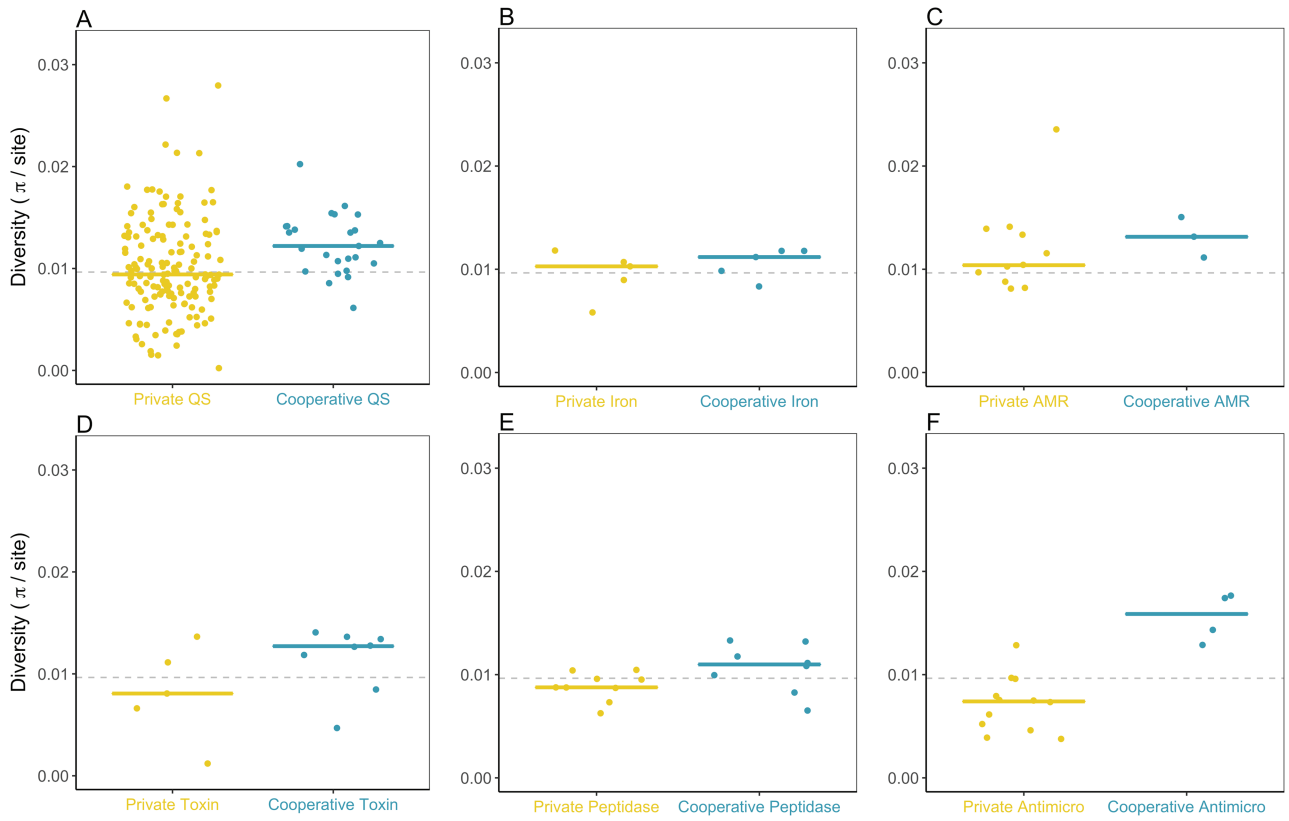


Figure 7. Private (yellow) versus cooperative (blue) polymorphism in genes for six traits. (A) The quorum sensing-controlled genes used in the main analysis. (B–F) The other cooperative traits.

polymorphism, as these sites are likely to be under weaker selection.

Nonsynonymous divergence is significantly greater in all cooperative genes compared to all private genes. This pattern is consistent across the three different ways that we can analyze our data: All genes comparison (Kruskal–Wallis $\chi^2(2) = 22.9$, $p < 10^{-4}$, Dunn test $p < 10^{-4}$); the genes for the social traits other than QS (Wilcoxon $V = 21$, $p = .031$ Supplementary Figure 4); and the six-category comparison (Kruskal–Wallis $\chi^2(2) = 14.7$, $p < .001$, Dunn test $p < .001$). As would be expected, the pattern is more mixed for synonymous divergence. While synonymous divergence is significantly greater in all cooperative genes compared to all private genes (ANOVA $F_{2,125} = 8.33$, $p = .001$; Games–Howell test $p < 001$), this is not the case when examining just the genes for the social traits other than QS (ANOVA $F_{2,45} = 2.10$, $p = .013$; Games–Howell test $p = .39$) or the six-category comparison (Wilcoxon $V = 13$, $p = .688$) (Supplementary Figure 6). Similarly, the ratio between nonsynonymous and synonymous divergence is significantly greater in all cooperative genes compared to all private genes (Kruskal–Wallis $\chi^2(2) = 25.0$, $p < 10^{-5}$, Dunn test $p = .012$), but not when we just looked at the genes for the social traits other than QS (Kruskal–Wallis $\chi^2(2) = 13.2$, $p < .01$, Dunn test $p = .011$) or the six-category comparison (Wilcoxon $V = 18$, $p = .156$) (Supplementary Figure 7). This may reflect differences in selection on synonymous variation on different traits or the weakness of some of these comparisons due to small sample sizes. While the fact that K_a tends to increase in cooperative genes, but K_s stays close to background levels may be a sign of weak positive selection, we confirmed that none of our cooperative versus private comparisons show significant differences in balancing or positive selection (Supplementary S10), suggesting that what we are seeing is a signature of kin selection and that we may just lack power in our other comparisons.

Overall, these results incorporating other traits in addition to QS-controlled genes provide support to the main result that there is a signature of kin selection for cooperation (Figures 3–5). Nonetheless, the a priori distinction between private and cooperative traits is weaker for some comparisons, and we could expect exceptions within the overall pattern. The main exception in our analyses was toxin comparison, where we compared contact-dependent LXG toxins ($N = 6$ genes) to the secreted antimicrobial bacilycin ($N = 8$ genes). Both of these sets of genes are involved in competition in biofilms, and both are controlled by the DegS–DegU system, so likely expressed at the same time (Kobayashi, 2021; Mariappan et al., 2012). Possible confounding factors in this case include the fact that although we classified the contact-dependent toxins as private, they also provide cooperative benefits to local cells by eliminating competitors. The LXG toxins also stand out because three of them are on phage elements (Kobayashi, 2021), and it may be that the strength or type of selection is different on these genes, masking any effect of sociality.

All traits: deleterious mutations

As an additional robustness test for our conclusions, we analyzed deleterious mutations, specifically those that cause loss of function, generate stop codons, or cause a frameshift. If kin selection is favoring cooperation, we should observe more deleterious mutations in genes controlling cooperative traits compared to private traits. This is because relaxed selection slows the rate at which deleterious mutations are purged from the population (Linksvayer & Wade, 2009, 2016; van Dyken & Wade, 2012). We

tested this prediction by looking for deleterious mutations in our SNP data and repeated this analysis with two different data sets.

First, we used all cooperative genes from our six comparisons in Table 1. We measured how many cooperative and private genes have deleterious mutations and compared this to an expectation based on their relative frequency across the genome. Cooperative genes were significantly more likely than private genes to have deleterious mutations, $\chi^2(1) = 12.3$, $p < .001$. This pattern also holds if we count total number of deleterious mutations, rather than just presence or absence, $\chi^2(1) = 11.0$, $p < .001$.

To test the robustness of this result, we repeated the analysis using the localization prediction tool PSORTb to categorize genes for extracellular proteins as "cooperative" and genes for proteins that are not secreted as "private" (Yu et al., 2010). This method has been previously used in many studies to estimate whether genes are for cooperative (social) or private traits (Dewar et al., 2021; Garcia-Garcera & Rocha, 2020; Nogueira et al., 2009, 2012). By using PSORTb, we are able to systematically analyze all genes, which increases our sample size and statistical power compared to the "artisan" approach we use in the main analysis. We removed 17% of all genes with unknown localization, leaving us with a set of genes of known sociality. We found deleterious mutations in 293 genes, of which 17 are cooperative (5.8%). This is significantly more than expected given that cooperative genes only make up 2.0% of genes (binomial test $p < .001$), matching our prediction that deleterious mutations should be biased toward cooperative genes. If we count total deleterious mutations (rather than number of genes with at least one), we see the same pattern, with cooperative genes making up 5.5% of mutations (20 of 361).

Relatedness estimation

The genetic relatedness between interacting cells (r) is a key parameter for social evolution. Relatedness can be very hard to estimate for natural populations of bacteria and other microbes, except in extreme cases where interactions take place in some physical structure such as a fruiting body or a filament (Flowers et al., 2010; Gilbert et al., 2012). Population genetic data allow relatedness to be estimated indirectly because the degree to which selection is relaxed, and greater polymorphism will be observed, depends on the relatedness between interacting cells. Consequently, if we assume that the patterns of polymorphism and divergence that we see are due to relatedness being <1, then we can work backwards from the polymorphism data to obtain an indirect estimate of relatedness (Belcher et al., 2022). This approach also requires assumptions about selection coefficients. As in previous studies, we have to assume that the magnitude of selection and the distribution of selection coefficients is the same on average for cooperative and private genes, which we cannot be sure is true. We also assume that we have identified the relevant private genes to compare our cooperative genes to. In *P. aeruginosa*, environmental isolates differ considerably in which genes are controlled by QS (Chugani et al., 2012), and the identity and number of genes can readily evolve under experimental evolution (Smalley et al., 2022). We also have to calculate relatedness across the whole genome because our calculation involves comparing the median polymorphism in cooperative genes to the median polymorphism in private genes. Although the correct measure of relatedness is that of the locus which controls the trait (West et al., 2006), this should correlate with whole-genome similarity. With these caveats in mind, by examining the polymorphism data from all genes, we estimated relatedness to be $r = .79$ (Supplementary S8).

An advantage of this indirect population genetic method for estimating relatedness is that it does not require knowledge about factors that would be hard or impossible to measure in natural populations. For example, the spatial details of how cooperative interactions play out, such as how far do public goods diffuse, and who benefits, as well as how much these vary in different environments, and the frequency with which different environments are encountered (Belcher et al., 2022; Nee et al., 2002). Indeed, cooperative traits in *B. subtilis* vary in the degree to which they are shared depending on whether groups are exhibiting sliding motility or growing in biofilms (Jautzus et al., 2022). In contrast, the indirect measure provided by population genetics represents an average of the different cooperative traits, over the different environments encountered, over evolutionary time. This population genetic approach represents an alternative or opposite way of looking at the data and so we should be careful not to overinterpret both at the same time—we can either test for relaxed selection to test for kin selection or we can look at the extent of relaxed selection to estimate relatedness.

Conclusions

We have found strong evidence of kin selection for cooperation in a natural population of *B. subtilis*. Our analyses controlled for possible confounding factors, such as expression rate, and eliminated alternative explanations for polymorphism and divergence, providing evidence that complements the numerous lab experiments demonstrating sociality in this species (Branda et al., 2001, 2004; Chai et al., 2008; Dragoš et al., 2018; Erskine et al., 2018; Kraigher et al., 2021; Lyons et al., 2016; Martin et al., 2020; Romero et al., 2010; Stefanic et al., 2015; Van Gestel et al., 2014). Taken together with a previous study, population genetic analyses have now provided evidence of kin selection for cooperation in both gram-positive (*B. subtilis*) and gram-negative (*P. aeruginosa*) bacteria (Belcher et al., 2022). These results suggest convergent, and potentially widespread, kin selection for cooperation, based on very different underlying mechanisms, across bacteria.

A possible complication with studying cooperation in bacteria is that the extent to which traits are cooperative, and their importance for fitness, can depend greatly on environmental conditions (Connelly et al., 2017; Kümmerli et al., 2009; Sexton & Schuster, 2017; West et al., 2012; Xavier et al., 2011). One advantage of the molecular population genetic approach that we have used is that it averages across different environments over evolutionary time. Consequently, rather than examining a specific environment, it provides an "average" answer. In the case of *B. subtilis*, for the traits that we have examined, we have found evidence of kin selection for cooperation, with an estimated average relatedness of $r = .79$.

Supplementary material

Supplementary material is available online at *Evolution Letters*.

Data availability

Data and code are available online at https://github.com/lauriebelch/bacillus.

Author contributions

L.J.B., M.G., and S.A.W. planned and designed the study. L.J.B. performed the bioinformatics, population genetic analysis, and

statistical analysis. A.E.D. and C.H. provided ideas and guidance throughout the project, particularly during the data analysis stage. L.J.B. and S.A.W. wrote the manuscript; all authors commented on and edited the manuscript.

Conflict of interest: The authors declare no conflict of interest.

Acknowledgments

We thank Ming Liu, Carolin Kobras, and Ákos Kovács for useful discussion and comments on the manuscript. This work was supported by the European Research Council (834164: L.J.B., A.E.D., and S.A.W.; SESE: M.G.).

References

Amanatidou, E., Matthews, A. C., Kuhlicke, U., Neu, T. R., McEvoy, J. P., & Raymond, B. (2019). Biofilms facilitate cheating and social exploitation of β-lactam resistance in *Escherichia coli*. *Biofilms and Microbiomes*, **5**(1), 1–10. https://doi.org/10.1038/s41522-019-0109-2

Andersen, S. B., Marvig, R. L., Molin, S., Krogh Johansen, H., & Griffin, A. S. (2015). Long-term social dynamics drive loss of function in pathogenic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(34), 10756–10761. https://doi.org/10.1073/pnas.1508324112

Andrews, S. (2010). *FastQC*. Babraham Bioinforma.

Arnaouteli, S., Bamford, N. C., Stanley-Wall, N. R., & Kovács, A. T. (2021). *Bacillus subtilis* biofilm formation and social interactions. *Nature Reviews Microbiology*, **19**(9), 600–614. https://doi.org/10.1038/s41579-021-00540-9

Arnaouteli, S., Ferreira, A. S., Schor, M., Morris, R. J., Bromley, K. M., Jo, J., Cortez, K. L., Sukhodub, T., Prescott, A. R., Dietrich, L. E. P., MacPhee, C. E., & Stanley-Wall, N. R. (2017). Bifunctionality of a biofilm matrix protein controlled by redox state. *Proceedings of the National Academy of Sciences of the United States of America*, **114**(30), E6184–E6191. https://doi.org/10.1073/pnas.1707687114

Azimi, S., Klementiev, A. D., Whiteley, M., & Diggle, S. P. (2020a). Bacterial quorum sensing during infection. *Annual Review of Microbiology*, **74**, 201–219.

Azimi, S., Roberts, A. E. L., Peng, S., Weitz, J. S., McNally, A., Brown, S. P., & Diggle, S. P. (2020b). Allelic polymorphism shapes community function in evolving *Pseudomonas aeruginosa* populations. *ISME Journal*, **15**, 1929–1942. https://doi.org/10.1038/s41396-020-0652-0

Belcher, L. J., Dewar, A. E., Ghoul, M., & West, S. A. (2022). Kin selection for cooperation in natural bacterial populations. *Proceedings of the National Academy of Sciences of the United States of America*, **119**: e2119070119. https://doi.org/10.1073/pnas.2119070119

Berka, R. M., Hahn, J., Albano, M., Draskovic, I., Persuh, M., Cui, X., Sloma, A., Widner, W., & Dubnau, D. (2002). Microarray analysis of the *Bacillus subtilis* K-state: Genome-wide expression changes dependent on ComK. *Molecular Microbiology*, **43**(5), 1331–1345. https://doi.org/10.1046/j.1365-2958.2002.02833.x

Branda, S. S., Chu, F., Kearns, D. B., Losick, R., & Kolter, R. (2006). A major protein component of the *Bacillus subtilis* biofilm matrix. *Molecular Microbiology*, **59**(4), 1229–1238. https://doi.org/10.1111/j.1365-2958.2005.05020.x

Branda, S. S., Gonzalez-Pastor, J., Ben-Yehuda, S., *et al.* (2001). Fruiting body formation by *Bacillus subtilis* Steven. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 11621–11626. https://www.pnas.org/doi/full/10.1073/pnas.191384198

Branda, S. S., González-Pastor, J. E., Dervyn, E., Ehrlich, S. D., Losick, R., & Kolter, R. (2004). Genes involved in formation of structured multicellular communities by *Bacillus subtilis*. *Journal of Bacteriology*, **186**(12), 3970–3979. https://doi.org/10.1128/JB.186.12.3970-3979.2004

Brito, P. H., Chevreux, B., Serra, C. R., Schyns, G., Henriques, A. O., & Pereira-Leal, J. B. (2018). Genetic competence drives genome diversity in *Bacillus subtilis*. *Genome Biology and Evolution*, **10**(1), 108–124. https://doi.org/10.1093/gbe/evx270

Broad Institute. (2019). *Picard toolkit*. Broad Institute, GitHub Repos.

Bucher, T., Keren-Paz, A., Hausser, J., Olender, T., Cytryn, E., & Kolodkin-Gal, I. (2019). An active β-lactamase is a part of an orchestrated cell wall stress resistance network of *Bacillus subtilis* and related rhizosphere species. *Environmental Microbiology*, **21**(3), 1068–1085. https://doi.org/10.1111/1462-2920.14526

Butaite, E., Baumgartner, M., Wyder, S., & Kümmerli, R. (2017). Siderophore cheating and cheating resistance shape competition for iron in soil and freshwater *Pseudomonas* communities. *Nature Communications*, **8**(1), 414. https://doi.org/10.1038/s41467-017-00509-4

Chai, Y., Chu, F., Kolter, R., & Losick, R. (2008). Bistability and biofilm formation in *Bacillus subtilis*. *Molecular Microbiology*, **67**(2), 254–263. https://doi.org/10.1111/j.1365-2958.2007.06040.x

Charif, D., & Lobry, J. R. (2007). SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In U. Bastolla, M. Porto, H. Roman, & M. Vendruscolo (Eds.), *Structural approaches to sequence evolution: Molecules, networks, populations*. Springer Verlag.

Chugani, S., Sik Kim, B., Phattarasukol, S., & Greenberg, E. P. (2012). Strain-dependent diversity in the Pseudomonas aeruginosa quorum-sensing regulon. *Proceedings of the National Academy of Sciences of the United States of America*, **109**: E2823–E2831. https://doi.org/10.1073/pnas.1214128109

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)*, pp. 80–92. https://doi.org/10.4161/fly.19695

Comella, N., & Grossman, A. D. (2005). Conservation of genes and processes controlled by the quorum response in bacteria: Characterization of genes controlled by the quorum-sensing transcription factor ComA in *Bacillus subtilis*. *Molecular Microbiology*, **57**(4), 1159–1174. https://doi.org/10.1111/j.1365-2958.2005.04749.x

Commichau, F. M., Pietack, N., & Stülke, J. (2013). Essential genes in *Bacillus subtilis*: A re-evaluation after ten years. *Molecular Biosystems*, **9**(6), 1068–1075. https://doi.org/10.1039/c3mb25595f

Connelly, B. D., Bruger, E. L., McKinley, P. K., & Waters, C. M. (2017). Resource abundance and the critical transition to cooperation. *Journal of Evolutionary Biology*, **30**(4), 750–761. https://doi.org/10.1111/jeb.13039

Cooper, G. A., Frost, H., Liu, M., & West, S. A. (2021). The consequences of group structure and efficiency benefits for the evolution of division of labour. *Elife*, **10:** e71968. https://doi.org/10.7554/ELIFE.71968

Cooper, G. A., & West, S. A. (2018). Division of labour and the evolution of extreme specialization. *Nature Ecology and Evolution*, **2**(7), 1161–1167. https://doi.org/10.1038/s41559-018-0564-9

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. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 20059–20064. https://doi.org/10.1073/pnas.1213344109

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li,

H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, **10**(2), 1–4. https://doi.org/10.1093/gigascience/giab008

de Oliveira, J. L., Morales, A. C., Stewart, B., Gruenheit, N., Engelmoer, J., Brown, S. B., de Brito, R. A., Hurst, L. D., Urrutia, A. O., Thompson, C. R L, & Wolf, J. B. (2019). Conditional expression explains molecular evolution of social genes in a microbe. *Nature Communications*, **10**(1), 3284. https://doi.org/10.1038/s41467-019-11237-2

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 and Evolution*, **5**, 1624–1636. https://doi.org/10.1038/s41559-021-01573-2

Diggle, S. P., Griffin, A. S., Campbell, G. S., & West, S. A. (2007). Cooperation and conflict in quorum-sensing bacterial populations. *Nature*, **450**(7168), 411–414. https://doi.org/10.1038/nature06279

Dragoš, A., Kiesewalter, H., Martin, M., Hsu, C. -Y., Hartmann, R., Wechsler, T., Eriksen, C., Brix, S., Drescher, K., Stanley-Wall, N., Kümmerli, R., & Kovács, A. T. (2018). Division of labor during biofilm matrix production. *Current Biology*, **28**(12), 1903–1913.e5.e5. https://doi.org/10.1016/j.cub.2018.04.046

Dugatkin, L. A., Perlin, M., Lucas, J. S., & Atlas, R. (2005). Group-beneficial traits, frequency-dependent selection and genotypic diversity: An antibiotic resistance paradigm. *Proceedings of the Royal Society B: Biological Sciences*, **272**, 79–83. https://doi.org/10.1098/rspb.2004.2916

Duncan, K. E., Ferguson, N., Kimura, K., Zhou, X., & Istock, C. A. (1994). Fine-scale genetic and phenotypic structure in natural populations of *Bacillus subtilis* and *Bacillus licheniformis*: Implications for bacterial evolution and speciation. *Evolution (N Y)*, **48**, 2002. https://doi.org/10.2307/2410523

Erskine, E., MacPhee, C. E., & Stanley-Wall, N. R. (2018). Functional amyloid and other protein fibers in the biofilm matrix. *Journal of Molecular Biology*, **430**(20), 3642–3656. https://doi.org/10.1016/j.jmb.2018.07.026

Ertekin, O., Kutnu, M., Taşkin, A. A., Demir, M., Karataş, A. Y., & Özcengiz, G. (2020). Analysis of a bac operon-silenced strain suggests pleiotropic effects of bacilysin in *Bacillus subtilis*. *Journal of Microbiology*, **58**(4), 297–313. https://doi.org/10.1007/s12275-020-9064-0

Fan, L., Bo, S., Chen, H., Ye, W., Kleinschmidt, K., Baumann, H. I, Imhoff, J. F., Kleine, M., & Cai, D. (2011). Genome sequence of *Bacillus subtilis* subsp. *spizizenii* gtP20b, isolated from the Indian ocean. *Journal of Bacteriology*, **193**(5), 1276–1277. https://doi.org/10.1128/JB.01351-10

Fisher, R. M., Cornwallis, C. K., & West, S. A. (2013). Group formation, relatedness, and the evolution of multicellularity. *Current Biology*, **23**(12), 1120–1125. https://doi.org/10.1016/j.cub.2013.05.004

Flowers, J. M., Li, S. I., Stathos, A., Saxer, G., Ostrowski, E. A., Queller, D. C, Strassmann, J. E., & Purugganan, M. D. (2010). Variation, sex, and social cooperation: Molecular population genetics of the social amoeba *Dictyostelium discoideum*. *PLoS Genetics*, **6**(7), e1001013. https://doi.org/10.1371/journal.pgen.1001013

Frost, I., Smith, W. P. J., Mitri, S., San Millan, A., Davit, Y. Osborne, J. M., Pitt-Francis, J. M., MacLean, R. C., & Foster, K. R. (2018). Cooperation, competition and antibiotic resistance in bacterial colonies. *ISME Journal*, **12**, 1582–1593. https://doi.org/10.1038/s41396-018-0090-4

Futo, M., Opašić, L., Koska, S., Čorak, N., Široki, T., Ravikumar, V., Thorsell, A., Lenuzzi, M., Kifer, D., Domazet-Lošo, M., Vlahoviček, K., Mijakovic, I., & Domazet-Lošo, T. (2021). Embryo-like features in developing *Bacillus subtilis* biofilms. *Molecular Biology and Evolution*, **38**(1), 31–47. https://doi.org/10.1093/molbev/msaa217

Garcia-Garcera, M., & Rocha, E. P. C. (2020). Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. *Nature Communications*, **11**, 1–11. https://doi.org/10.1038/s41467-020-14572-x

Ghoul, M., Andersen, S. B., & West, S. A. (2017). Sociomics: Using omic approaches to understand social evolution. *Trends in Genetics*, **33**(6), 408–419. https://doi.org/10.1016/j.tig.2017.03.009

Ghoul, M., Griffin, A. S., & West, S. A. (2014). Toward an evolutionary definition of cheating. *Evolution (N Y)*, **68**, 318–331. https://doi.org/10.1111/evo.12266

Gilbert, O. M., Foster, K. R., Mehdiabadi, N. J., Strassmann, J. E., & Queller, D. C. (2007). High relatedness maintains multicellular cooperation in a social amoeba by controlling cheater mutants. *Proceedings of the Royal Society B: Biological Sciences*, **104**, 8913–8917. https://doi.org/10.1073/pnas.0702723104

Gilbert, O. M., Strassmann, J. E., & Queller, D. C. (2012). High relatedness in a social amoeba: The role of kin-discriminatory segregation. *Proc R Soc B Biol Sci*, **279**, 2619–2624. https://doi.org/10.1098/rspb.2011.2514

González-Pastor, J. E. (2011). Cannibalism: a social behavior in sporulating *Bacillus subtilis*. *FEMS Microbiology Review*, **35**(3), 415–424. https://doi.org/10.1111/j.1574-6976.2010.00253.x

Griffin, A. S., West, S. A., & Buckling, A. (2004). Cooperation and competition in pathogenic bacteria. *Nature*, **430**(7003), 1024–1027. https://doi.org/10.1038/nature02744

Hahn M. W. (2018). *Molecular population genetics*. Oxford University Press.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, **7**(1), 17–52. https://doi.org/10.1016/0022-5193(64)90039-6

Harwood, C. R., & Kikuchi, Y. (2022). The ins and outs of *Bacillus* proteases: Activities, functions and commercial significance. *FEMS Microbiology Review*, **46**, 1–20. https://doi.org/10.1093/femsre/fuab046

Jautzus, T., van Gestel, J., & Kovács, A. T. (2022). Complex extracellular biology drives surface competition during colony expansion in *Bacillus subtilis*. *ISME Journal*, **16**, 2320–2328. https://doi.org/10.1038/s41396-022-01279-8

Kalamara, M., Abbott, J. C., Macphee, C. E., & Stanley-Wall, N. R. (2021). Biofilm hydrophobicity in environmental isolates of *Bacillus subtilis*. *Microbiol (United Kingdom)*, **167**, 863–878. https://doi.org/10.1099/mic.0.001082

Kalamara, M., Spacapan, M., Mandic-Mulec, I., & Stanley-Wall, N. R. (2018). Social behaviours by *Bacillus subtilis*: Quorum sensing, kin discrimination and beyond. *Molecular Microbiology*, **110**(6), 863–878. https://doi.org/10.1111/mmi.14127

Kobayashi, K. (2008). SlrR/SlrA controls the initiation of biofilm formation in *Bacillus subtilis*. *Molecular Microbiology*, **69**(6), 1399–1410. https://doi.org/10.1111/j.1365-2958.2008.06369.x

Kobayashi, K. (2021). Diverse LXG toxin and antitoxin systems specifically mediate intraspecies competition in *Bacillus subtilis* biofilms. *PLoS Genetics*, **17**(7), e1009682. https://doi.org/10.1371/journal.pgen.1009682

Kobayashi, K. (2007). Gradual activation of the response regulator DegU controls serial expression of genes for flagellum formation and biofilm formation in *Bacillus subtilis*. *Molecular Microbiology*, **66**(2), 395–409. https://doi.org/10.1111/j.1365-2958.2007.05923.x

Konkol, M. A., Blair, K. M., & Kearns, D. B. (2013). Plasmid-encoded ComI inhibits competence in the ancestral 3610 strain of *Bacillus subtilis*. *Journal of Bacteriology*, **195**(18), 4085–4093. https://doi.org/10.1128/JB.00696-13

Koo, B. M., Kritikos, G., Farelli, J. D., Todor, H., Tong, K., Kimsey, H., Wapinski, I., Galardini, M., Cabal, A., Peters, J. M., Hachmann, A. -B., Rudner, D. Z., Allen, K. N., Typas, A., & Gross, C. A. (2017). Construction and analysis of two genome-scale deletion libraries for *Bacillus subtilis*. *Cell Systems*, **4**(3), 291–305.e7. https://doi.org/10.1016/j.cels.2016.12.013

Kraigher, B., Butolen, M., Stefanic, P., & Mandic Mulec, I. (2021). Kin discrimination drives territorial exclusion during *Bacillus subtilis* swarming and restrains exploitation of surfactin. *ISME Journal*, **16**(3), 833–841. https://doi.org/10.1038/s41396-021-01124-4

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. https://doi.org/10.1098/rspb.2009.0861

Kümmerli, R., Santorelli, L. A., Granato, E. T., Dumas, Z., Dobay, A., Griffin, A. S., & West, S. A. (2015). Co-evolutionary dynamics between public good producers and cheats in the bacterium *Pseudomonas aeruginosa*. *Journal of Evolutionary Biology*, **28**(12), 2264–2274. https://doi.org/10.1111/jeb.12751

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Linksvayer, T. A., & Wade, M. J. (2009). Genes with social effects are expected to harbor more sequence variation within and between species. *Evolution (N Y)*, **63**, 1685–1696. https://doi.org/10.1111/j.1558-5646.2009.00670.x

Linksvayer, T. A., & Wade, M. J. (2016). Theoretical predictions for sociogenomic data: The effects of kin selection and sex-limited expression on the evolution of social insect genomes. *Frontiers in Ecology and Evolution*, **4**, 1–10. https://doi.org/10.3389/fevo.2016.00065

Liu, M., West, S. A., & Cooper, G. A. (2021). Relatedness and the evolution of mechanisms to divide labor in microorganisms. *Ecology and Evolution*, **11**(21), 14475–14489. https://doi.org/10.1002/ece3.8067

Logan N. A., & De Vos, P. (2015). *Bacillus*. In M.E. Trujillo, S. Dedysh, P. DeVos, B. Hedlund, P. Kämpfer, F. A. R., & W. B. W. (Eds.), *Bergey's manual of systematics of archaea and bacteria*. Wiley.

López, D., & Kolter, R. (2010). Extracellular signals that define distinct and coexisting cell fates in *Bacillus subtilis*. *FEMS Microbiology Review*, **34**(2), 134–149. https://doi.org/10.1111/j.1574-6976.2009.00199.x

Lyons, N. A., Kraigher, B., Stefanic, P., Mandic-Mulec, I., & Kolter, R. (2016). A combinatorial kin discrimination system in *Bacillus subtilis*. *Current Biology*, **26**, 1–10. https://doi.org/10.1016/j.cub.2016.01.032

Mariappan, A., Makarewicz, O., Chen, X. H., & Borriss, R. (2012). Two-component response regulator DegU controls the expression of bacilysin in plant-growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42. *Microb Physiol*, **22**(2), 114–125. https://doi.org/10.1159/000338804

Martin, M., Dragoš, A., Otto, S. B., Schäfer, D., Brix, S., Maróti, G., & Kovács, A. T. (2020). Cheaters shape the evolution of phenotypic heterogeneity in *Bacillus subtilis* biofilms. *ISME Journal*, **14**(9), 2302–2312. https://doi.org/10.1038/s41396-020-0685-4

McKenna, A., Hanna, M., Banks, E., *et al*. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1997–1303. https://doi.org/10.1101/gr.107524.110

McNally, L., Bernardy, E., Thomas, J., Kalziqi, A., Pentz, J., Brown, S. P., Hammer, B. K., Yunker, P. J., & Ratcliff, W. C. (2017). Killing by Type VI secretion drives genetic phase separation and correlates with increased cooperation. *Nature Communications*, **8**, 14371. https://doi.org/10.1038/ncomms14371

McNally, L., Viana, M., & Brown, S. P. (2014). Cooperative secretions facilitate host range expansion in bacteria. *Nature Communications*, **5**(1), 4594. https://doi.org/10.1038/ncomms5594

Meir, M., Harel, N., Miller, D., Gelbart, M., Eldar, A., Gophna, U., & Stern, A. (2020). Competition between social cheater viruses is driven by mechanistically different cheating strategies. *Science Advances*, **6**(34), eabb7990. https://doi.org/10.1126/sciadv.abb7990

Miethke, M., Klotz, O., Linne, U., May, J. J., Beckering, C.L., & Marahiel, M. A. (2006). Ferri-bacillibactin uptake and hydrolysis in *Bacillus subtilis*. *Molecular Microbiology*, **61**(6), 1413–1427. https://doi.org/10.1111/j.1365-2958.2006.05321.x

Molle, V., Fujita, M., Jensen, S. T., Eichenberger, P., González-Pastor, J. E., Liu, J. S., & Losick, R. (2003). The Spo0A regulon of *Bacillus subtilis*. *Molecular Microbiology*, **50**(5), 1683–1701. https://doi.org/10.1046/j.1365-2958.2003.03818.x

Müller, S., Strack, S. N., Hoefler, B. C., Straight, P. D., Kearns, D. B., & Kirby, J. R. (2014). Bacillaene and sporulation protect *Bacillus subtilis* from predation by *Myxococcus xanthus*. *Applied and Environment Microbiology*, **80**(18), 5603–5610. https://doi.org/10.1128/AEM.01621-14

Nakano, M. M., Xia, L., & Zuber, P. (1991). Transcription initiation region of the srfA operon, which is controlled by the comP–comA signal transduction system in *Bacillus subtilis*. *Journal of Bacteriology*, **173**, 5487. https://doi.org/10.1128/JB.173.17.5487-5493.1991

Nee, S., West, S. A., & Read, A. F. (2002). Inbreeding and parasite sex ratios. *Proceedings of the Royal Society B: Biological Sciences*, **269**, 755–760. https://doi.org/10.1098/rspb.2001.1938

Noguchi, N., Sasatsu, M., & Kono, M. (1993). Genetic mapping in *Bacillus subtilis* 168 of the aadK gene which encodes arninoglycoside 6-adenylyltransferase. **114**, 47–52.

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**(20), 1683–1691. https://doi.org/10.1016/j.cub.2009.08.056

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**(11), e49403–e49410. https://doi.org/10.1371/journal.pone.0049403

Noh, S., Geist, K. S., Tian, X., Strassmann, J. E., & Queller, D. C. (2018). Genetic signatures of microbial altruism and cheating in social amoebas in the wild. *Proceedings of the National Academy of Sciences of the United States of America*, **115**(12), 3096–3101. https://doi.org/10.1073/pnas.1720324115

Ostrowski, A., Mehert, A., Prescott, A., Kiley, T. B., & Stanley-Wall, N. R. (2011). YuaB functions synergistically with the exopolysaccharide and TasA amyloid fibers to allow biofilm formation by *Bacillus subtilis*. *Journal of Bacteriology*, **193**(18), 4821–4831. https://doi.org/10.1128/JB.00223-11

Ostrowski, E. A., Shen, Y., Tian, X., Sucgang, R., Jiang, H., Qu, J., Katoh-Kurasawa, M., Brock, D. A., Dinh, C., Lara-Garduno, F., Lee, S. L, Kovar, C. L, Dinh, H. H., Korchina, V., Jackson, L. R., Patil, S., Han, Y., Chaboub, L., Shaulsky, G., … David, C. (2015). Genomic signatures of cooperation and conflict in the social amoeba.

Current Biology, **25**(12), 1661–1665. https://doi.org/10.1016/j.cub.2015.04.059

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient Swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, **31**(7), 1929–1936. https://doi.org/10.1093/molbev/msu136

Pi, H., & Helmann, J. D. (2017). Sequential induction of Fur-regulated genes in response to iron limitation in *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America*, **114**(48), 12785–12790. https://doi.org/10.1073/pnas.1713008114

Pisithkul, T., Schroeder, J. W., Trujillo, E. A., Yeesin, P., Stevenson, D. M., Chaiamarit, T., Coon, J. J., Wang, J. D., & Amador-Noguez, D. (2019). Metabolic remodeling during biofilm development of *Bacillus subtilis*. *MBio*, **10**(3), e00623–e00719. https://doi.org/10.1128/mbio.00623-19

Pohl, S., Bhavsar, G., Hulme, J., Bloor, A. E., Misirli, G., Leckenby, M. W., Radford, D. S., Smith, W., Wipat, A., Williamson, E. D., Harwood, C. R., & Cranenburgh, R. M. (2013). Proteomic analysis of *Bacillus subtilis* strains engineered for improved production of heterologous proteins. *Proteomics*, **13**(22), 3298–3308. https://doi.org/10.1002/pmic.201300183

Poole, K. (2005). Aminoglycoside resistance in *Pseudomonas aeruginosa*. *Antimicrobial Agents and Chemotherapy*, **49**(2), 479–487. https://doi.org/10.1128/AAC.49.2.479-487.2005

Quentin, Y., Fichant, G., & Denizot, F. (1999). Inventory, assembly and analysis of *Bacillus subtilis* ABC transport systems. *Journal of Molecular Biology*, **287**(3), 467–484. https://doi.org/10.1006/jmbi.1999.2624

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 841–842. https://doi.org/10.1093/bioinformatics/btq033

R. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Roberts, M., & Cohan, F. (1995). Recombination and migration rates in natural populations of *Bacillus Subtilis* and *Bacillus mojavensis*. *Evolution (N Y)*, **49**, 1081–1094.

Romero, D., Aguilar, C., Losick, R., & Kolter, R. (2010). Amyloid fibers provide structural integrity to *Bacillus subtilis* biofilms. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(5), 2230–2234. https://doi.org/10.1073/pnas.0910560107

Romero, D., De Vicente, A., Rakotoaly, R. H., Dufour, S. E., Veening, J. -W., Arrebola, E., Cazorla, F. M., Kuipers, O. P., Paquot, M., & Pérez-García, A. (2007). The iturin and fengycin families of lipopeptides are key factors in antagonism of *Bacillus subtilis* toward *Podosphaera fusca*. *MPMI*, **20**(4), 430–440. https://doi.org/10.1094/mpmi-20-4-0430

Rutherford, S. T., & Bassler, B. L. (2012). Bacterial quorum sensing: Its role in virulence and possibilities for its control. *Cold Spring Harbor Perspectives in Medicine*, **2**, 1–25. https://doi.org/10.1101/cshperspect.a012427

Salverda, M. L. M., de Visser, J. A. G. M., & Barlow, M. (2010). Natural evolution of TEM-1 β-lactamase: Experimental reconstruction and clinical relevance. *FEMS Microbiology Review*, **34**(6), 1015–1036. https://doi.org/10.1111/j.1574-6976.2010.00222.x

Schuster, M., Sexton1, D. J., & Hense, B. A. (2017). Why quorum sensing controls private goods. *Frontiers in Microbiology*, **8**, 1–16. https://doi.org/10.3389/fmicb.2017.00885

Sexton, D. J., & Schuster, M. (2017). Nutrient limitation determines the fitness of cheaters in bacterial siderophore cooperation. *Nature Communications*, **81**(8), 1–8. https://doi.org/10.1038/s41467-017-00222-2

Simonet, C., & McNally, L. (2021). Kin selection explains the evolution of cooperation in the gut microbiota. *Proceedings of the National Academy of Sciences of the United States of America*, **118**, e2016046118. https://doi.org/10.1073/pnas.2016046118

Smalley, N. E., Schaefer, A. L., Asfahl, K. L., Perez, C., Greenberg, E. P., & Dandekar, A. A. (2022). Evolution of the quorum sensing regulon in cooperating populations of *Pseudomonas aeruginosa*. *MBio*, **13**(1), 1–19. https://doi.org/10.1128/MBIO.00161-22

Špacapan, M., Danevčič, T., Štefanic, P., Porter, M., Stanley-Wall, N. R., & Mandic-Mulec, I. (2020). The ComX quorum sensing peptide of *Bacillus subtilis* affects biofilm formation negatively and sporulation positively. *Microorg*, **8**, 1131. https://doi.org/10.3390/MICROORGANISMS8081131

Stefanic, P., Kraigher, B., Lyons, N. A., Kolter, R., and Mandic-Mulec, I. (2015). Kin discrimination between sympatric *Bacillus subtilis* isolates. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 14042–14047. https://doi.org/10.1073/pnas.1512671112

Strassmann, J. E., Gilbert, O. M., & Queller, D. C. (2011). Kin discrimination and cooperation in microbes. *Annual Review of Microbiology*, **65**, 349–367. https://doi.org/10.1146/annurev.micro.112408.134109

Strassmann, J. E., Zhu, Y., & Queller, D. C. (2000). Altruism and social cheating in the social amoeba *Dictyostelium discoideum*. *Nature*, **408**(6815), 965–967. https://doi.org/10.1038/35050087

Tam, N. K. M., Uyen, N. Q., Hong, H. A., Duc, Le H, Hoa, T. T., Serra, C. R., Henriques, A. O., & Cutting, S. M. (2006). The intestinal life cycle of *Bacillus subtilis* and close relatives. *Journal of Bacteriology*, **188**(7), 2692–2700. https://doi.org/10.1128/JB.188.7.2692-2700.2006

Torres, C., Galián, C., Freiberg, C., Fantino, J. -R., & Jault, J. -M. (2009). The YheI/YheH heterodimer from *Bacillus subtilis* is a multidrug ABC transporter. *Biochimica et Biophysica Acta*, **1788**, 1788615–1788622. https://doi.org/10.1016/J.BBAMEM.2008.12.012

Urrutia, A. O., & Hurst, L. D. (2001). Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics*, **159**(3), 1191–1199. https://doi.org/10.1093/genetics/159.3.1191

Urrutia, A. O., & Hurst, L. D. (2003). The signature of selection mediated by expression on human genes. *Genome Research*, **13**(10), 2260–2264. https://doi.org/10.1101/gr.641103

Van Dyken, J. D., Linksvayer, T. A., & Wade, M. J. (2011). Kin selection-mutation balance: A model for the origin, maintenance, and consequences of social cheating. *American Naturalist*, **177**(3), 288–300. https://doi.org/10.1086/658365

Van Dyken, J. D., & Wade, M. J. (2010). The genetic signature of conditional expression. *Genetics*, **184**(2), 557–570. https://doi.org/10.1534/genetics.109.110163

Van Dyken, J. D., & Wade, M. J. (2012). Detecting the molecular signature of social conflict: Theory and a test with bacterial quorum sensing genes. *American Naturalist*, **179**(4), 436–450. https://doi.org/10.1086/664609

Van Gestel, J., Weissing, F. J., Kuipers, O. P., & Kovács, A. T. (2014). Density of founder cells affects spatial pattern formation and cooperation in *Bacillus subtilis* biofilms. *ISME Journal*, **8**(10), 2069–2079. https://doi.org/10.1038/ismej.2014.52

Veening, J. W., Stewart, E. J., Berngruber, T. W., Taddei, F., Kuipers, O. P., & Hamoen, L. W. (2008). Bet-hedging and epigenetic inheritance in bacterial cell development. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(11), 4393–4398. https://doi.org/10.1073/pnas.0700463105

Velicer, G. J., Kroos, L., & Lenski, R. E. (2000). Developmental cheating in the social bacterium *Myxococcus xanthus*. *Nature*, **404**(6778), 598–601. https://doi.org/10.1038/35007066

Wang, Q., Wei, S., Silva, A. F., & Madsen, J. S. (2023). Cooperative antibiotic resistance facilitates horizontal gene transfer. *ISME Journal*, **17**, 846–854. https://doi.org/10.1038/s41396-023-01393-1

West, S. A., & Cooper, G. A. (2016). Division of labour in microorganisms: An evolutionary perspective. *Nature Reviews Microbiology*, **1411**(14), 716–723. https://doi.org/10.1038/nrmicro.2016.111

West, S. A., Griffin, A. S., Gardner, A., & Diggle, S. P. (2006). Social evolution theory for microorganisms. *Nature Reviews Microbiology*, **4**(8), 597–607. https://doi.org/10.1038/nrmicro1461

West, S. A., Winzer, K., Gardner, A., & Diggle, S. P. (2012). Quorum sensing and the confusion about diffusion. *Trends in Microbiology*, **20**(12), 586–594. https://doi.org/10.1016/j.tim.2012.09.004

Whiteley, M., Diggle, S. P., & Greenberg, E. P. (2017). Progress in and promise of bacterial quorum sensing research. *Nature*, **551**(7680), 313–320. https://doi.org/10.1038/nature24624

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag

Xavier, J. B., Kim, W., & Foster, K. R. (2011). A molecular mechanism that stabilizes cooperative secretions in *Pseudomonas aeruginosa*. *Molecular Microbiology*, **79**(1), 166–179. https://doi.org/10.1111/j.1365-2958.2010.07436.x

Yoshida, K. I., Fujita, Y., & Ehrlich, S. D. (1999). Three asparagine synthetase genes of *Bacillus subtilis*. *Journal of Bacteriology*, **181**(19), 6081–6091. https://doi.org/10.1128/jb.181.19.6081-6091.1999

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., & Brinkman, F. S. L. (2010). PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**(13), 1608–1615. https://doi.org/10.1093/bioinformatics/btq249

Zwart, M. P., Schenk, M. F., Hwang, S., Koopmanschap, B., de Lange, N., van de Pol, L., Nga, T. T. T., Szendro, I. G., Krug, J., & de Visser, J. A. G. M. (2018). Unraveling the causes of adaptive benefits of synonymous mutations in TEM-1 β-lactamase. *Heredity (Edinb)*, **121**(5), 406–421. https://doi.org/10.1038/s41437-018-0104-z

Bacterial lifestyle shapes pangenomes

Authors and affiliations

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

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

*Corresponding author, anna.dewar@biology.ox.ac.uk

Abstract

Pangenomes vary across bacteria. Some species have fluid pangenomes, with a high proportion of genes varying between individual genomes. Other species have less fluid pangenomes, with different genomes tending to contain the same genes. Two main hypotheses have been suggested to explain this variation: differences in species' bacterial lifestyle and effective population size. However, previous studies have not been able to test between these hypotheses, because the different features of lifestyle and effective population size are highly correlated with each other, and phylogenetically conserved, making it hard to disentangle their relative importance. We used a phylogeny-based path analysis, across 126 bacterial species, to tease apart the causal role of different factors. We found that pangenome fluidity was lower in: (i) host-associated compared with free-living species; (ii) host-associated species that are obligately dependent on a host, live inside cells, are more pathogenic and less motile. In contrast, we found no support for the competing hypothesis that larger effective population sizes lead to more fluid pangenomes. Effective population size appears to correlate with pangenome variation because it is also driven by bacterial lifestyle, rather than because of a causal relationship.

Main text

At the turn of the 21st century, a rapid increase in bacterial genome sequencing revealed something surprising: individuals of the same bacterial species often vary considerably in the set of genes they carry[1,2]. This led to the concept of 'pangenomes', which refers to all the genes that have been sequenced in a species[3–5]. While a pangenome is, by-definition, a species-level measure, the variation it captures is the product of individual-level processes. Gene gain, either by duplication or horizontal gene transfer, and differential gene loss together generate gene content differences among individual genomes of the same species, influencing the size and variability of a species' pangenome[4–10]. Furthermore, this gene gain plays a key role in the evolution of pathogenesis, facilitating the spread of antimicrobial resistance and the emergence of novel pathogens, while gene loss potentially facilitates the streamlining pathogen genomes[9,11–14].

Pangenome structure varies considerably across bacterial species. For example, 84% of genes (loci) sequenced in the pathogen *Chlamydia trachomatis* are present in all genomes of the species, compared with only 16% of genes sequenced in the environmental generalist and gut pathogen *Salmonella enterica*[3]. How can we explain this variation that has been observed across species in pangenome structure? One hypothesis is that this variation in pangenome structure reflects different bacterial lifestyles[3–5]. Species that live in more variable environments may have greater opportunities to acquire genes horizontally, or be selected to retain different genes in different environments. An alternate hypothesis is that variation in pangenome structure reflects the consequences of different effective population sizes[15–19]. Larger population sizes could lead to more open pangenomes through either allowing more neutral genes to be maintained in the population, or by allowing more effective selection for different genes in different environments.

However, the different factors invoked by these hypotheses are highly correlated, and so their relative importance is unclear: species that live in more variable environments also tend to have larger effective population sizes. In addition, many different aspects of lifestyle that could influence environmental variability and effective population size, such as living within eukaryote hosts and motility, are also correlated. Lifestyle correlates with genome size, and smaller genomes are likely to carry a higher proportion of 'core genes', which could constrain pangenome fluidity. Consequently, even when correlations with pangenome structure are observed, the relative importance of different factors and underlying causality are unclear. For example, if free-living bacteria have more variable genomes this could be because they encounter more variable environments, have increased motility, have greater effective population sizes, or because they have larger genomes. Or it could be some combination of these factors.

Another problem is that species and genomes can share characteristics through common descent, rather than through independent evolution[20,21]. This means that species cannot be considered independent data points, analogous to the problem of pseudoreplication in experimental studies[22]. In addition, the number and distribution of available genome sequences is not representative of bacteria in nature, both across and within species. Genome sequences are biased towards species that have a greater impact on humans[23]. Phylogenetically-controlled bioinformatic analyses are required to control for these potential problems of non-independence and bias, while also disentangling the role of different factors[24].

We addressed this problem with a phylogeny-based comparative genomics analysis across 126 species where there was data on pangenome structure. We used a combination of phylogenetic correlation and phylogenetic path-analysis to determine the relative importance of the different factors that have been hypothesised to shape the pangenome. These analyses provided several complementary approaches that allowed us to tease apart the underlying causality. We used a phylogeny-based approach to control for the fact that closely related species tend to be similar because they share traits by common descent and so cannot be treated as independent data points. This approach also controls for variation across taxa in the number of species for which data are available, and variation across species in the number of genomes sequenced. We first examined the extent to which pangenome variation can be explained by bacterial lifestyle. Then, we examined whether lifestyle shapes the pangenome directly, or indirectly via an influence on genome size and effective population size.

Bacterial lifestyle and pangenome fluidity

We measured pangenome structure by calculating 'pangenome fluidity', which is the average proportion of genes which are not shared between any two genomes of the same species. This measure has also been called 'genomic/genome fluidity'[25]. A higher value of pangenome fluidity corresponds to a species with a larger and more variable pangenome. Across the 126 species, pangenome fluidity varied from 0.01 in *Chlamydia muridarum* to 0.41 in *Pseudomonas fluorescens*, corresponding to between 99% and 59% of genes shared between an average pair of genomes, respectively (Figure 1a). We found that this measure of pangenome fluidity was highly correlated with other possible measures of pangenome structure and variability (S1; section 4).

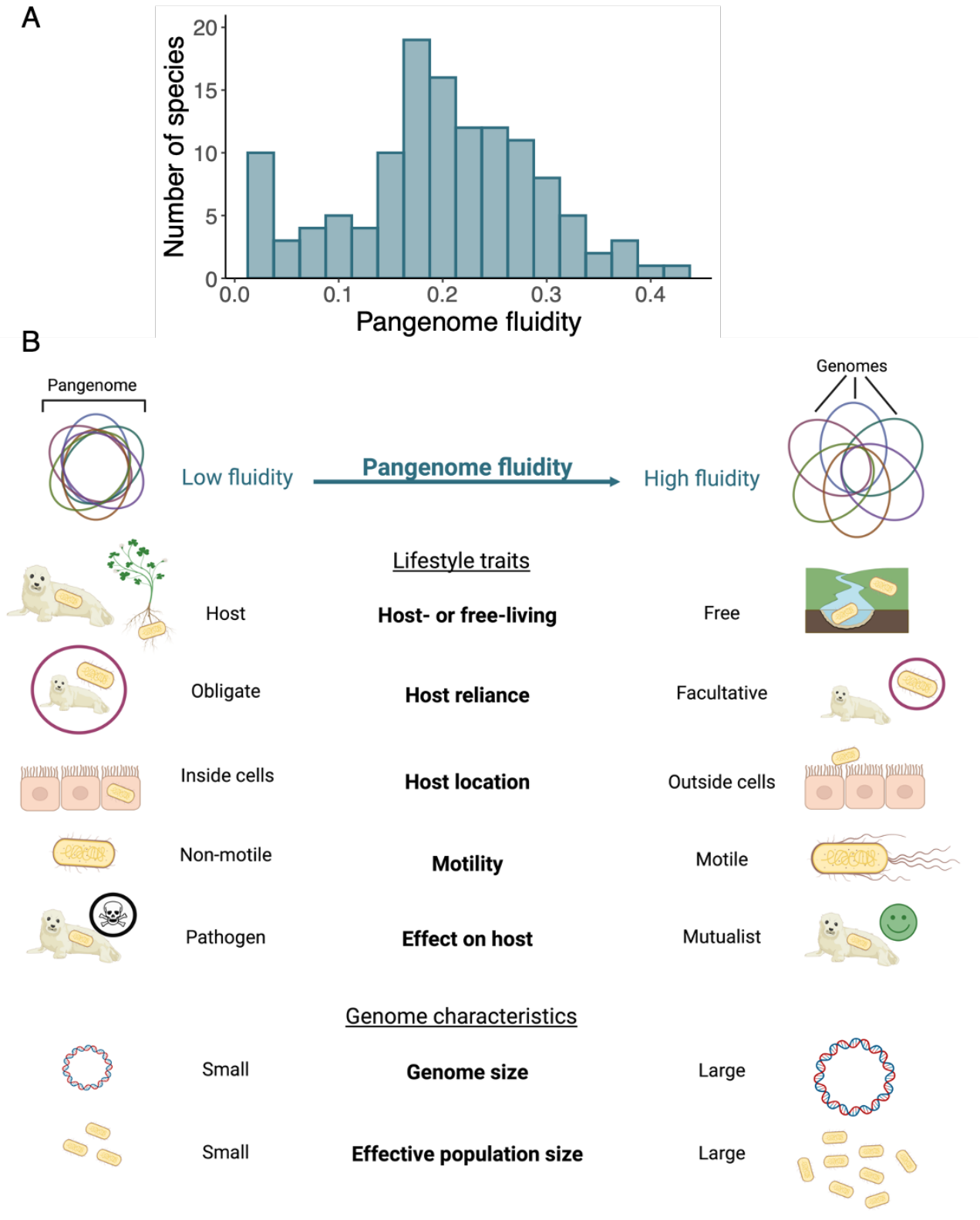


Figure 1. Hypothesised determinants of pangenome fluidity?

A. Histogram of pangenome fluidity across our 126 species, ranging from 0.012 to 0.41, with a median of 0.20. B. Conceptual figure illustrating hypothesised

determinants of pangenome fluidity. The intersecting ovals represent two illustrative pangenomes: each oval is a genome, and the overlap between ovals represents the proportion of genes shared between those genomes. The left illustrates low pangenome fluidity and the right high pangenome fluidity. Each of the rows below correspond to a factor which has been predicted to influence pangenome fluidity: whether a species is free-living or host-associated; four additional lifestyle traits that vary across host-associated species (host reliance, host location, effect on host and motility); genome size; and effective population size. For each factor, we illustrate how the biology of species varies between those which have lower compared to higher pangenome fluidity. Made with Biorender.com.

Our first aim was to examine how bacterial lifestyle influenced pangenome fluidity. There are multiple ways to define and estimate the lifestyle, ecology and/or environment of a bacterial species. The location in which species' 16s rRNA has been sequenced is often used as a source of information about the lifestyle of a species but this can be a potentially misleading. This is due to biases in the relative frequency that different environments are sequenced and potential discrepancies in how particular environments are defined. For example, out of 491 distinct 'biomes' defined in the MGnify metagenome database, 55 correspond to locations of the human body (https://www.ebi.ac.uk/). In contrast, only 12 biomes correspond to all fish (35,000 species).

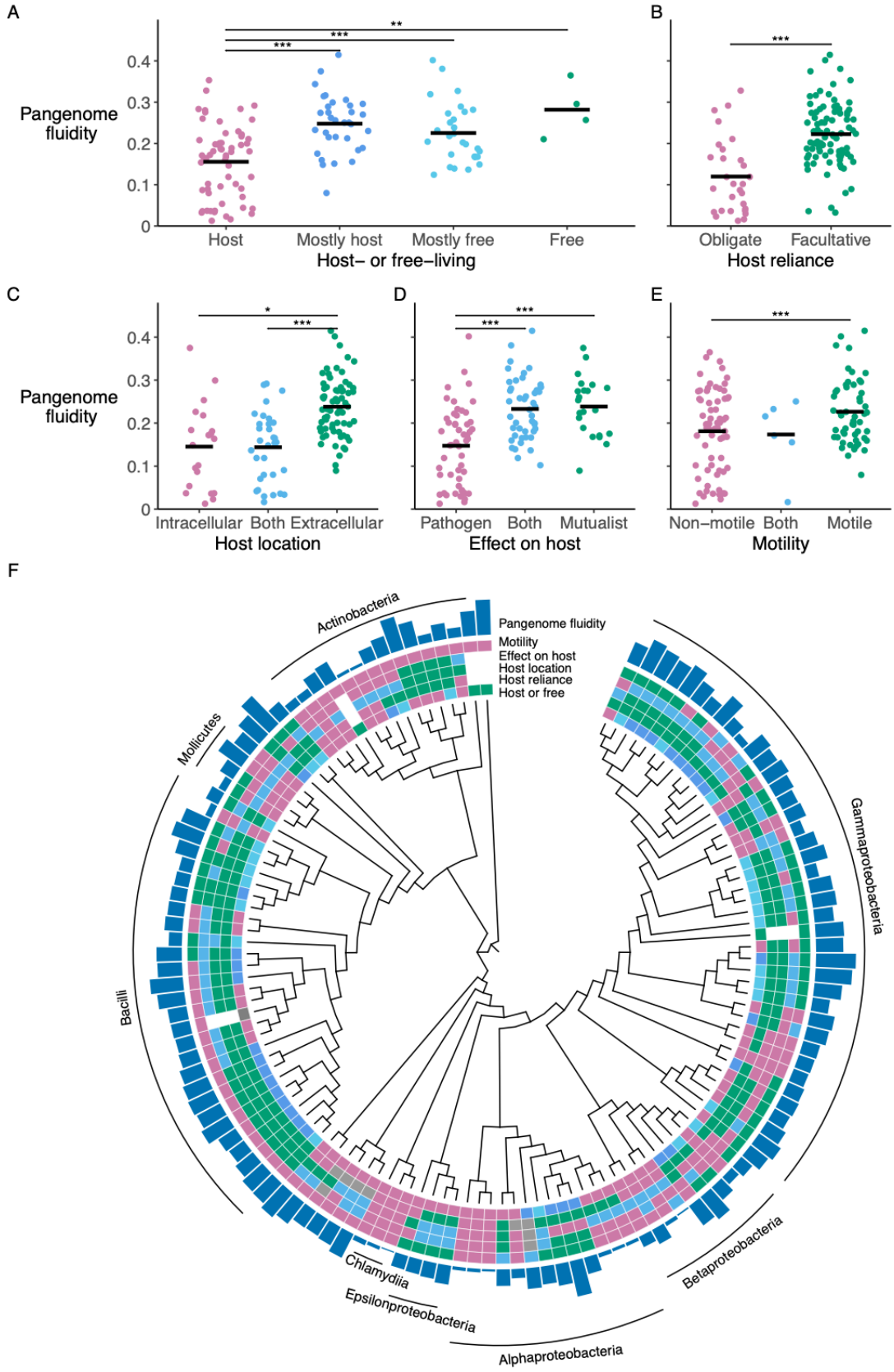
We instead classified species' lifestyle based on their biology. We drew a consensus from multiple sources including species descriptions, genome sequence metadata, published research articles and reviews, and several reference manuals and books on prokaryotes (S2). Using these, we categorised each of our 126 species based on five distinct lifestyle traits: (i)

host-associated or free-living; (ii) nature of host reliance (obligate or facultative); (iii) location within host(s) (intracellular, extracellular or both); (iv) effect on host(s) (pathogen, mutualist or both); (v) motility (non-motile, motile or both) (Figure 1b). We predicted that these would each influence the number and variability of environments encountered by a species, the potential to encounter novel genes, and therefore what kinds of genes bacteria gain via HGT and how frequently genes are lost (S1, sections 1 & 2).

We carried out two analyses. First, we compare free-living vs host-associated species. Second, we examined only host-associated species, examining the role of the other four lifestyle traits: host reliance; location in host; effect on host; and motility.

Free-living versus host-associated species

We found that free-living species had a higher genome fluidity than host-associated species (Figure 2a; MCMCglmm, R^2=0.139: 'Free' vs. 'host': pMCMC=0.004). Species which were sometimes free-living and sometimes host-associated had an intermediate fluidity, between that of strictly free-living or strictly host-associated species (Figure 2a; MCMCglmm, 'Mostly host' vs. 'host': pMCMC=<0.001; 'Mostly free' vs. 'host': pMCMC=<0.001). Overall, whether a species was free-living, host-associated, or both, explained around 14% of the total variance in genome fluidity across species. These results were robust to how we categorised 'host-associated' and 'free-living' (S1; section 1).


A
Pangenome fluidity
0
0.1
0.2
0.3
0.4
**
***
***
Host
Mostly host
Mostly free
Free
Host– or free–living
B
***
Obligate
Facultative
Host reliance
C
Pangenome fluidity
*
***
Intracellular
Both
Extracellular
Host location
D
***
***
Pathogen
Both
Mutualist
Effect on host
E
***
Non–motile
Both
Motile
Motility
F
Pangenome fluidity
Motility
Effect on host
Host location
Host reliance
Host or free
Actinobacteria
Mollicutes
Bacilli
Chlamydiia
Epsilonproteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria

149

Figure 2. Pangenome fluidity is correlated with multiple lifestyle traits and phylogeny.

Panels a-e shows how pangenome fluidity correlates with each of five lifestyle factors. Each dot is a species. Significance bars indicate results from MCMCglmm analyses (pMCMC: <0.05 '*'; <0.01 '**'; <0.001 '***'). (a) Host- or free-living (n=125). (b) Obligate or facultative (N=115, R^2=0.141; pMCMC<0.001). (c) Location in host: inside or outside cells (N=120, R^2=0.128; 'extracellular' vs. 'intracellular': pMCMC=0.046). (d) Effect on host: pathogen or mutualist (N=119, R^2=0.143; 'mutualist' vs. 'pathogen': pMCMC<0.001). (e) Motility (N=126, R^2=0.046; 'motile' vs. 'non-motile'; pMCMC=0.004). Panel (f) shows a tree of all 126 species, with rings corresponding to their lifestyle traits from panels a-e, and pangenome fluidity illustrated by the height of the outer blue bars. Grey squares correspond to where the lifestyle trait was unknown for a species, and white squares correspond to a within-host trait that was non-applicable for a free-living species. Clade annotations correspond to taxonomic classes which had at least three species' representatives in our dataset. A full list of species in our dataset is available in S1. Pink and green squares, corresponding to low and high lifestyle variability, respectively, are clustered together, both within and across species.

Variation across host-associated species

We then examined the variation in pangenome fluidity across host-associated species in further detail, using four discrete lifestyle traits: (i) nature of host reliance (obligate or facultative); (ii) location within host(s) (intracellular, extracellular or both); (iii) effect on host(s) (pathogen, mutualist or both); and (iv) motility (non-motile, motile or both) (Figure

1b). Information on all four of these traits was available for 115 species. Phylogenetic correlations showed that all four of these variables were significantly correlated with pangenome fluidity (Figure 2b-e, S1 (section 2.1-2.4). Species had more fluid (open) pangenomes when they were facultatively reliant on their hosts, extracellular, mutualists and motile (Figure 2). Overall, these four variables together explained 25.7% of the variation in pangenome fluidity across host-associated species (S1, section 2.6.1; MCMCglmm with four traits as fixed effects, R^2=0.257, N=115 species).

We also examined the overall explanatory power, considering all lifestyle variables, both host-associated versus free-living and variation within host-associated species. This means all the five lifestyle traits in Figures 1 and 2. Taken together, these variables were able to explain 29.9% of the variance in pangenome fluidity across species (S1, section 2.6.2: MCMCglmm with all five traits as fixed effects, R^2=0.299; N=119 species). This is a relatively large amount of variance explained for an across species comparative study, and approximately eight times the amount explained compared to the average of 3.6% for evolutionary and ecological studies (3.6%)[26,27].

Correlations between lifestyle traits

However, the different lifestyle traits were highly correlated (S1, section 2.6). Species which were obligately reliant on hosts were also more likely to live inside host cells, act as pathogens and be non-motile, while species which lived inside cells were more likely to act as pathogens (Phylogenetic regressions, S1, section 2.6.3; Host reliance vs: Host location, p=0.028; Effect on host, p=0.017; Motility, p=0.005; Host location vs. effect on host: p<0.001). Additionally, obligate host reliance was correlated with the evolution of living inside cells and with reduced motility (Pagel's correlated evolution model, S1, section 2.6.4;

Host association vs: Host location, $p=0.011$; Motility, $p<0.001$). This non-independence between lifestyle traits can be seen on a phylogenetic level in Figure 2f – pink squares, which correspond to the category for each lifestyle trait associated with lower pangenome fluidity, are clearly clustered, both across the tree and across the four lifestyle traits.

These correlations across the different lifestyle traits mean that phylogenetic correlations alone are unable to reveal the underlying causality. For example, intracellular lifestyle could lead to less fluid (closed) pangenomes, or this pattern could just be an artifact, with pangenome fluidity shaped by another factor that just happens to correlated with living inside cells. Alternatively, it is possible that intracellular life could influence pangenome fluidity indirectly, by influencing another lifestyle trait which then influences pangenome fluidity directly. Our point here is not to argue for a certain causal relationship, but instead to point out that phylogenetic correlations alone are unable to distinguish between different possibilities.

Causality and Phylogenetic Path Analysis

We compared the likelihood of alternative causal relationships for the traits in host-associated species with phylogenetic path analysis[28]. This method is based upon the theory of causal inference, which suggests that while correlation does not equal causation, correlation, if not due to chance, always implies an underlying causal structure[29]. Using a path analysis, one can compare support for multiple hypothesised causal models by constructing a set of phylogenetic linear models which must be supported for the model to not be rejected. For example, if A caused both B and C, a linear model could be constructed in which B would no longer correlate with C once A was taken into account. If many variables are included, linear

models can be constructed for each causal pathway included in a potential model of causation. Support for any models not rejected can then be compared.

Using this approach, we found the model with the highest support was one in which all four lifestyle traits had a direct causal influence on pangenome fluidity (Figure 3). A higher pangenome fluidity, and consequently a more open pangenome, was influenced by species: (i) having a facultative compared to an obligate host reliance; (ii) living outside compared to inside host cells; (iii) acting as a mutualist compared to a pathogen; (iv) being motile compared to being non-motile.

In addition, our path analysis found significant causal links between lifestyle traits, suggesting that the evolution of one trait might lead to the evolution of other traits, in a way that subsequently influenced pangenome fluidity (S1, section 2.7). For example, species which lived inside their hosts' cells were more likely to be obligately reliant on their hosts, which in turn led to less fluid pangenomes (Figure 3). The presence of these additional links between lifestyle traits meant that three of the four lifestyle traits also had an indirect influence on pangenome fluidity via their effect on at least one other trait (Figure 3). We were able to reject a simpler model which included no links between the four lifestyle traits (S1, section 2.7.1).

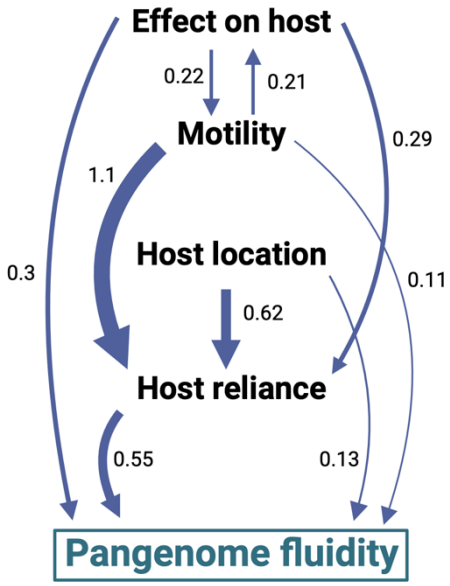


Figure 3. Phylogenetic path analysis suggests multiple lifestyle traits influence pangenome fluidity.

The average best model from a phylogenetic path analysis comparing how four within-host lifestyle traits influence both pangenome fluidity and each other. Arrows indicate the direction of causation and the width of arrows are proportional to the size of the standardised regression coefficients, which are printed next to each path. All four lifestyle traits have a direct influence on pangenome fluidity, and are influenced by or have influence on at least one other lifestyle trait.

These results were robust to additional analyses. We obtained the same patterns when using a number of alternative phylogenetic approaches, including Bayesian mixed effects models, phylogenetic regressions and correlated evolution models (S1, section 2). All four lifestyle traits had a direct influence on pangenome fluidity when we also defined each trait as binary (S1, section 2.8).

While our results demonstrate a role of bacterial lifestyle, they do not differentiate between the influence of adaptive or neutral genetic variation[18,30–32]. Species with more variable lifestyles and environments could have increased adaptive variation because different individuals would be under selection to adapt to different niches; this local adaptation could lead to patchy gene presence and more variation across the pangenome. Additionally, species with more variable lifestyle could have increased neutral variation because genes that individuals gain via horizontal gene transfer will be from a wider variety of bacteria, meaning genomes could vary in their gene content even in the absence of strong selection for carriage of those genes. Whether pangenomes are a product of adaptive or neutral variation could vary depending on the species, and both scenarios are not mutually exclusive.

To summarise, our results show that bacteria with more variable lifestyles also have more fluid (variable) pangenomes, with a larger fraction of gene differing between genomes. Free-living bacterial species have more fluid pangenomes than species which are associated with hosts. Whether a species is free-living or host-associated can explain 14% of the variance in pangenome fluidity across 125 species. Considering only host-associated species, those which are facultatively host-reliant, live outside host cells, act as mutualists and are motile have more fluid pangenomes. These features of lifestyle can together explain 26% of the variation in pangenome fluidity across the 115 host-associated species.

Genome characteristics and pangenome fluidity

Bacterial lifestyle could shape pangenome fluidity directly, or via other factors which are influenced by or correlate with lifestyle. One possibility is that bacterial lifestyle determines effective population size, which then determines pangenome fluidity. Species with larger effective population sizes tend to have more fluid pangenomes, and this has been

hypothesised to reflect causality[31]. Larger effective population sizes tend to have a higher level of neutral genetic diversity and a higher efficiency of natural selection, both of which could increase variation in gene content, and thus increase pangenome fluidity. Species with more variable lifestyle could have larger effective population sizes[3,5]. Consequently, bacterial lifestyle could influence pangenome fluidity just because it determines effective population size.

Another possibility is that bacterial lifestyle determines genome size, which then determines pangenome fluidity. Species which are obligately reliant on their hosts, particularly those which live inside cells, tend to have smaller genomes, as do species which more frequently act as pathogens[13,33]. Given that each species has a set of 'core' genes, which all individuals carry, a smaller genome could limit the number of genes which can vary between individuals of the same species, leading to less fluid genomes. Consequently, bacterial lifestyle could influence pangenome fluidity just because it determines genome size.

Consistent with previous analyses, we found that both genome size and effective population size were significantly positively correlated with pangenome fluidity[15–17] (Figure 4b & 4c, S1, section 3.1; MCMCglmm; (i) genome size, R^2=0.066; pMCMC<0.001; (ii) effective population size, R^2=0.0811; pMCMC<0.001). However, as stressed above, these correlations could result from either a causal relationship, or because they are associated with features of the lifestyle that shape pangenome fluidity. For example, obligate host reliance could lead to smaller effective population sizes which then could lead to lower pangenome fluidity; or obligate host reliance could lead to both smaller effective population sizes and lower pangenome fluidity. In this second case, the correlation between effective population size and

genome fluidity would just reflect their independent relationships with obligate host reliance, and not causality.

We used a phylogenetic path analysis to distinguish the most likely causal explanation. Estimates of effective population size were available for 75 of the 115 species where we had explored the role of bacterial lifestyle[15]. To characterise a species' lifestyle as a single variable we used several different methods, which all lead to the same conclusions. In our main analysis, we converted each lifestyle trait into a discrete numerical variable, and then summed these four values for each species. This simple additive lifestyle variable had a minimum value of 0, corresponding to an obligate, intracellular, pathogenic, non-motile species, predicted to have the least variable lifestyle, and a maximum value of 4, corresponding to a facultative, extracellular, mutualistic, motile species, which would have them most variable lifestyle (Figure 4c). We obtained the same conclusions with alternative methods for scoring lifestyle, including using Multiple Component Analyses to define a single lifestyle trait (S1, section 3.3).

We found that both lifestyle and genome size shape a species' pangenome fluidity, with little evidence of a direct influence of effective population size on pangenome fluidity (Figure 4d). Species with more variable lifestyles have more fluid pangenomes and larger genome sizes. Species with larger genomes had more fluid pangenomes. Species with more variable lifestyles had larger effective population sizes, but effective population size had no direct causal influence on pangenome fluidity. We found some evidence that effective population size might have a small influence on species' genome size, thereby indirectly influencing pangenome fluidity, but this effect was weak. Our results were robust to alternative analyses, including when we compared a set of very simple models which compared only the influence

of host-association and effective population size on pangenome fluidity and when we defined the single lifestyle trait using alternative methods (S1, section 3.4).

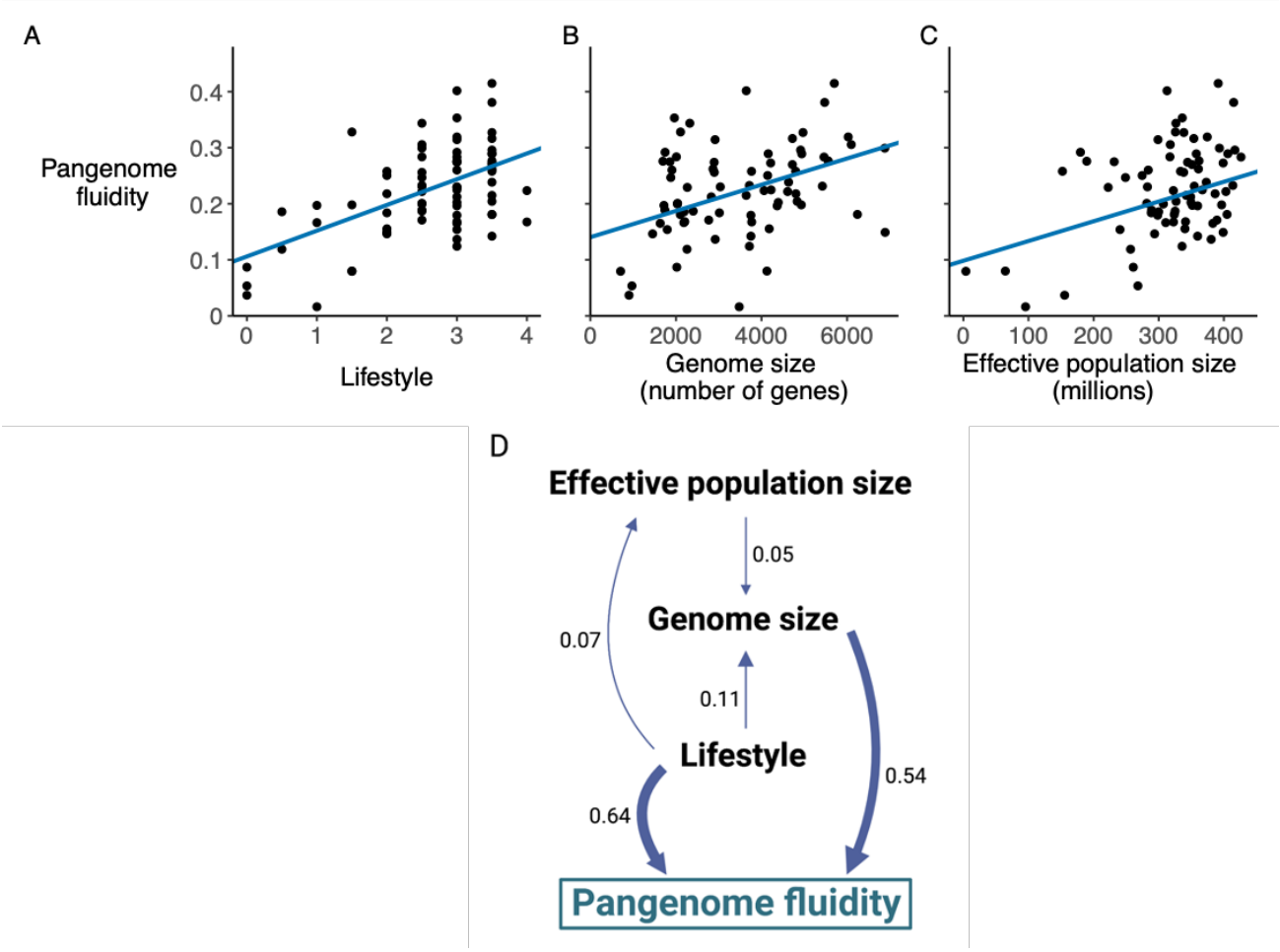


Figure 4. Lifestyle shapes pangenome fluidity, genome size and effective population size. Panels a-c show how pangenome fluidity correlates with three key factors. (a) Lifestyle, defined here by combining four traits into a single variable. A higher value corresponds to a more variable lifestyle. (b) Genome size, calculated by the mean number of genes across all genomes of a species. (c) Effective population size, units are millions (100 = 100 million individuals). Dots represent species and N=75 for all three panels. (d) The average best model from a phylogenetic path analysis comparing how lifestyle (a combination of the four traits in Figure 3), genome size and effective population size influence both pangenome fluidity and the evolution of each other. Arrows indicate the direction of causation and the

width of arrows are proportional to the size of the standardised regression coefficients, which are printed next to each path. Our best model was an average of four models which were similar in structure, and had similar support.

Conclusions

Our analyses suggest that bacterial lifestyle plays a key role in shaping the variation in pangenome structure across species. We consistently found that species which live in more variable environments have more fluid pangenomes (Figures 2-4). This influence of environmental variability could reflect adaptive and/or neutral processes. Species which live in more variable environments could both encounter a greater diversity of genes, and be under greater selection to gain and lose genes, depending upon the environment.

In contrast, we found less support for an influence of effective population size on pangenome fluidity. Species with larger effective population sizes had more fluid genomes, but this appears to be an artifact of how lifestyle influences both effective population size and pangenome fluidity (Figure 4). Effective population sizes were relatively large in most species that we studied, which is consistent with the conclusion that effective population size does not consistently increase or limit variation in pangenome fluidity across species[34–36](Figure 4c). We are not saying that effective population size could never influence pangenome fluidity, just that it does not significantly explain any of the variation in pangenome fluidity that has been observed across species which have been studied. Genome size did, however, influence pangenome fluidity (Figure 4d). Species with larger genomes had more fluid pangenomes. This is consistent with the hypothesis that smaller genomes contain a higher fraction of essential genes and so are constrained to have less variable genomes.

More generally, our results illustrate the insights that can be gained from applying behavioural and evolutionary ecology methods to the study of genome evolution in bacteria. We analysed the influence of bacterial lifestyle with phylogenetic methods that were developed to study behaviour and physical phenotypes. Previous work has shown the influence of symbiotic and pathogenic lifestyles on genome sizes and codon usage[13,33,37]. Our work builds upon this by showing that bacterial lifestyle not only influences the structure of genomes, but also how they vary across different individuals to produce the pangenome.

Methods

Collection of pangenome data

We collected bacterial pangenome data from panX in August 2021 (https://pangenome.org/). PanX is a web-based pangenome database that uses a pipeline to identify genes in genomes and then clusters them into orthologous groups. As of 2022, the database included species that had a minimum of 10 complete genomes in the RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/). We retrieved data for 126 bacterial pangenomes composed of 6221 genomes. PanX stores data in JSON format and we downloaded this using GNU Wget (https://www.gnu.org/software/wget/). We used the R package 'jsonlite' to convert the JSON data files into R objects. The pangenome data for all subsequent analyses included information on orthologous genes and the genomes in which they were found.

Pangenome fluidity

We used pangenome fluidity as our main measure of within-species genome variability – this measure has also been referred to as 'genome/genomic fluidity'[25].

$$Pangenome\ fluidity = \frac{2}{N\ (N-1)} \sum_{\substack{k,l=1\ldots N \\ k<l}} \frac{U_k + U_l}{M_k + M_l}$$

Where U_k and U_l are the number of genes found only in genomes *k* and *l*, M_k and M_l are the total number of genes found in *k* and *l*, and *N* is the number of genomes. The number of genes not shared between a random pair of genomes, k and l, is divided by the total genes found in both genomes, giving the proportion of genes which are not shared. This proportion is then averaged across all pairs of genomes within the species. This gives a value between 0 and 1, where values of 0.1 and 0.3 would mean that pairs of genomes did not share on average 10% and 30% of their genes respectively. For more information on this measure, please see Kislyuk *et al.* 2011, where it was first defined and where its relative benefits compared to other pangenome variability measures is assessed in detail[25].

Categorisation of species' lifestyles

We categorised the lifestyle of our species by defining distinct categorical lifestyle traits which we expected might influence opportunities for gene gain and selection for gene loss. We categorised lifestyle based on a consensus from multiple sources including species descriptions, genome sequence metadata, published research articles and reviews, and several reference manuals and books on prokaryotes including 'Bergey's Manual of Systematics of Archaea and Bacteria' (S2).

In the main analyses presented in this paper we first categorised species into those that were host- or free-living: 'host' if the species was host-associated for substantial part of lifecycle, and if rarely found in environment this is transient and/or a state to facilitate entry into a new host; 'primarily host' or 'primarily free' if the species was found in both hosts and in the

environment, but mostly lived in hosts or free, respectively; ‘free’ if the species lived independent of any host, with little or no evidence of host association.

Next, for those species that lived at least sometimes in hosts, and so were not assigned as ‘free’, we categorised species based on the following within-host traits: (i) Host reliance: ‘obligate’ if required host for survival and/or propagation; ‘facultative’ if the species did not require a host and can frequently survive outside its host(s); (ii) Host location: ‘intracellular’ if the species primarily lived inside its host’s cells; ‘extracellular’ if the species lived outside its host’s cells; ‘both’ if the species frequently lived both inside and outside its host’s cells; (iii) Effect on host: ‘pathogen’ if the species usually and consistently had a negative effect on host fitness; ‘mutualist’ if the species usually and consistently had a positive and/or neutral effect on host fitness; ‘both’ if the species frequently had a positive and negative effect on host fitness. We also categorised species into a lifestyle trait based on their motility: a species was ‘motile’ if flagella/other motility was always present; ‘non-motile’ if flagella/motility never present; ‘both’ if motility sometimes present and sometimes not. We used this motility lifestyle trait along with the three within-host traits described above to examine how multiple lifestyle traits may each correlate with and potentially cause variation in a species’ pangenome fluidity.

We also recorded additional information regarding the lifestyle, ecology and environment of our species. These were: host range (both in terms of ‘broad’ or ‘narrow’ and the taxonomic details of that range); primary and secondary locations within a host (i.e. ‘cell’, ‘root’, ‘gut’) ; whether the species lived inside or outside its host(s); host type (‘animal’, ‘plant’ or ‘both’); primary and secondary environment (i.e. ‘cell’, ‘soil’ ‘leaf’); whether the primary environment was host-associated or free-living; whether the species could undergo

sporulation; the number of broad environments the species has been recorded to live in. We hope that our comprehensive, literature-based approach to categorising species' lifestyle and ecology will prove useful for future comparative analyses on micro-organisms.

Genome size

We extracted the number of genes found in each of the 6221 genomes in our dataset, and then calculated the mean number of genes for all genomes in of our 126 species. Genome size ranged from 502 genes per average genome for the obligate, intracellular endosymbiont species *Buchnera aphidicola* to 6933 genes per average genome for the for the facultative plant mutualist *Rhizobium leguminosarum*.

Effective population size

To investigate a potential influence of effective population size on genome fluidity we used previously calculated estimates of effective population size, available for 77 of our species, which were based on dN/dS ratios in a set of universal genes[15]. Effective population size varied by several orders of magnitude, from 3.8 million in the obligate human pathogen *Mycoplasma pneumoniae* to 426 million in the broad-range plant pathogen *Pseudomonas syringae*. Of the 77 species for which effective population size estimates were available, there were 75 which we also had information on all four traits contributing to our single lifestyle variable in Figure 4a.

Phylogeny

To examine how genome fluidity and other factors have evolved across bacteria, and to control for any phylogenetic non-independence of these factors, we generated a phylogeny of the 126 species in our dataset, using methods as described in Dewar et al. 2021[22]. We used a

recently published maximum likelihood tree generated with a dataset of sixteen ribosomal proteins as the basis for our phylogeny[38]. We used the R package 'ape' to identify all branches that matched either a species or a genus in our dataset[39]. In cases where we had multiple species within a single genus, we used the R package 'phytools' to add these species as additional branches in the tree[40]. We used published phylogenies from the literature to add any within-genus clustering of species' branches (details and references in S3). We used this phylogeny for all our statistical analyses, as described below.

Correlations controlling for phylogenetic relationships

We used the R package MCMCglmm to examine correlations between genome fluidity and factors including lifestyle, genome size and effective population size[41]. We used this package to run Markov Chain Monte Carlo generalised linear mixed models, which are Bayesian statistical models used across the field of evolutionary biology for phylogenetic regressions. The evolutionary history of bacteria could mean that closely related species have more similar genome fluidity, regardless of other factors[20]. Consequently, we controlled for the phylogenetic relationships between species by setting the phylogeny as a random effect in our model. In the main text we have reported the pMCMC value (which for simplicity can be interpreted as one interprets a 'p-value') and the R^2 of the fixed effect for each model. Full results for all models can be found in S1, and code for all models is available at: https://github.com/AnnaEDewar/pangenome_lifestyle.git.

Phylogenetic path analysis

To disentangle causation from correlation across multiple factors which might influence genome fluidity, we used phylogenetic path analyses[28]. Briefly, this method compares support for multiple hypothesised causal models by constructing a set of phylogenetic linear

models which must be supported in order for the model to not be rejected. For example, if A caused both B and C, a linear model could be constructed in which B would no longer correlate with C once A was taken into account. If many variables are included, linear models can be constructed for each causal pathway included in a potential model of causation. By using a path analysis, one can identify any models of causation which can be rejected, and then compare support for any models not rejected. The phylogenetic component of the path analysis allowed us to address how lifestyle traits might have caused changes in genome fluidity across our species tree, again controlling for non-independence. We used the R package 'phylopath' to run our phylogenetic path analyses[42]. More details, including model support and comparisons, can be found in S1.

References

1. Welch, R. A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences* **99**, 17020–17024 (2002).
2. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial 'pan-genome'. *Proceedings of the National Academy of Sciences* **102**, 13950–13955 (2005).
3. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat Microbiol* **2**, 17040 (2017).
4. Domingo-Sananes, M. R. & McInerney, J. O. Mechanisms That Shape Microbial Pangenomes. *Trends in Microbiology* **29**, 493–503 (2021).
5. Brockhurst, M. A. *et al.* The Ecology and Evolution of Pangenomes. *Current Biology* **29**, R1094–R1103 (2019).

6. Puigbò, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y. I. & Koonin, E. V. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol* **12**, 66 (2014).
7. Cummins, E. A., Hall, R. J., McInerney, J. O. & McNally, A. Prokaryote pangenomes are dynamic entities. *Current Opinion in Microbiology* **66**, 73–78 (2022).
8. McInerney, J. O., Whelan, F. J., Domingo-Sananes, M. R., McNally, A. & O'Connell, M. J. Pangenomes and Selection: The Public Goods Hypothesis. in *The Pangenome: Diversity, Dynamics and Evolution of Genomes* (eds. Tettelin, H. & Medini, D.) (Springer, 2020).
9. Hall, J. P. J., Brockhurst, M. A. & Harrison, E. Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**, 20160424 (2017).
10. Hall, R. J. *et al.* Gene-gene relationships in an Escherichia coli accessory genome are linked to function and mobility. *Microbial Genomics* **7**, 000650 (2021).
11. MacLean, R. C. & San Millan, A. The evolution of antibiotic resistance. *Science* **365**, 1082–1083 (2019).
12. Sheppard, A. E. *et al.* Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene blaKPC. *Antimicrob Agents Chemother* **60**, 3767–3778 (2016).
13. Murray, G. G. R. *et al.* Genome Reduction Is Associated with Bacterial Pathogenicity across Different Scales of Temporal and Ecological Divergence. *Molecular Biology and Evolution* **38**, 1570–1579 (2021).
14. Merhej, V. & Raoult, D. Rickettsial evolution in the light of comparative genomics. *Biological Reviews* **86**, 379–405 (2011).
15. Bobay, L.-M. & Ochman, H. Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evolutionary Biology* **18**, 153 (2018).

16. Maistrenko, O. M. *et al.* Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *The ISME Journal* **14**, 1247–1259 (2020).
17. Andreani, N. A., Hesse, E. & Vos, M. Prokaryote genome fluidity is dependent on effective population size. *The ISME Journal* **11**, 1719–1721 (2017).
18. McInerney, J. O., McNally, A. & O'Connell, M. J. Reply to 'The population genetics of pangenomes'. *Nat Microbiol* **2**, 1575–1575 (2017).
19. McInerney, J. O. Prokaryotic Pangenomes Act as Evolving Ecosystems. *Mol Biol Evol* **40**, (2023).
20. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology*. (Oxford University Press, 1991).
21. Ridley, M. Why not to use species in comparative tests. *Journal of Theoretical Biology* **136**, 361–364 (1989).
22. Dewar, A. E. *et al.* Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range. *Nat Ecol Evol* **5**, 1624–1636 (2021).
23. Blackwell, G. A. *et al.* Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLOS Biology* **19**, e3001421 (2021).
24. Felsenstein, J. Phylogenies and the Comparative Method. *The American Naturalist* **125**, 1–15 (1985).
25. Kislyuk, A. O., Haegeman, B., Bergman, N. H. & Weitz, J. S. Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* **12**, 32 (2011).
26. Jennions, M. D. & Møller, A. P. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology* **14**, 438–445 (2003).

27. West, S. A., Shuker, D. M. & Sheldon, B. C. Sex-Ratio Adjustment When Relatives Interact: A Test of Constraints on Adaptation. *Evolution* **59**, 1211–1228 (2005).
28. Hardenberg, A. von & Gonzalez-Voyer, A. Disentangling Evolutionary Cause-Effect Relationships with Phylogenetic Confirmatory Path Analysis. *Evolution* **67**, 378–387 (2013).
29. Garamszegi, L. Z. *Chapter 8 in Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*. (Springer, 2014).
30. Vos, M. & Eyre-Walker, A. Are pangenomes adaptive or not? *Nat Microbiol* **2**, 1576–1576 (2017).
31. Shapiro, B. J. The population genetics of pangenomes. *Nat Microbiol* **2**, 1574–1574 (2017).
32. Douglas, G. M. & Shapiro, B. J. Genic Selection Within Prokaryotic Pangenomes. *Genome Biol Evol* **13**, (2021).
33. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* **10**, 13–26 (2011).
34. Lynch, M. Streamlining and Simplification of Microbial Genome Architecture. *Annual Review of Microbiology* **60**, 327–349 (2006).
35. Lynch, M. Phylogenetic divergence of cell biological features. *eLife* **7**, e34820 (2018).
36. Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* **17**, 704–714 (2016).
37. Fisher, R. M., Henry, L. M., Cornwallis, C. K., Kiers, E. T. & West, S. A. The evolution of host-symbiont dependence. *Nat Commun* **8**, 15973 (2017).
38. Hug, L. A. *et al.* A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).
39. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

590 40. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other
591 things). *Methods in Ecology and Evolution* **3**, 217–223 (2012).
592 41. Hadfield, J. D. MCMC Methods for Multi-Response Generalized Linear Mixed Models:
593 The MCMCglmm R Package. *Journal of Statistical Software* **33**, 1–22 (2010).
594 42. van der Bijl, W. phylopath: Easy phylogenetic path analysis in R. *PeerJ* **6**, e4718 (2018).
595

596 Acknowledgements

597 We thank Ashleigh Griffin, Craig MacLean, Thomas Scott, William Matlock and Zohar Katz
598 for helpful comments and suggestions on the manuscript. We also thank Juliet Turner, Louis
599 Bell-Roberts and Ming Liu for help and discussion regarding causal inference using path
600 analysis and other phylogenetic methods. **Funding**: We thank the European Research
601 Council for funding (834164 and SESE 647586). Figure 1b was created with BioRender.com.
602

603 Author Contributions

604 A.E.D. and S.A.W. conceived of the study. A.E.D and C.H. collected and curated lifestyle
605 and genomic data, respectively. A.E.D. completed the formal analysis, and A.E.D., C.H.,
606 L.J.B. and S.A.W. contributed to interpretation and visualisation of results. A.E.D. wrote the
607 original draft of the manuscript; all authors reviewed the manuscript and suggested edits and
608 revisions. S.A.W. supervised the project and acquired funding.
609

610 Competing Interest Declaration

611 The authors declare no competing interests.
612

613 Code Availability

614 Code for all analyses is available to download at:

https://github.com/AnnaEDewar/pangenome_lifestyle.git

Data Availability

Data will be to available download from Dryad upon publication, and is currently available to download at:

https://github.com/AnnaEDewar/pangenome_lifestyle.git

SOCfinder: a genomic tool for identifying cooperative genes in bacteria

Laurence J. Belcher[1a*], Anna E. Dewar[1a], Chunhui Hao[1], Zohar Katz[1], Melanie Ghoul[1b], & Stuart A. West[1b]

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

[a] Joint first author

[b] Joint last author

*Corresponding author

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

Abstract

Bacteria cooperate by working collaboratively to defend their colonies, share nutrients, and resist antibiotics. Nevertheless, our understanding of these remarkable behaviours primarily comes from studying a few well-characterized species. Consequently, there is a significant gap in our understanding of microbial cooperation, particularly in natural environments. To address this gap, we can use bioinformatic tools to identify cooperative traits and their underlying genes across diverse species. Existing tools address this challenge through two approaches. One approach is to identify genes that encode extracellular proteins, which can provide benefits to neighbouring cells. An alternative approach is to predict gene function using annotation tools. However, these tools have several limitations. Not all extracellular proteins are cooperative, and not all cooperative behaviours are controlled by extracellular proteins. Furthermore, existing functional annotation methods frequently miss known cooperative genes. Here, we introduce SOCfinder as a new tool to find cooperative genes in bacterial genomes. SOCfinder combines information from several methods, considering if a gene is likely to (1) code for an extracellular protein, (2) have a cooperative functional annotation, or (3) be part of the biosynthesis of a cooperative secondary metabolite. We use data on two extensively-studied species (*P. aeruginosa* & *B. subtilis*) to show that SOCfinder is better at finding known cooperative genes than existing tools. We also use theory from population genetics to identify a signature of kin selection in SOCfinder cooperative genes, which is lacking in genes identified by existing tools. SOCfinder opens up a number of exciting directions for future research, and is available to download from https://github.com/lauriebelch/SOCfinder.

Data Summary

All code and associated files are available at https://github.com/lauriebelch/SOCfinder.

Impact Statement

Bacteria cooperate by secreting many molecules outside the cell, where they can provide benefits to other cells. While we know much about how bacteria cooperate in the lab, we know much less about bacterial cooperation in nature. Is cooperation equally important in all species? Are all cooperations equally vulnerable to cheating? To answer these questions, we need a way of identifying cooperative genes across a wide range of genomes. Here, we provide such a method – which we name SOCfinder. SOCfinder allows users to find cooperative genes in any bacterial genome. SOCfinder opens up a number of exciting directions for future research. It will allow detailed studies of non-model species, as well as broad comparative studies across species. These studies will allow cooperation in the wild to be studied in new ways.

Introduction

The last twenty years has seen a revolution in our understanding of microbial sociality. We have moved from thinking that bacteria and other microbes live relatively independent unicellular lives, to discovering that they cooperate and communicate to perform a stunning array of social behaviours (1–6). This revolution has been largely driven by laboratory-based experiments in a small number of model species, especially *Pseudomonas aeruginosa, Escherichia coli,* and *Bacillus subtilis* (7–11) (Supplement S1). In contrast, we know little about social behaviours in natural populations outside of model species, and we don't know how the importance of cooperation varies across populations and species. For example, we know that division of labour underpins *Bacillus subtilis* cooperation (12, 13), but we don't know whether this is true in other species. We know that cheating is important in *Pseudomonas aeruginosa* iron-scavenging (6, 14–16), but we don't know why it doesn't appear to be important for the same behaviour in *Burkholderia cenocepacia* (17).

Relatively new genomic approaches offer several ways to study cooperative behaviours in natural populations. These genomic approaches rely on methodologies for identifying genes that control cooperative behaviours. One way to identify such 'cooperative genes' is to study the behaviour experimentally, and test whether it is cooperative (18, 19). While these experiments are relatively decisive, they are labour intensive and so not feasible for non-model organisms or large scale across species studies. An alternative approach is to use bioinformatic tools to identify genes for cooperative behaviours (28–33). Comparisons can then be made across species in order to examine how the number or proportion of cooperative genes varies, and if this can be explained by evolutionary theory (20–25). For example, do species where interacting individuals are more likely to be clonally related have more cooperative genes (20)? Alternatively, population genetic approaches can be used to test for 'signatures' (footprints) of selection for cooperation, to test if putatively cooperative behaviours really are cooperative in natural populations (26, 27). Other possibilities include comparisons between populations, between species with different lifestyles, or between genes that can undergo different rates of horizontal transfer (25).

The most commonly used bioinformatic tool is PSORTb, which can be used to identify genes that code for extracellular proteins (also known as 'extracellular genes') (28). These genes are likely to be cooperative because the proteins can diffuse away from the cell. Any effect of the

protein, such as breaking down food or neutralising antibiotics, can therefore provide benefits to the whole group of cells (21–25). Another tool is PANNZER, which predicts the function of any gene based on sequence similarity to known proteins (a process known as 'functional annotation') (29). Some functions, like 'extracellular biofilm matrix' are known to be cooperative (19).

However, there are several problems with these current methods. First, not all extracellular proteins are cooperative, and not all cooperative behaviours are controlled by extracellular proteins. Some important cooperative behaviours like siderophores are produced by many genes (30), none of which encode extracellular proteins. Second, these methods ignore information about a gene's location in the genome. Many secondary metabolite genes, including those for siderophores, are clustered together in the genome (30). Functional annotation might label the first and third gene in a cluster as cooperative, but miss the middle gene. Third, existing methods don't use contextual information on the quality and significance of functional annotation. This can make it difficult to compare across species, as there may be variation in the quality of annotations in different taxa. Fourth, existing methods can be slow to implement on bacterial genomes. Fifth, existing methods don't account for overlap between methods that are being combined, which can lead to mischaracterization or double-counting of genes.

To address these problems, we provide SOCfinder, a bioinformatics tool to find cooperative genes in bacterial genomes (Figure 1). SOCfinder combines information from several methods, considering if a gene is likely to: (1) code for an extracellular protein; (2) have a cooperative functional annotation; or (3) be part of the biosynthesis of a cooperative secondary metabolite. SOCfinder uses information on the quality and significance of database matches and annotations, and takes around 10 minutes to find cooperative genes in an average bacterial genome on a laptop. A separate list of cooperative genes from each tool is provided as an output, along with a total that avoids double-counting genes. SOCfinder version 1.0 is available as an easy-to-use command line tool, with tutorials, R scripts, and python scripts freely available at github.com/lauriebelch/SOCfinder.

We then examine the accuracy of SOCfinder, relative to other bioinformatic tools. We test the ability of different methods to identify genes for cooperation in two species: *Pseudomonas aeruginosa* and *Bacillus subtilis*. We focus on these two species because laboratory

experiments have been used to identify a number of cooperative behaviours, including the production of iron scavenging siderophores, quorum sensing and biofilm matrix proteins (7, 31–33). This allows us to test the accuracy and power of the different bioinformatic tools against direct experimental tests. We also test SOCfinder by applying it to >1000 bacterial genomes from 51 species, to see how cooperative gene repertoires vary among and between-species. Finally, we also carry out a population genetic analysis on the genes for cooperation identified by these different tools. This allows us to compare the power provided by the different methods for detecting signatures of selection.

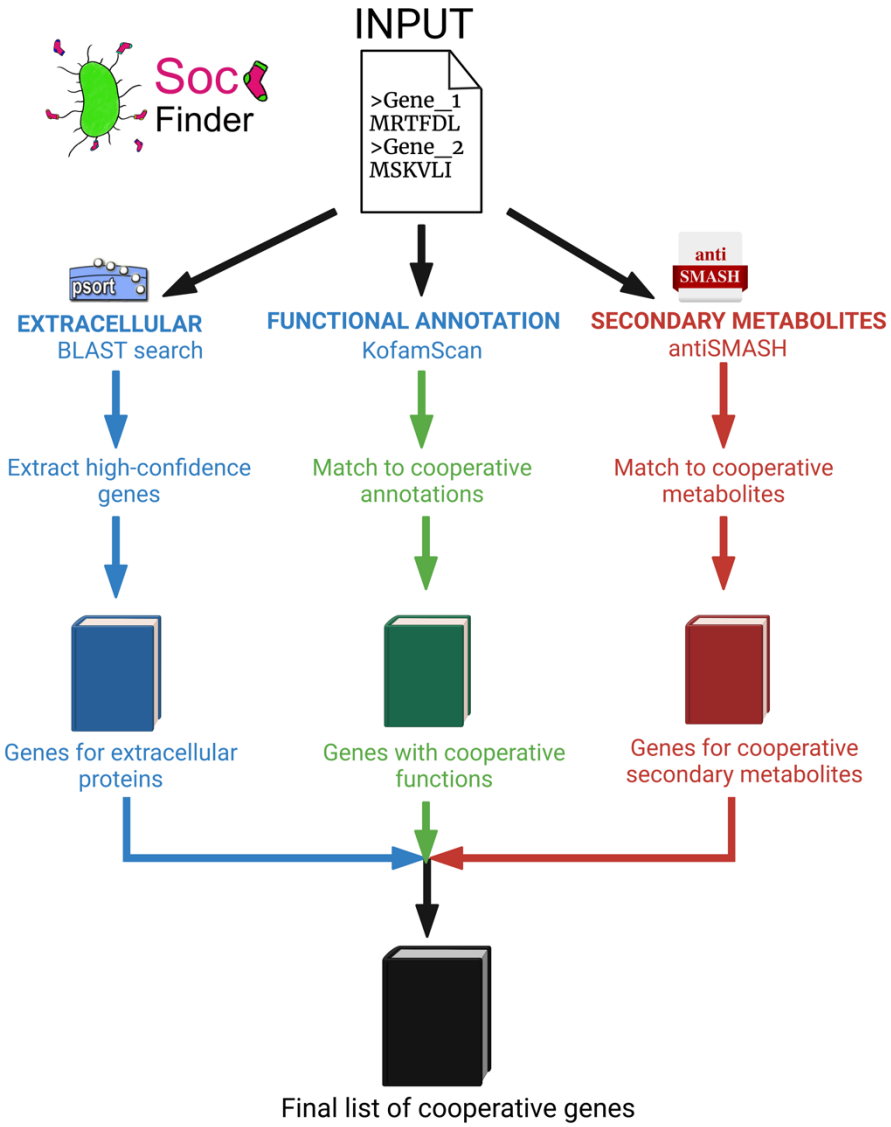


Figure 1: Overview of SOCfinder. We input a genome sequence, and cooperative genes are found based on three modules: (1) Extracellular genes. (2) Genes annotated with functions known to be cooperative, based on sequence similarity. (3) Genes for secondary metabolites that are known to be cooperative. We output a list of cooperative genes for each module, and a final list that combines all three.

Methods

Defining cooperative genes

Before describing our methodology for identifying cooperative genes, we need to define exactly what kind of genes we are looking for. A behaviour is social if it has fitness consequences for both the actor and the recipient (1, 34). Cooperation is a social behaviour where the recipient receives a benefit, and where the behaviour has been selectively favoured at least partially because of that benefit (35). This definition highlights the evolutionary problem of cooperation. Cooperators pay a cost by helping others, so are potentially vulnerable to cheats who benefit from cooperation without paying the cost (36, 37).

In animals, cooperative behaviours tend to be complex traits controlled by many genes, such as worker ants defending the colony (38), vampire bats sharing food (39), or meerkats helping others to rear young (40). As we move from meerkats to microbes the genetics is often simpler, with behaviours involving the production of molecules by one or few known genes. Bacteria produce a range of these molecules that provide benefits to the local group of cells (public goods), including iron scavenging molecules (41), enzymes to digest proteins (42), and toxins to eliminate competitors (43, 44).

We define a cooperative gene in bacteria as a gene which codes for a behaviour that provides a benefit to other cells, and has evolved at least partially because of this benefit. This can be tested for experimentally, by comparing the relative fitness of strains that do and don't perform a putatively cooperative behaviour both alone and in a mixed culture (1, 7). This contrasts with a 'private' gene, which has fitness consequences only for the individual expressing the gene (Figure 2).

A simple example is *lasB* in the opportunistic pathogen *P. aeruginosa*. This gene codes for the protein elastase, which is secreted outside the cell where it breaks down large structural proteins such as elastin and collagen (45). The digested products can then be taken up by the cell and used for nutrition (46). Lab experiments have compared the growth of the wildtype with a knockout mutant lacking *lasB*. The knockout strain grows slower than the wildtype when grown alone, but outcompetes the wildtype when both are grown together, because it can

exploit the elastase produced by the wildtype, while avoiding paying the costs (18, 31, 46, 47). The wildtype is therefore a cooperator and the knockout a 'cheat'.

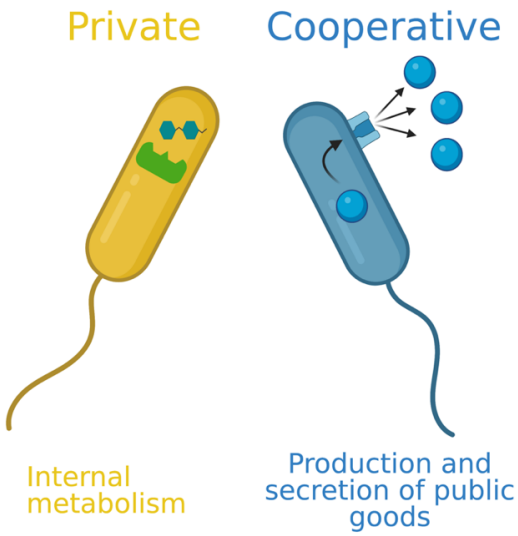


Figure 2: Categorisation of cooperative and private behaviours in bacteria. Cooperative behaviours are involved in the production and secretion of molecules that provide benefits that can be shared with other cells. Private behaviours give fitness benefits only to the individual expressing the gene.

Methods for identifying cooperative genes

In order to assess their validity and usefulness, we examined the methods used by researchers to identify cooperative genes, which vary from simply collating results from experimental work to genome-mining (Figure 3). We examine both the concept behind each method, and the tools used.

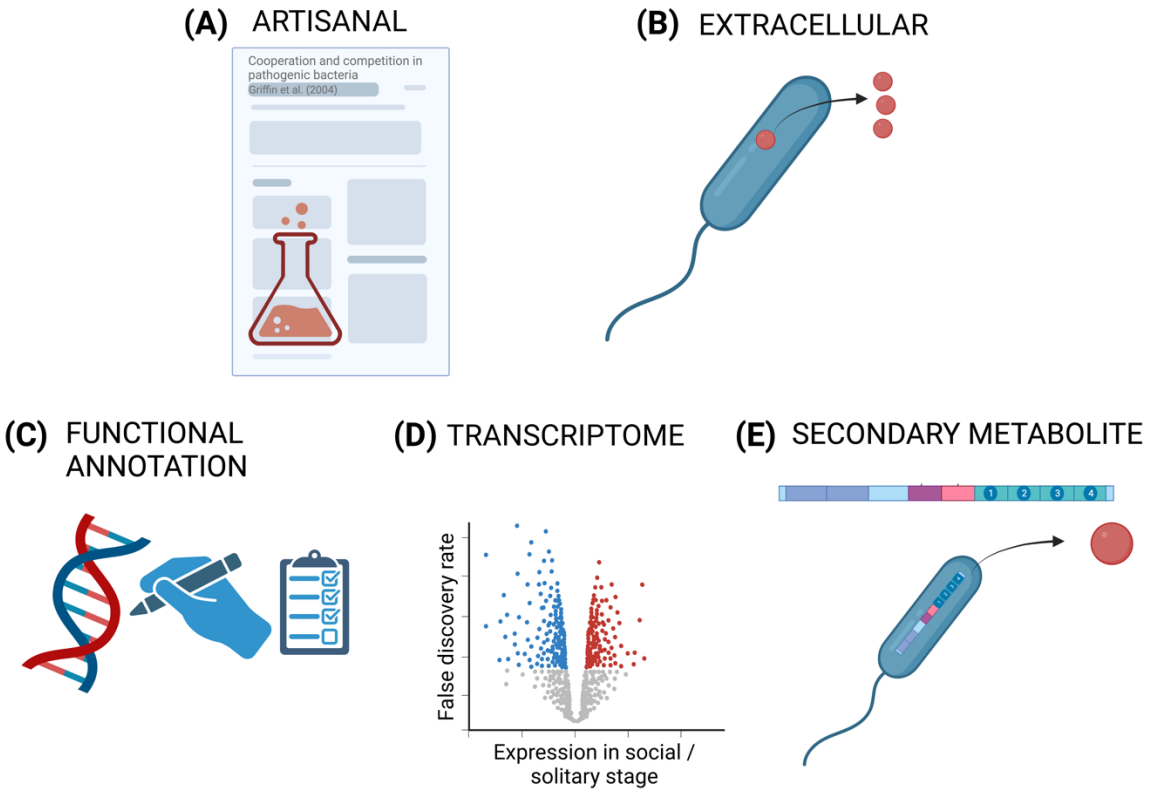


Figure 3: Principles of existing methods to find cooperative genes in genomes. We can look for: (A) Genes that have been shown to be cooperative in lab experiments (**artisan**). (B) **Extracellular** proteins that are secreted from the cell. (C) Genes that are **annotated** with functions that we know are cooperative, based on sequence similarity to proteins of known function. (D) Genes that are significantly upregulated when individuals are cooperating (**transcriptome**). (E) Genes for the biosynthesis of **secondary metabolites** that are known to be cooperative. A table of specific tools that can be used to find cooperative genes according to these principles is in Supplement S2.

Artisanal curation

In some species we can determine the genes for cooperative behaviours, based on upon the results of detailed laboratory experiments. If a species is sufficiently well-studied then we can identify cooperative genes using a literature search for papers conducting these experiments. For example, in *P. aeruginosa*, we could add the gene for elastase *lasB* to our list of cooperative genes based on experimental evidence (18, 31, 46, 47). This method, which we term the 'Artisanal' method, has been used in two studies on *P. aeruginosa* (26, 48), and one in *B. subtilis* (27).

Extracellular proteins

Many proteins produced by bacteria are extracellular (secreted outside the cell). Genes encoding extracellular proteins are likely to be cooperative because the proteins can diffuse away from the cell and provide a benefit to other cells in the population (22, 25). There are several tools to look for extracellular proteins. For instance, we can use simple BLAST searches to identify extracellular proteins based on similarity to proteins known from lab assays to be secreted, or more sophisticated tools like PSORTb, which also looks at the presence of known sequence motifs (28). This method is the most established for finding cooperative genes, having been used in a number of studies (20–25, 49). One recent study of 51 diverse bacterial species found that on average ~2% of genes code for extracellular proteins (25).

Gene functional annotation

Many gene functions are known to be cooperative, such as the production of extracellular matrix proteins in biofilms. Gene function can be predicted, based on homology and sequence similarity across species for the genes encoding for these behaviours (29, 50, 51). We can use our knowledge of cooperation from model species to make a list of cooperative functional annotation terms, using standardised systems such as gene ontology (GO) or KEGG orthology (KO). For example, Simonet & McNally curated a list of 118 cooperative gene ontology (GO) terms, that can be further split into five categories (secretion systems, siderophores, quorum sensing, biofilm, and antibiotic degradation) (20). They then used PANNZER (29) to predict the function of bacterial genes, which works by looking for homologous sequences which already have GO annotations. Other tools such as KOFAMscan (50) or eggnog-mapper (51) can also be used to predict gene function.

PanSort: A combined method

As well as looking at methods in isolation, we can combine the results of multiple methods. This kind of 'consensus' method might give better results than any one method in isolation, allowing multiple sources of information to be integrated. This innovative approach was used by Simonet & McNally, who combined a search for extracellular proteins with functional annotation of genes across human microbiome bacteria (20). They used PSORTb to count the number of genes coding for extracellular proteins. They then used PANNZER to annotate gene functions, with the top hit for each gene compared to a curated list of 'cooperative' gene ontology (GO) annotation terms. These two totals were then summed to give a total count of

the number of cooperative genes in a genome, which could potentially lead to double-counting. We refer to this method, which combined PSORTb and PANNZER, as 'PanSort'.

Transcriptomes

In some microbes there is a distinct social life stage, and we can find the genes controlling this switch in sociality by comparing gene expression between different stages of the life cycle. For example, the bacteria *Myxococcus xanthus* lives in swarms when food is abundant, but upon starvation forms a fruiting body where cells aggregate together. Some cells sacrifice themselves to cooperatively form the stalk that holds up the remaining cells as dispersing spores (52, 53). Similarly, the social amoeba *Dictyostelium discoideum* also has a division between solitary and social life stages (54, 55), with altruistic self-sacrifice in the social stage (56–58). Researchers have used transcriptome data to define cooperative genes as those that are highly expressed in the social stage of the lifecycle, but not in the solitary stage (59).

Secondary metabolites

Several known cooperative behaviours in bacteria are not simple extracellular proteins, but are complex molecules developed from several biosynthesis and modification steps. One example is iron-scavenging siderophores such as pyoverdine in *P. aeruginosa* (6, 7, 41). Whilst pyoverdine itself is secreted, none of the proteins controlling its production and export are. Instead, it is a secondary metabolite, defined as a compound that is not required for normal cell growth, but does provide some other benefit (60). We can use bioinformatic tools such as antiSMASH to look for genes that produce secondary metabolites in any genome sequence by looking at sequence similarity and the presence of certain conserved protein domains (61). This tool has been used to help find the cooperative genes that allow *Pseudomonas* and *Paenibacillus* strains to be cooperatively resistant to predation by amoebae when grown together, but susceptible when grown alone (62).

SOCfinder

Our new method SOCfinder draws on several of these methods. Given an assembled bacterial whole genome, SOCfinder runs three separate modules, and combines the predictions to produce a list of cooperative genes.

Module 1: Extracellular proteins

We designed our own method for finding genes that code for extracellular proteins, using the same principles as PSORTb (28). PSORTb gives a prediction of the localization of a protein across the cell, such as the periplasm of cytoplasmic membrane, whereas we only want to know if a protein is secreted or not. We therefore simplified and adapt the BLAST approached used by PSORTb to find gens for extracellular proteins, with some controls to check if a protein matches better to another location. This approach allows SOCfinder to be much quicker than PSORTb.

In our extracellular module, a BLAST search is performed against three out of four custom BLAST databases, based on the subcellular localisation of proteins as determined by PSORTb (Table 1). Depending on whether the species is gram negative or gram positive, either database 1 (gram-positive) or database 2 (gram-negative) is used, whereas databases 3 & 4 are always used (Figure 4).

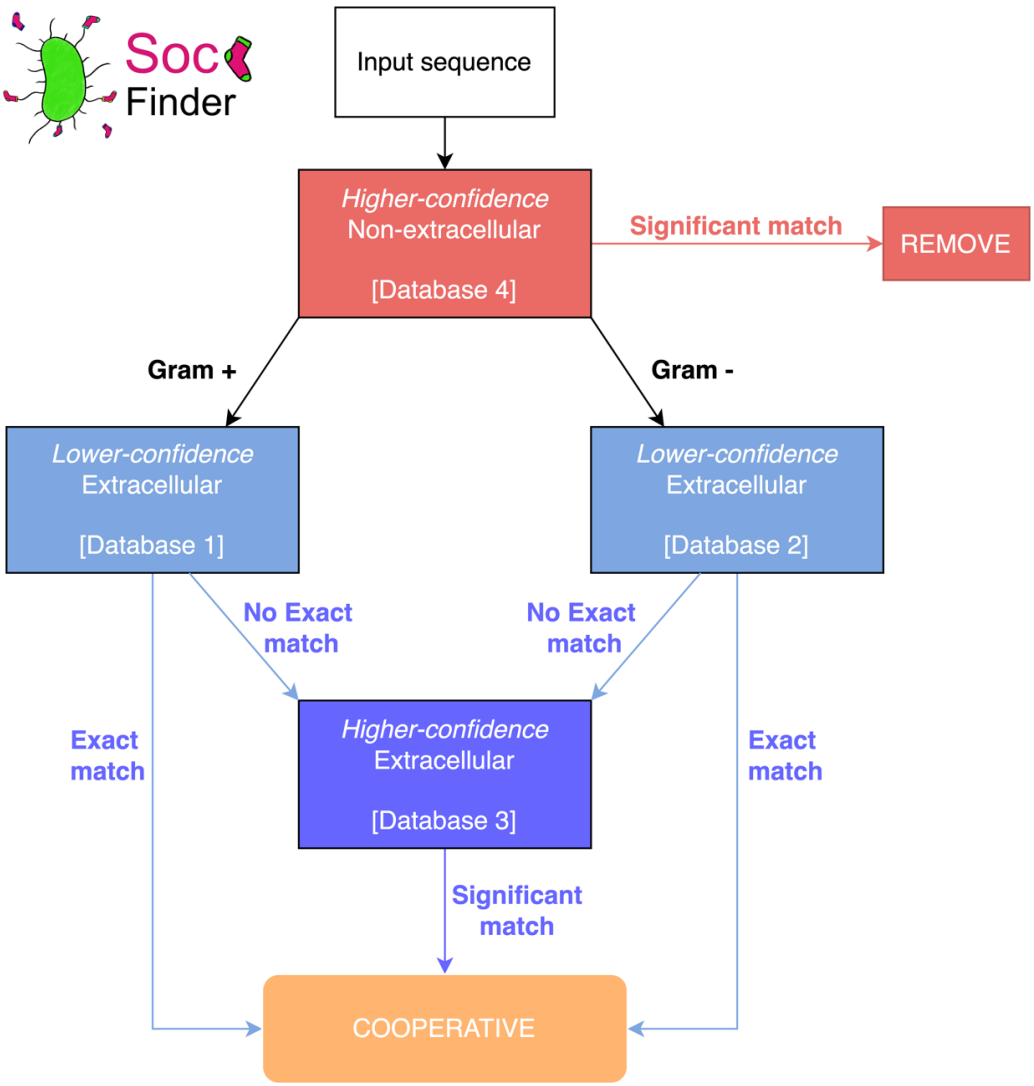


Figure 4: Flow diagram of the BLAST process for finding cooperative genes. Gram-positive and Gram-negative genomes are run against their own databases of high-confidence non-extracellular proteins (database 1 or 2), but both are run against the same databases of higher- and lower-confidence extracellular proteins (databases 3 & 4). Full information of the databases, as well as the definition of a significant match are found in tables 1-3.

316

317

318

319

320

321

Table 1: BLAST databases for finding extracellular genes. cPSORT refers to proteins that have been assigned a location based on the PSORTb algorithm. ePSORT refers to proteins with experimental evidence for their localisation.

Number	Name	Description	Proteins
1	cPSORTdbP extracellular	All the proteins from gram-positive bacteria that are computationally categorised as extracellular by PSORTb3	122,392
2	cPSORTdbN extracellular	All the proteins from gram-positive bacteria that are computationally categorised as extracellular by PSORTb3	156,076
3	ePSORTdb extracellular	All the proteins that are categorised as extracellular by the experimentally-derived version of PSORTb4	751
4	ePSORTdb non-extracellular	All the proteins that are categorised as not extracellular by the experimentally-derived version of PSORTb4	9,502

We first remove some genes from consideration in this module, based on strong evidence that they have a localization that isn't extracellular (Table 2). This step is important to avoid being too lenient with categorising genes as cooperative. Proteins will often have matches to proteins from multiple localizations, and within a species the same gene can be assigned to different localisations in different strains. We want to have a conservative approach, which is why we apply a stricter significance threshold to include a gene than we do to remove it from consideration, however this can be easily modified by users.

Table 2: Rules to remove a gene from consideration as cooperative (extracellular).

Test	Database	Action
Query protein has an exact match to a known non-extracellular protein	4	Remove from consideration
Query protein has a significant* match to a known non-extracellular protein	4	Remove from consideration

*e-value $<10^{-8}$, and query and database protein have the same length ±10%

We then test the remaining genes, and categorise genes as cooperative if it meets one or more of the conditions (Table 3). The databases can be found online at https://github.com/lauriebelch/SOCfinder and can be modified by users, and updated as tools such as PSORTb update their own databases to include more genes that have been experimentally or computationally categorised by location.

Table 3: Rules to categorise a gene as cooperative (extracellular).

Test	Database	Action
Query protein has an exact match to a high-confidence extracellular protein	1 or 2	List as cooperative
Query protein has an exact match to a known extracellular protein	3	List as cooperative
Query protein has a significant* match to a known extracellular protein	3	List as cooperative

*e-value $<10^{-20}$, and query and database protein have the same length ±20%

Module 2: Functional annotation

In the functional annotation module, we annotate the genome using KOFAMScan (50). The function of many bacterial genes is known, often because lab experiments have compared the phenotypes of a wildtype and a knock-out mutant that lacks the gene. For any query gene, we can assign it a function based on sequence similarity and machine-learning models that compare our query gene to proteins of known function. The number of matches and the closeness of each match can also be used to assign a score reflecting how confident we are that the query gene really does have that function. The full list of possible functional annotations is held by a database of KEGG orthology (KO) terms, each of which corresponds to a given function (63).

KOFAMScan annotates each protein with any matching KO terms, and each annotation is also given a score as well as an e-value which represents the number of hits it would expect to see by chance for that gene (50). KOFAMScan combines this information to determine whether a given annotation meets its threshold for significance. We can then categorise a gene as cooperative if it has a significant annotation for a KEGG orthology term that is cooperative.

To do this, we have created a curated list of cooperative KO terms, generated using a search of all KO terms for keywords corresponding to known cooperative behaviours in bacteria, followed by manual curation to remove KO terms that aren't likely to be cooperative. The full list of 321 cooperative KO terms is available at https://github.com/lauriebelch/SOCfinder/. Some examples include "exopolysaccharide biosynthesis", "beta lactamase", and "pyochelin biosynthesis protein", and they can be split into nine distinct categories including "siderophore", "biofilm formation", and "quorum sensing" (Table 4). For species where we know the full set of genes controlled by quorum sensing, we can use this method to separate cooperative from private quorum sensing genes. Cooperative genes are those highlighted by SOCfinder, and private genes are those not highlighted by SOCfinder. Similar to the extracellular module, we again take a conservative approach. For example, we exclude Type VI secretion systems, which are possibly social (64). However, the user can freely alter this list based on their own criteria.

Table 4: Categories of cooperative genes captured by functional annotation

Category	Description	Number of KO annotations
Beta-lactamase	Secreted enzymes that provide resistance against beta-lactam antibiotics	63
Biofilm formation	Genes that cause cells to collectively assemble in biofilms	46
Exopolysaccharide	Secreted molecules that form the main part of the biofilm matrix in many species	56
Extracellular matrix	Secreted molecules that form the biofilm matrix	8
Quorum sensing	Genes that regulate or are regulated by quorum sensing, where gene expression changes in response to population density	88
Rhamnolipid	Secreted biosurfactants that allow bacteria to collectively move and disperse over surfaces	3
Siderophore	Secreted molecules that bind to iron, allowing bacteria to scavenge iron from their hosts	22
Type II secretion	Genes secreted by the Type II secretion system used by many gram-negative bacteria to secrete exoproteins into the extracellular environment	16
Type IV pili	Genes for Type IV pili, which are used for collective "twitching motility"	18

Module 3: Secondary Metabolites

In the secondary metabolites module, we use antiSMASH (61) to find gene clusters that produce secondary metabolites. The aim here is to ensure that we can capture the entire region for complex social behaviours like iron-scavenging siderophores, where each gene codes for an intracellular protein, but the final product is secreted extracellularly. Functional annotation approaches often capture some, but not all, of these genes. We filter the antiSMASH output to remove all genes which have NA for their 'type' (e.g. core biosynthesis, transport, regulation), and then include a gene as cooperative if it matches our custom list of a small number of known social secondary metabolites. Our list includes beta-lactamases and metallophores such as siderophores, which allow bacteria to obtain iron and other metal ions from their hosts (41) (available at https://github.com/lauriebelch/SOCfinder/). Again, this is a conservative approach, but users can easily adjust the list to include other types of secondary metabolite, or as tools such as antiSMASH update their own categorisation.

One of the main strengths of SOCfinder is that it uses three different modules, which tend to capture separate genes. We control for any issues of double-counting by always outputting a final list of cooperative genes that avoids this, whilst still allowing flexibility by outputting separate lists for each module

Molecular Population Genetics

We followed the approach used in our previous research of analysing signatures of selection on genes whose expression is controlled by quorum-sensing (26, 27). Population genetic theory predicts that, in non-clonal populations (genetic relatedness $r<1$) that traits favoured by kin selection for cooperation will exhibit increased polymorphism and divergence, relative to traits that provide private benefits (65–69). When comparing putatively cooperative and private traits it is useful to compare traits which are likely to be expressed at similar rates (26, 27). We controlled for expression rates by examining genes controlled by the quorum sensing network. We use published datasets on which genes are controlled by quorum sensing in two species: *P. aeruginosa* and *B. subtilis* (70–73). Within quorum-sensing controlled genes, we assign a gene as 'cooperative' if it is found by whichever cooperative method we are testing (SOCfinder, PSORTb, or PanSort). We assign all other quorum-sensing controlled gene as 'private'.

To analyse a given population genetic measure, we compare three groups of genes: (1) cooperative quorum sensing genes; (2) private quorum sensing genes; and (3) background genes, which are those encoding proteins that localize to the cytoplasm. This set of background genes is least likely to have a cooperative function, and acts as another 'private genes' comparison.

Results

A test of SOCfinder on 51 species

We first tested our method by applying it to 1,301 bacterial genomes from 51 species that were used in a recent study on whether horizontal gene transfer can favour cooperation (25). This allowed us to look at how the number of cooperative genes varies both within- and between species. We found substantial variation across species in the proportion of a genome that is dedicated to cooperative genes, with an average of 2.8% (Figure 5). At one end of the scale, with only 1.2% of its genome dedicated to cooperation is *Buchnera aphidicola,* a symbiont that lives inside aphids (74). At the other end of the scale, with 5.3% its genome dedicated to cooperation is *Chlamydia trachomatis*, an obligate intracellular pathogen (75). Both species have tiny genomes (<1000 proteins), but very different lifestyles. *B. aphidicola* is vertically transmitted and synthesizes amino acids for its host (76). Our estimate here for cooperative genes in *B. aphidicola* is based upon cooperation between bacterial cells, and not cooperative behaviours that it performs to aid its aphid host. However, the search terms in SOCfinder could be expanded to also look at genes for such mutualistic cooperation. *C. trachomatis* has to enter cells, scavenge for nutrients, and fight a hostile immune system – all of which allow lots of opportunity for cooperation (77). Our results also suggest that there can be considerable variation within some species. For example, in *Escherichia coli*, the percentage of cooperation genes varies from 2.3-3.3%, with a median of 2.7%.

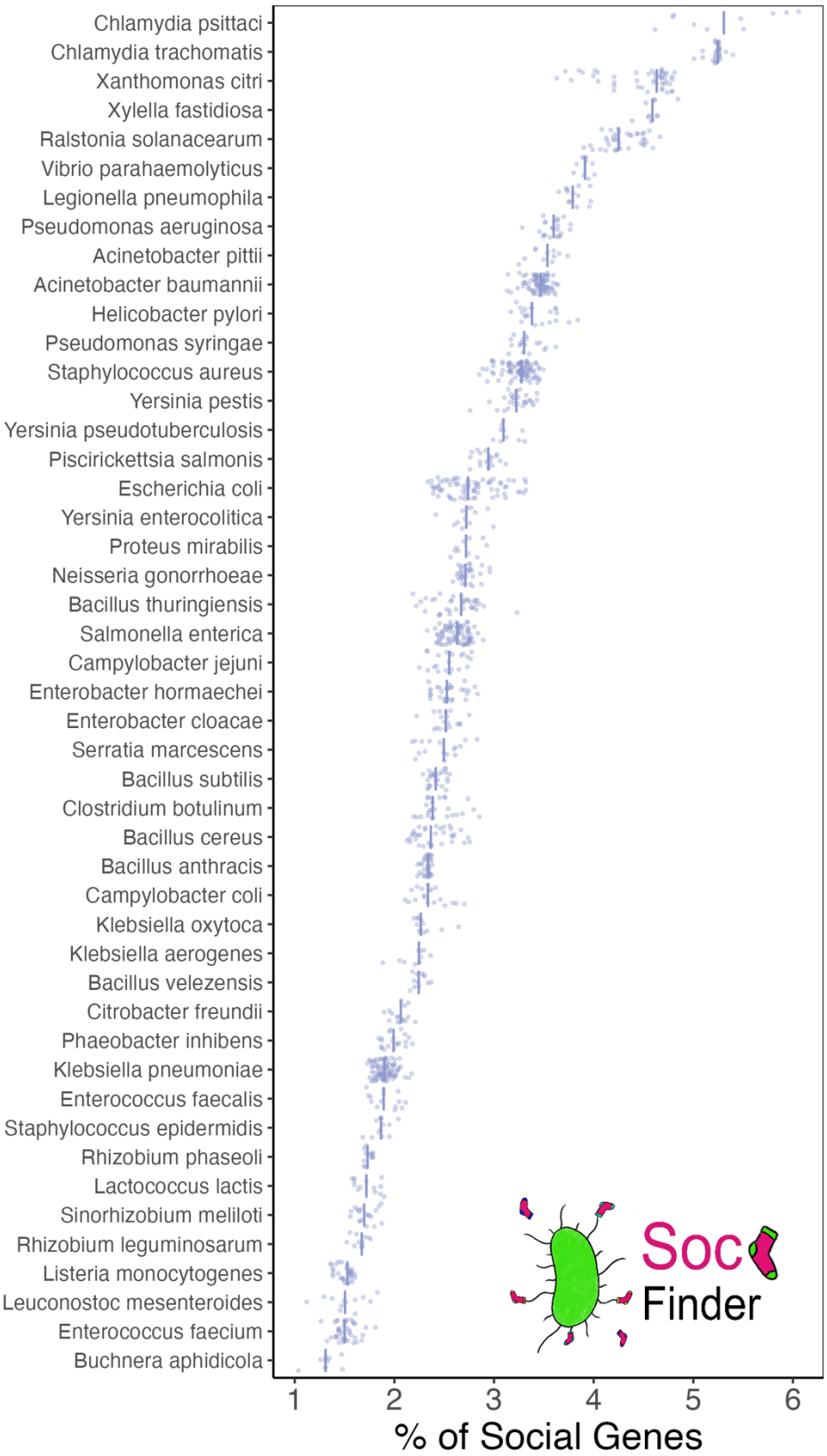


Figure 5: SOCfinder on 1,301 genomes of 51 species. The x-axis shows the proportion of the genes in a genome that are categorised by SOCfinder as cooperative. For each species, a point represents the proportion for one genome, and the bar represents the median proportion.

Comparison of methods in model species

The artisanal method has been used to identify genes for cooperative behaviours in two well studied species: (1) the gram-negative opportunistic pathogen *Pseudomonas aeruginosa* (26); and (2) the gram-positive soil-dwelling *Bacillus subtilis* (27). In both these species, data from laboratory experiments have identified a number of cooperative behaviours, for which the genes have been determined. We used these artisanal data sets to test the ability and accuracy of other automated methods for identifying genes for cooperative behaviours. We compared three automated methods: (1) the most common previously used method – PSORTb (28); (2) a recent combined method – PanSort (combines PSORTb and PANNZER) (20); and (3) our new method - SOCfinder.

We start by looking at how many genes are captured by each method (Figure 6A&B). SOCfinder captures the most genes. Artisanal captures the fewest genes, because it requires detailed experimental evidence. PanSort and PSORTb are intermediate, with PanSort capturing almost as many genes as SOCfinder, while PSORTb captured many less.

We next look at how many of the Artisanal genes are captured by each method (Figure 6C&D). SOCfinder does much better than the other method in both species. In *P. aeruginosa*, SOCfinder captures 68% of the 40 Artisanal genes, which is significantly more than the next best method (38% by PanSort and only 23% by PSORTb, binomial test $p<0.001$). In *B. subtilis*, SOCfinder captures 68% of the 25 Artisanal genes, which is also significantly more than the next best method (PanSort 8%, PSORTb 0%, binomial test $p<10^{-12}$).

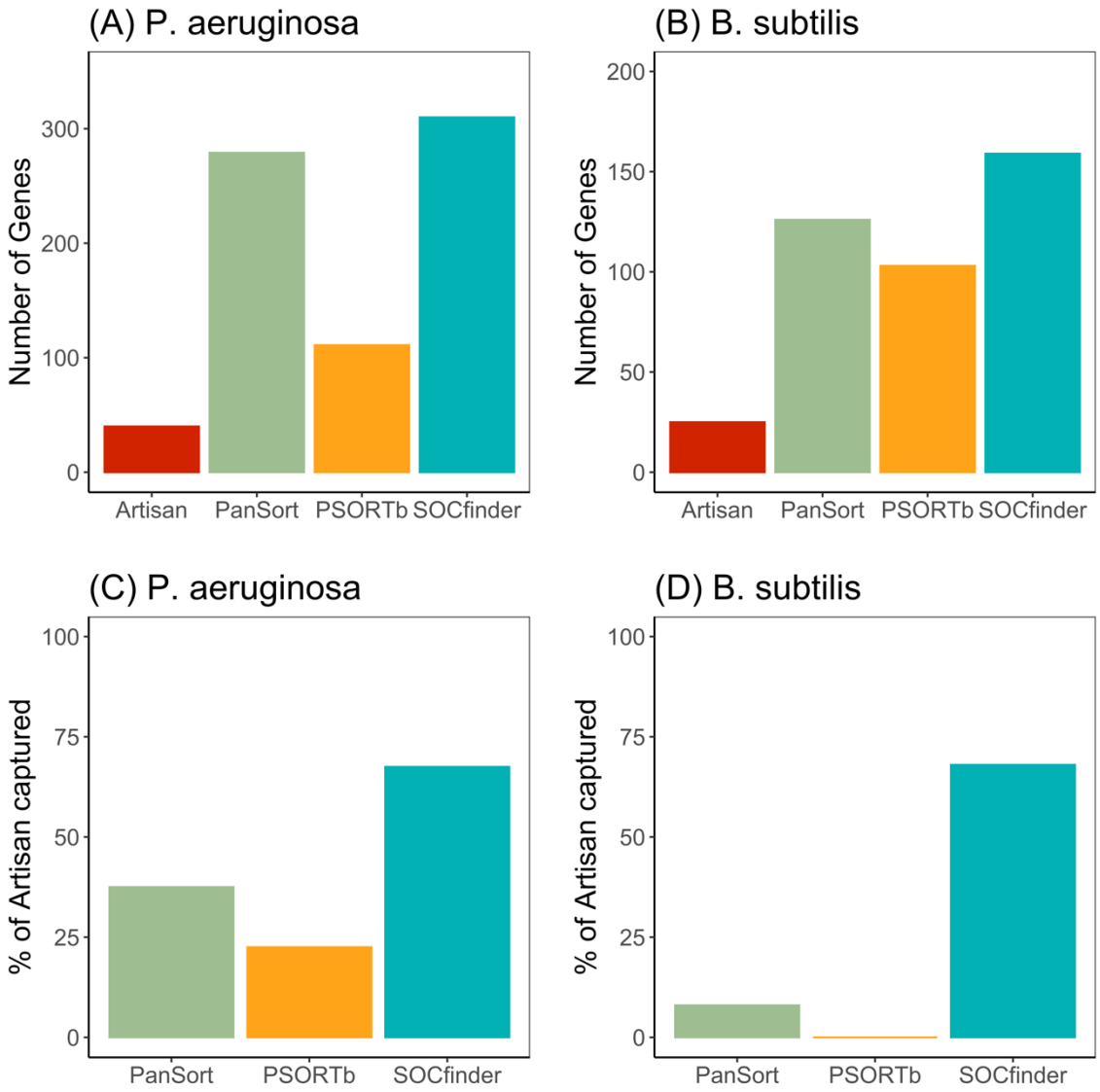


Figure 6: (A&B) Number of genes captured by each method. (C&D) Percentage of artisanal cooperative genes captured by each method. The left panels (A&C) are for *P. aeruginosa*, and the right panels (B&D) are for *B. subtilis*.

One key cooperative trait in *P. aeruginosa* is the production of iron scavenging pyoverdine molecules (6, 7, 41). SOCfinder is more than three times better than PanSort at capturing pyoverdine genes, (24/34 = 71%, compared to 7/34 = 21%, binomial test $p<10^{-9}$) (Supplementary Figure 3). PSORTb does not capture any of the pyoverdine genes (Supplementary Figure 4).

Can we explain why different methods give different results?

There are a number of possible explanations for the lack of overlap, in terms of genes identified, between the different methods (Figure 7). We now examine the explanatory power of these different explanations, to both test the usefulness of different methods, and guide possible future adjustments to SOCfinder.

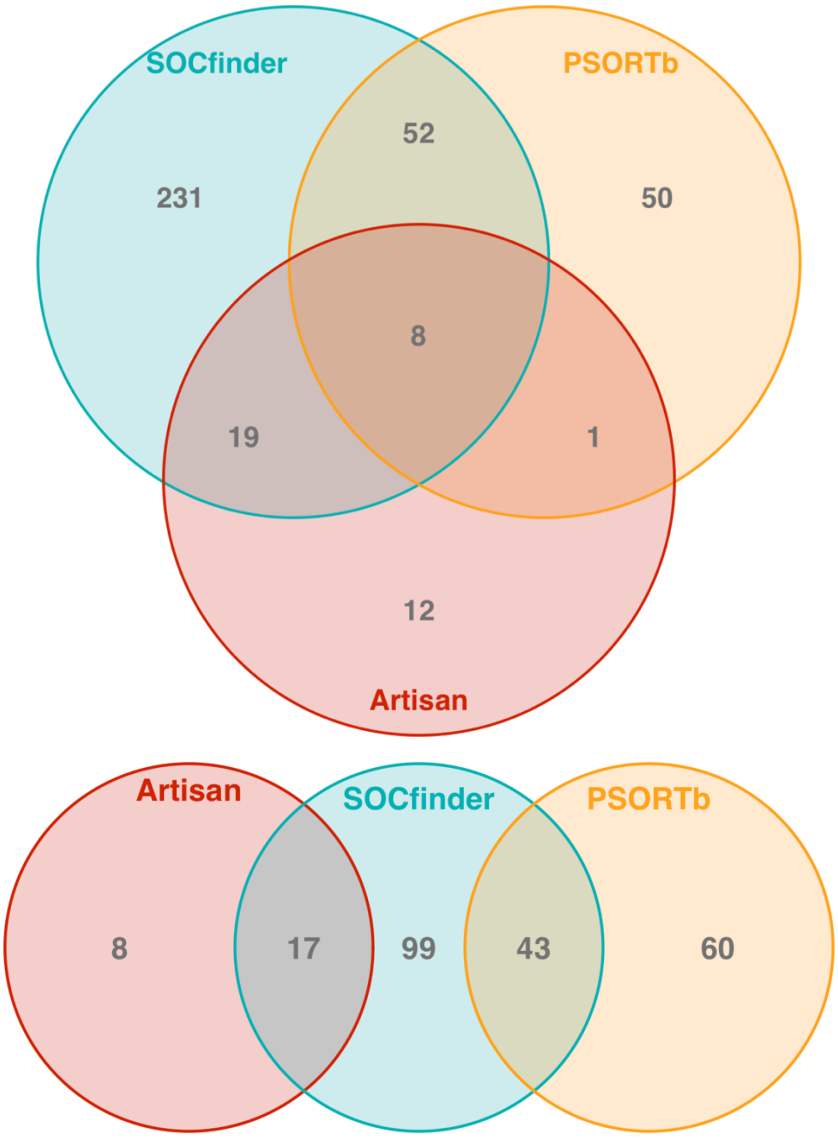


Figure 7: Overlap between methods to find cooperative genes. The top Venn diagram is for *P. aeruginosa*, and the bottom Venn diagram is for *B. subtilis*. The red circle is genes categorised as cooperative by the Artisanal approach. The blue circle is genes categorised as cooperative by SOCfinder. The yellow circle is genes categorised as cooperative (extracellular) by PSORTb.

Which known cooperative genes are not found by PSORTb?

There are many known cooperative genes are not extracellular based on PSORTb (31 genes in *P. aeruginosa,* and 25 in *B. subtilis*). Many of these will be intracellular (such as pyoverdine biosynthesis genes), however it is also possible that PSORTb is too conservative in deciding if a gene is extracellular. If this is true, then PSORTB will list the genes as 'Unknown' localization (21% of all genes in *P. aeruginosa*, 19% in *B. subtilis*). We tested if the missed cooperative genes are more likely to be listed as 'unknown' than the average across the genome. In *P. aeruginosa,* missed cooperative genes aren't overrepresented for unknown genes (16% of missed genes are unknown, binomial test $p=0.66$), but in *B. subtilis* they are (32% of missed genes are unknown, binomial test $p<0.01$).

In gram-negative bacteria which have an outer membrane, another possibility is that PSORTb mistakenly categorises some artisanal cooperative genes as 'outer membrane'. We tested this in *P. aeruginosa*, and found that cooperative genes missed by PSORTb aren't overrepresented for 'outer membrane' genes (2/31 = 6.5% of missing cooperative genes are outer membrane, compared to 3.1% of all genes: binomial test $p=0.244$). This isn't surprising, as we know that many intracellular genes are involved in producing extracellular traits, such as pyoverdine (Supplementary Figures 3&4).

Which extracellular genes are missed by SOCfinder?

SOCfinder doesn't include many genes that are identified by PSORTb as extracellular (51 in *P. aeruginosa*, 55 in *B. subtilis*). This may be because these 'extracellular but non-cooperative' genes have no known function, and so wouldn't have been caught by the functional annotation module of SOCfinder. We found some support for this hypothesis. In both *P. aeruginosa* and *B. subtilis* extracellular genes missed by SOCfinder are significantly more likely to produce a "hypothetical protein" than extracellular genes which are included by SOCfinder (*P. aeruginosa* 21/51 = 41.2% compared to 14/60 = 23.3%, binomial test $p<0.01$; *B. subtilis* 18/55 = 32.7% compared to 0/48 = 0%, binomial test $p<10^{-15}$).

Why are some Artisanal cooperative genes missed by both PanSort and SOCfinder?

There are several known cooperative genes which are missed by both PanSort and SOCfinder (11 genes in *P. aeruginosa,* and 7 in *B. subtilis*). These genes are missed because the annotations they are given don't match a known cooperative function, although most have a significant annotation (10/12 = 83.3% in *P. aeruginosa;* 3/7 = 42.9% in *B. subtilis*). Often these annotations are too broad to be useful for our purposes, such as "protease I". Future work is likely to improve functional annotation pipelines, which may allow these missing genes to be eventually captured.

Can we detect kin selection for cooperation in genes for cooperative behaviours?

Another way to test the usefulness of the different approaches for identifying genes for cooperation is with population genetics. Population genetic theory suggests that selection is relaxed on cooperative genes relative to private genes, making deleterious mutations more likely to fix, and beneficial mutations less likely to fix (65–69). This is because cooperative genes only provide a benefit to carriers of the gene a certain proportion of the time, based on the likelihood that the recipient shares the cooperative gene (genetic relatedness, r). Consequently, genes for cooperative behaviours favoured by kin selection, in non-clonal populations ($r<1$) should show increased polymorphism and divergence relative to genes for private behaviours.

Studies on both *P. aeruginosa* and *B. subtilis* have supported this prediction (26, 27). However, these studies used the artisanal approach to identify cooperative and private genes. The artisanal approach was used in these studies because accuracy of identification of cooperative genes is required to be able to pick up possibly subtle population genetic patterns, that could be missed by larger but potentially more messy data sets, compiled with other approaches. In this section, we ask whether other methods to identify cooperative genes give similar results. If the results of an approach do not agree with an analysis on artisanal selected genes, then it could suggest a possible problem with that alternative approach. We examined patterns of polymorphism and divergence for cooperative and private genes identified with three methods: (1) PSORTb; (2) PanSort; and (3) SOCfinder.

When examining genes identified by PSORTb we did not find the expected pattern of increased polymorphism (Figure 8 C&F) and divergence (Supplementary Figures 1&2). There was no significant difference in polymorphism between cooperative and private genes in *P. aeruginosa*

549 (Kruskal–Wallis X2=0.45, p=0.80) or in *B. subtilis* (Kruskal–Wallis X2=2.37, p=0.31). Non-
550 synonymous divergence was significantly higher in cooperative genes compared to private
551 genes in *P. aeruginosa* (Kruskal–Wallis X2=13.2, p<0.01, Dunn Test p=0.03), but not in *B*
552 *subtilis* (Kruskal–Wallis X2=0.51, p=0.77). Synonymous divergence was not significantly
553 different in cooperative genes compared to private genes in *P. aeruginosa* (Kruskal–Wallis
554 X2=2.86, p=0.24), or in *B. subtilis* (Kruskal–Wallis X2=5.74, p=0.06).
555
556 When examining genes identified by PanSort we also did not find the expected pattern of
557 increased polymorphism (Figure 8 B&E) and divergence (Supplementary Figures 1&2). There
558 was no significant difference in polymorphism between cooperative and private genes in *P.*
559 *aeruginosa* (Kruskal–Wallis X2=1.35, p=0.51) or in *B. subtilis* (Kruskal–Wallis X2=3.81,
560 p=0.15). Non-synonymous divergence was significantly higher in cooperative genes compared
561 to private genes in *P. aeruginosa* (Kruskal–Wallis X2=24.3, p<0.0001, Dunn Test p=0.03), but
562 not in *B subtilis* (Kruskal–Wallis X2=2.28, p=0.32). Synonymous divergence was significantly
563 higher in cooperative genes compared to private genes in *P. aeruginosa* (Kruskal–Wallis
564 X2=9.46, p<0.01, Dunn Test p<0.01), but not in *B subtilis* (Kruskal–Wallis X2=14.73,
565 p<0.001, Dunn Test p=0.26). This indicates that PanSort may be performing better in *P.*
566 *aeruginosa* than it does in *B. subtilis*.
567
568 In contrast, when we identified cooperative and private genes with SOCfinder, we did find that
569 cooperative genes had the signature of kin selection for cooperation, with elevated
570 polymorphism (Figure 8 A&D) and divergence (Supplementary Figures 1&2) compared to
571 private genes. Polymorphism was significantly higher in cooperative genes compared to private
572 genes in both species (*P. aeruginosa*: Kruskal–Wallis X2=6.12, p<0.05, Dunn Test p=0.04. *B.*
573 *subtilis* Kruskal–Wallis X2=8.48, p<0.02, Dunn Test p=0.01). Non-synonymous divergence
574 was significantly higher in cooperative genes compared to private genes in both species (*P.*
575 *aeruginosa*: Kruskal–Wallis X2=21.1, p<0.0001, Dunn Test p=0.006. *B. subtilis* Kruskal–
576 Wallis X2=8.26, p<0.02, Dunn Test p=0.02). Synonymous divergence was significantly higher
577 in cooperative genes compared to private genes in *P. aeruginosa* (Kruskal–Wallis X2=9.60,
578 p<0.01, Dunn Test p<0.01), but the trend was significant in *B. subtilis* (Kruskal–Wallis
579 X2=16.70, p<0.001, Dunn Test p=0.08).
580

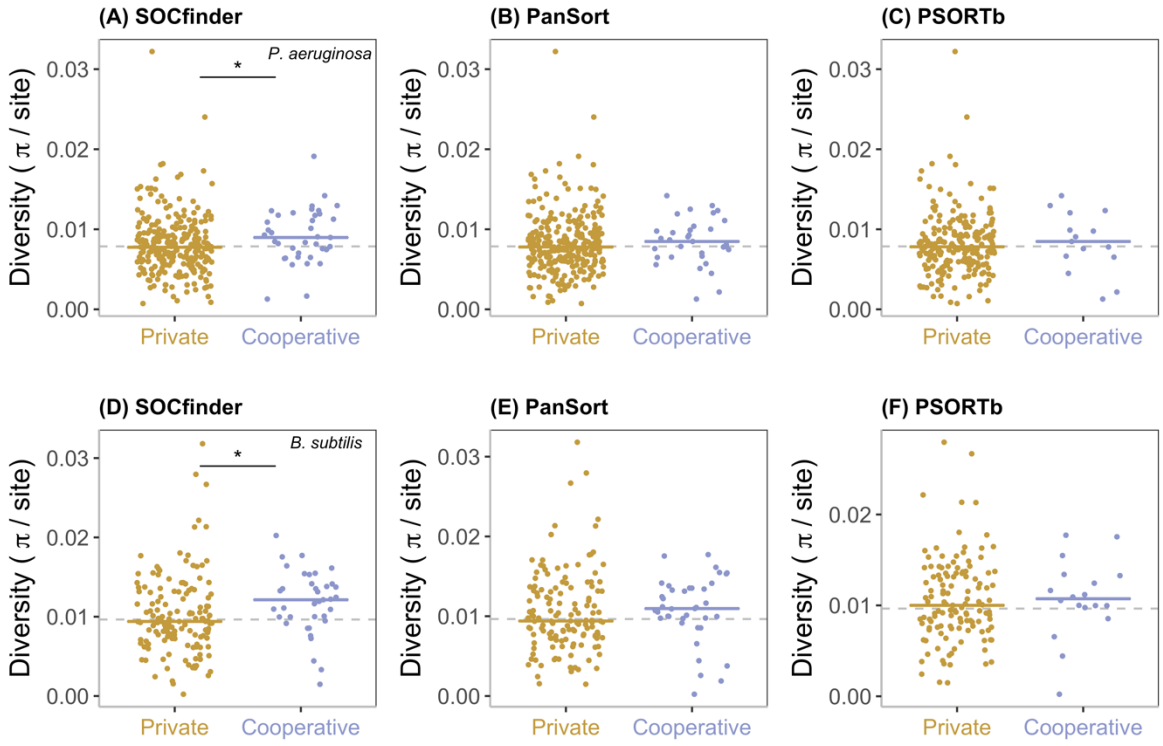


Figure 8: Nucleotide polymorphism for private (gold) and cooperative (blue) quorum-sensing controlled genes. The top three graphs (A-C) show *P. aeruginosa*, and the bottom three graphs (D-F) show *B. subtilis*. The left graphs (A&D) show cooperative genes identified by SOCfinder. The middle graphs (B&E) show cooperative genes identified by PanSort. The right graphs (C&F) show cooperative genes identified by PSORTb. For each graph, the dotted line shows the background level of nucleotide polymorphism for a set of private genes. The black line and * shows a significant difference between cooperative and private genes.

Discussion

We have developed a bioinformatic tool for identifying genes for cooperative behaviours in bacteria. SOCfinder combines information from several methods, and still only takes less than 10 minutes to identify cooperative genes in an average bacterial genome (Supplement S3). Our analyses suggest that SOCfinder both identifies cooperative genes more accurately, and finds more cooperative genes, compared with previous methods such as PSORTb or a combination of PSORTb with functional annotation (PanSort). In addition, these other methods appear to mis-assign genes, to the extent that they are unable to capture the underlying population genetic processes.

The different methods for identifying cooperative genes each have different pros and cons (Table 4). The artisanal method, based on the results of examining behaviours with laboratory experiments represents the relative gold standard in terms of accuracy. It is for this reason that we used it previously when carrying out population genetic analyses, where any incorrect assignments would have introduced noise that could have concealed underlying patterns (26, 27). However, this approach is labour intensive, produces a limited number of genes, and is restricted to species where there has been considerable experimental work, such as *P. aeruginosa* and *B. subtilis*. For example, it identified 40 genes for cooperation in *P. aeruginosa* and 25 genes in *B. subtilis.* Consequently, this approach cannot be applied across the whole genome, to a wide range of species, or to facilitate broad comparative studies.

Methods such as PSORTb are potentially less accurate, but can be automated, and applied across the whole genome of a wide range of species. PSORTb has been used to identify genes for cooperation in a number of studies, for both studies of single species, and broad across species studies (21–23, 25). This has allowed many more genes and many more species to be analysed in a single study. However, PSORTb introduces some inaccuracies with how it identifies cooperative genes, capturing none of the artisanal identified cooperative genes in *B. subtilis*, and only 23% in *P. aeruginosa.* In addition, our population genetic analyses that the level of inaccuracy is sufficient that the noise introduced prevents us from observing the signature (footprint) of kin selection for cooperation at the genomic level.

The importance of the potential problems with using PSORTb can depend upon the kind of question being asked. For example, if you want to know if cooperative genes evolve fast in symbionts, then you need to categorise ('bin') genes as either cooperative or private. You don't want to miss many cooperative genes, because they would then be categorised as private and introduce noise to any comparison. PSORTb could be a problematic approach for such questions. In contrast, if you just wanted to know which intracellular pathogens have the most cooperative genes ('counting'), then it is less important if you miss some cooperative behaviours. Extracellular genes are likely to be a good proxy for this, and so using PSORTb could be less problematic. The PanSort method developed by Simonet & McNally fixes some of the problems of PSORTb by including some functional annotation (20). However, we show that PanSort doesn't make full use of power of functional annotation, and still performs badly

on the best studied cooperative traits like pyoverdine (Supplementary Figures 3&4), and when comparing to the gold-standard artisanal method.

SOCfinder allows large scale analyses, across whole genomes, and across a broad range of species, but without the same level of problems introduced by PSORTb. SOCfinder is more accurate in identifying cooperative genes because it uses contextual information on the quality of functional annotations, and includes antiSMASH to capture full clusters of biosynthetic genes for key cooperative traits like pyoverdine (Supplementary Figures 3&4). SOCfinder captures variation in the cooperative gene repertoire of bacteria. SOCfinder performs better than other methods in replicating the signature of kin selection that we know exists from studies that have used the gold-standard artisanal approach. To a large extent therefore, SOCfinder has the advantages of methods such as PSORTb, while significantly reducing the disadvantages.

Table 4: Advantages and disadvantages of methods

Issue	SOCfinder	PSORTb	ARTISANAL
Key advantage	Highly flexible Can capture all known types of gene for cooperative behaviour.	Not subjective Doesn't require judgement about which behaviours are cooperative.	Certainty Experimental evidence gives us high confidence that a gene is cooperative
Behaviours captured	Any	Extracellular proteins	Any
Bias	Potential taxonomic bias in training set for extracellular and functional annotation modules Depends on subjective categorisation of behaviours	Misses intracellular cooperative behaviours (e.g. siderophores or exopolysaccharides) Includes some known private behaviours (e.g. proteins tethered to membrane)	Requires culturing a species in the lab, knowledge of the environment in which the trait is favoured, and ability to edit the genome to generate cheaters
Precision	Adjustable – can adjust parameters to force prediction or apply high confidence threshold	High precision – doesn't force a prediction for each gene (~25% of genes annotated as "Unknown")	Very high precision
Adaptability	Users can adjust; - Cooperative annotation list	Users can use the 'Extracellular score' to exclude lower-confidence genes	Standard methodology is applied to all species

	- Score and significance thresholds - Cooperative metabolite list		
Output	Can be split into categories (e.g. by function)	One list of cooperative genes	One list of cooperative genes
Speed	10 minutes per genome	30 minutes per genome	Very slow (years)
Ease of use	Easy: Command line	Very easy: Interactive webpage, or Command line	Simple experiments
Available species	All	All	Very limited

To conclude, SOCfinder opens up a number of exciting directions for future research. It will allow both detailed studies of non-model species, and broad across species studies. These studies will allow cooperation, and how cooperation shapes the genome, to be studied in new ways, such as in natural populations of bacteria. As one example, we could investigate if species that use greenbeards (37, 78, 79) or genetic kin recognition mechanisms (2, 80, 81) have more cooperative genes than those that use environmental kin recognition. In addition, SOCfinder could be used to reassess the results of previous studies which used methods such as PSORTb. We have shown how such methods could lead to limited or inaccurate identification of gene function, and that this could be particularly important if 'binning' approaches were used to compare 'cooperative' to 'non-cooperative' genes. It is still unknown whether the unavoidable inaccuracies imposed by methodologies such as PSORTb have led to biased conclusions.

Acknowledgements

We thank Carolin Kobras & Ming Liu for useful discussion and comments on the manuscript, and Sarah Flint and Harvey Jeffrey for testing the tool. This work was supported by the European Research Council (834164: LJB, AED & SAW; SESE: MG), and the Biotechnology and Biological Sciences Research Council (BBSRC Oxford Interdisciplinary Bioscience Doctoral Training Partnership ZK). The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work.

Conflict of Interest

The authors declare that there are no conflicts of interest.

Supplement S1: Taxonomic bias

We conducted a Web of Science search for papers on cooperation in bacteria. Specifically, we searched for [Topic = "cooperation" OR "public good" AND "bacteria"] AND [Year = since 2000] AND [Type = "article"] AND [Category = "microbiology" or "evolutionary biology"]. This gave n=464 papers.

We took the list of bacteria genera from the Approved List of Bacterial Names (https://lpsn.dsmz.de/). We included only those genera which are culturable, validly published, and have a correct name (*n*=4180 genera).

We then looked for genus names in the title, keywords, and abstract of the papers. 308 of the papers have a genus name mentioned in one of those three places. More than 28% of these papers mention *Pseudomonas*, with even *Escherichia* and *Bacillus* lagging behind (Figure S1.1)

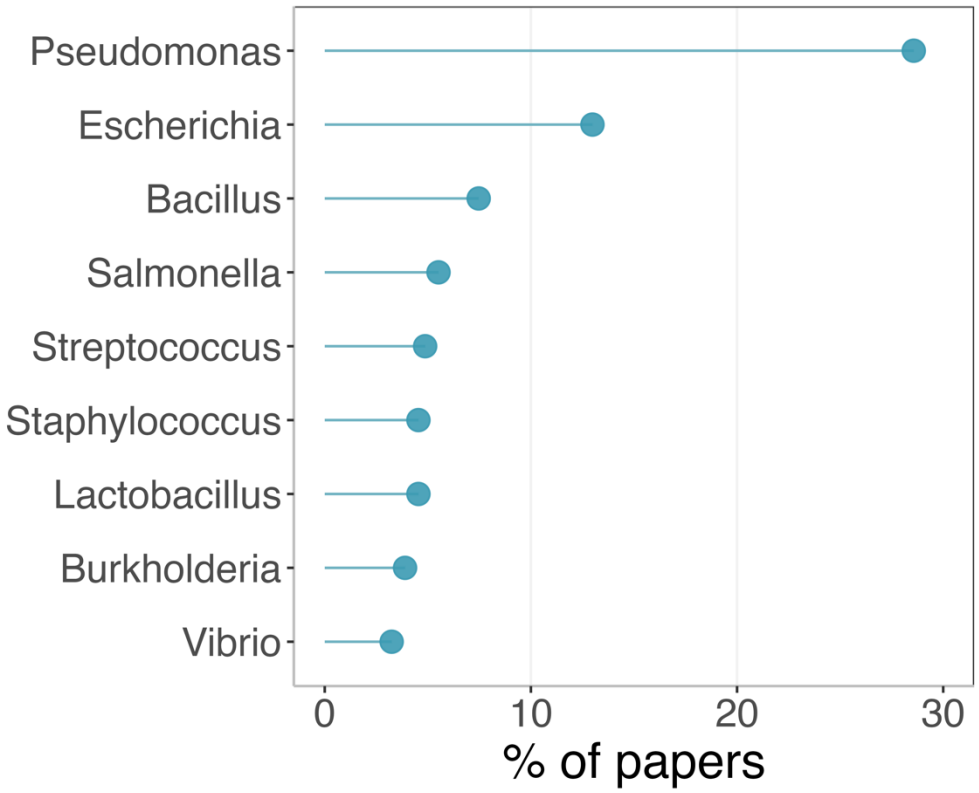


Figure S1.1: The percentage of papers about microbial cooperation that mention each genus in the title, abstract, or keywords.

Supplement S2: Other tools

Table S2: List of alternative tools and methods to find cooperative genes. Tools are separated based on what they find, how they find it, and what species they work on.

Tool / Method	What does it do?	Reference
FUNCTIONAL ANNOTATION		
eggnog-mapper	Functional annotation based on a large database of orthologous relationships	(51)
DeepFRI	Functional annotation based on structure, using a machine learning approach	(82)
Microbe Annotator	Combines functional annotation of several tools, including both GO and KO databases	(83)
EXTRACELLULAR PROTEINS		
PSO-LocBact	Combines the predictions of multiple tools that predict subcellular localization (e.g. PSORTb) to reach a consensus prediction	(84)
Psortm	Finds extracellular genes in metagenomes. Combines PSORTb with a computational classification of gram-stain.	(85)
SignalP	Predicts the presence of signal peptides, including secretory signal peptides	(86)
SPECIALIST TOOLS		
PathoFact	Finds virulence factors, toxins, and resistance genes in metagenomes, using sequence homology and machine learning	(87)
EffectiveDB	Finds intact secretion systems, based on known protein domains and secretion signals	(88)
Metage2Metabo	Reconstructs metabolic networks from sequence data. Can be used to look for mutualistic cross-feeding	(89)
CAZy	Finds carbohydrate-active enzymes, some of which will be cooperative (e.g. rhamnolipid biosynthesis)	(90)
Machine learning algorithms	Find cheats that lack the core genes for a cooperative behaviour, but maintain the genes that accompany it (e.g. receptors).	(91)
VIRAL COOPERATION		
DI-tector	Detects certain types of defective interfering genome in viruses	(92)
ViReMa	Can detect deletions in viral genomes	(93)
VODKA	Detects certain types of defective interfering genome in viruses	(94)
ANTIVIRAL COOPERATION		
PADLOC	Detects antiviral defence mechanisms in bacterial genomes, using sequence homology	(95)
DefenseFinder	Detects antiviral defence mechanisms in bacterial genomes, using sequence homology	(96)
EXPERIMENTAL METHODS		
Experimental evolution	Evolve a population under low relatedness, and sequence the cheats that emerge	(97)

Supplement S3: Speed test

We ran a small test to compare the speed of SOCfinder with the speed of PanSort. PSORTb is the slowest part of PanSort, so the time for PanSort is equivalent to the time for PSORTb.

We chose ten *Escherichia coli* genomes for our speed test (accession numbers GCA_013357365.1, GCA_005221885.1, GCA_004358365.1, GCA_030013595.1, GCA_008931135.1, GCA_009650035.1, GCA_006874785.1, GCA_001612475.1, GCA_024223415.1, GCA_024223415.1). The test was run on a 2020 iMac with a 3.8GHz 8-core intel i7 processor and 32GB RAM.

SOCfinder is significantly quicker than PanSort (t-test, t=109.95, df=17, $p<10^{-15}$), taking 8 minutes on average per genome, compared to 32 minutes for PanSort (Figure S3.1).

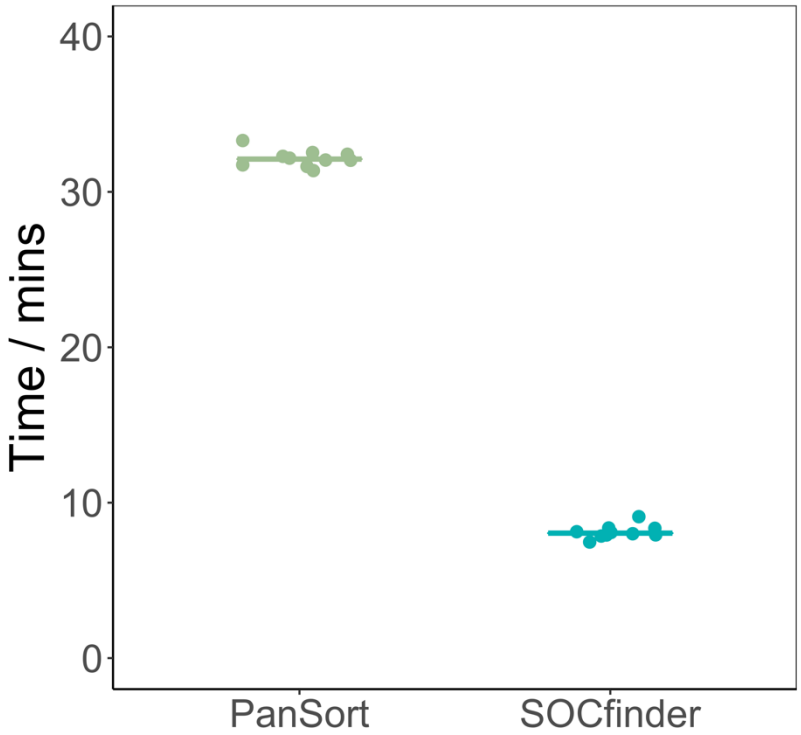
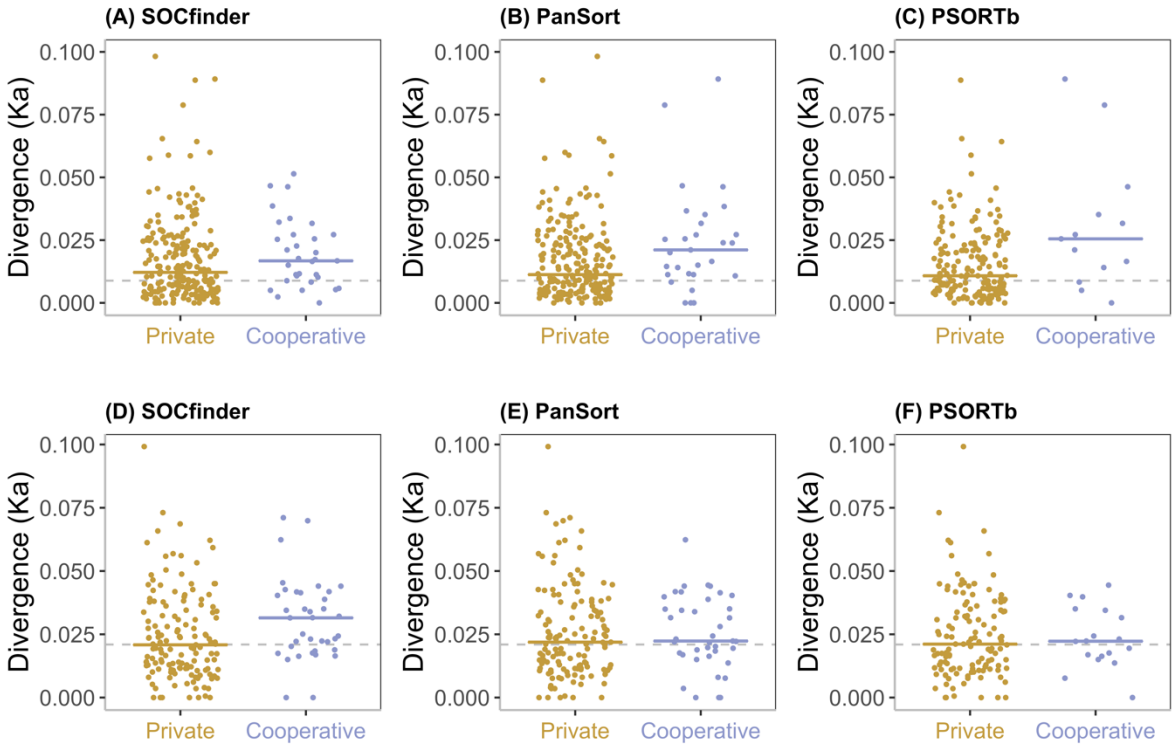
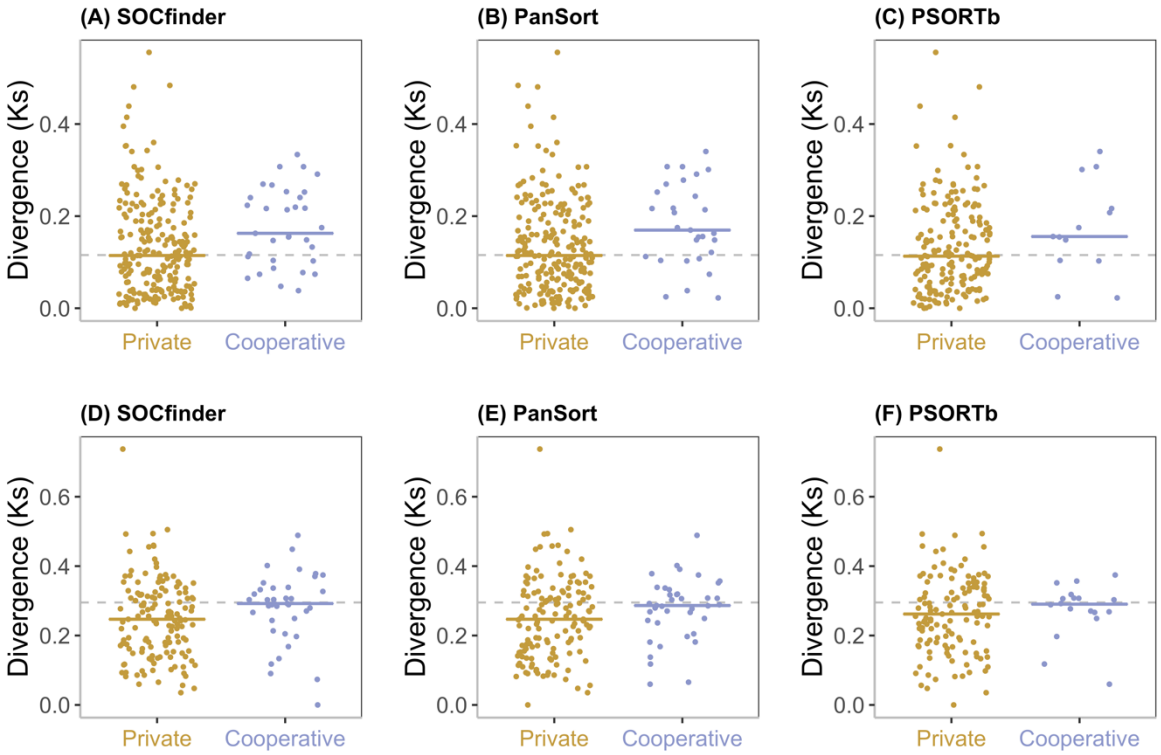


Figure S3.1: Time in minutes for PanSort (green) and SOCfinder (blue) to run on ten *E. coli* genomes.

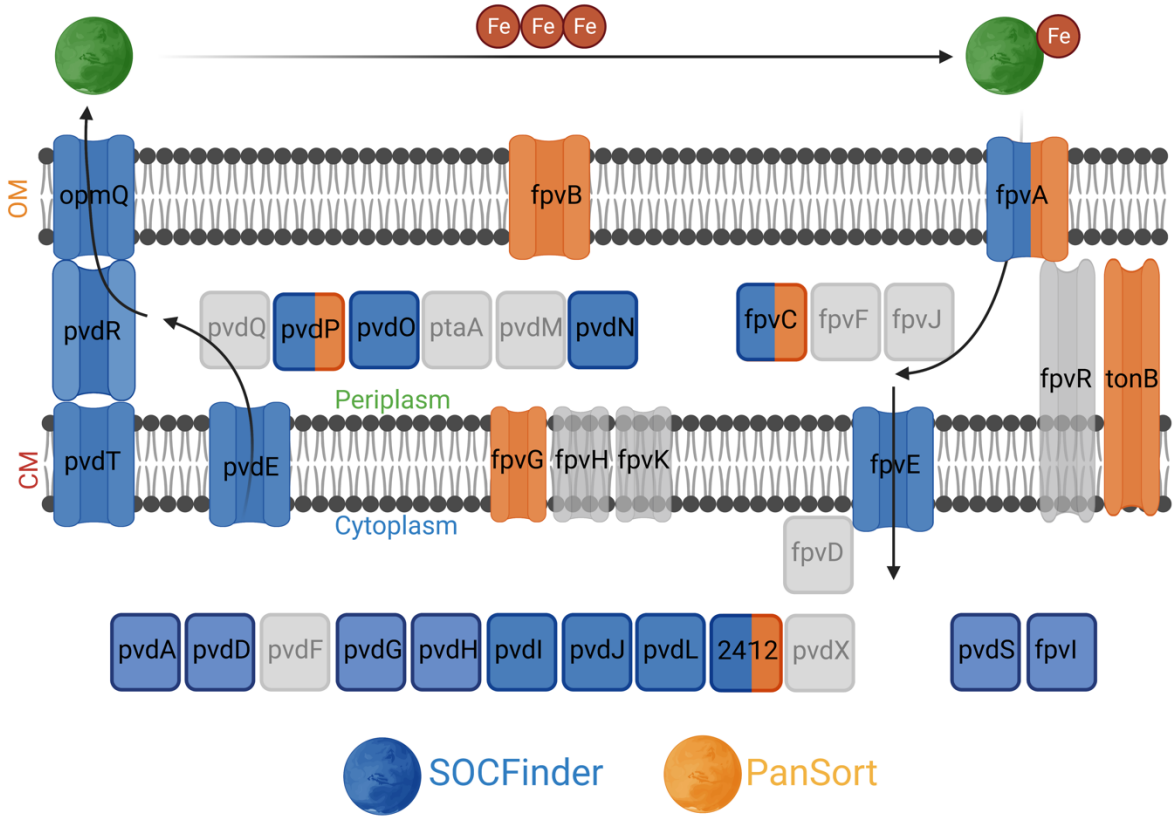
Supplementary Figures



Supplementary Figure 1: Non-synonymous divergence for private (gold) and cooperative (blue) quorum-sensing controlled genes. The top three graphs (A-C) show *P. aeruginosa*, and the bottom three graphs (D-F) show *B. subtilis*. The left graphs (A&D) show cooperative genes identified by SOCfinder. The middle graphs (B&E) show cooperative genes identified by PanSort. The right graphs (C&F) show cooperative genes identified by PSORTb. For each graph, the dotted line shows the background level of non-synonymous divergence for a set of private genes.

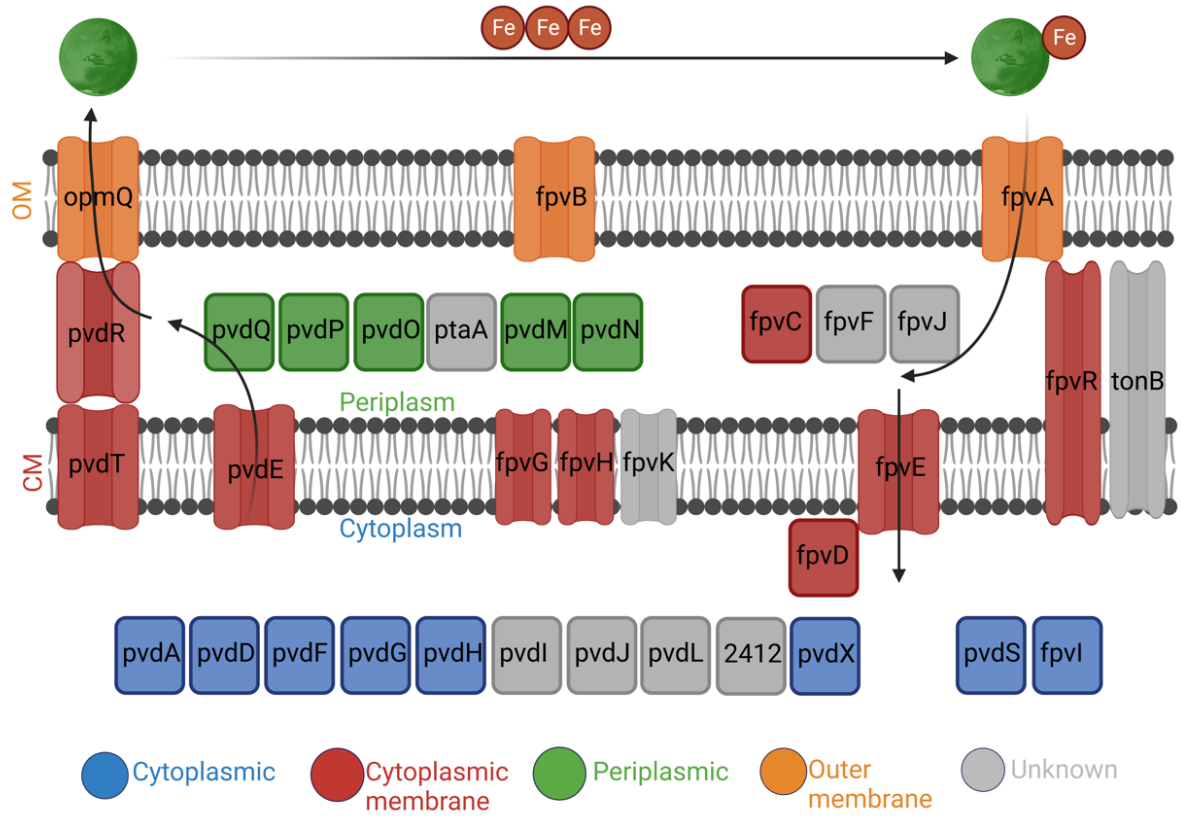


Supplementary Figure 2: Synonymous divergence for private (gold) and cooperative (blue) quorum-sensing controlled genes. The top three graphs (A-C) show *P. aeruginosa*, and the bottom three graphs (D-F) show *B. subtilis*. The left graphs (A&D) show cooperative genes identified by SOCfinder. The middle graphs (B&E) show cooperative genes identified by PanSort. The right graphs (C&F) show cooperative genes identified by PSORTb. For each graph, the dotted line shows the background level of synonymous divergence for a set of private genes.



Supplementary Figure 3: Schematic of which genes involved in the biosynthesis, export, intake, and use of pyoverdine are captured by SOCfinder and PanSort. Genes in blue are captured by SOCfinder. Genes in orange are captured by PanSort. Genes which are half blue and half orange are captured by both tools. Genes in grey are captured by no tool. Layout of genes is adapted from Ringel & Bruser (2018).

727



Supplementary Figure 4: Schematic of the PSORTb subcellular localisation of proteins produced by genes involved in the biosynthesis, export, intake, and use of pyoverdine. Genes in blue code for proteins that are predicted to be cytoplasmic. Genes in red code for cytoplasmic membrane proteins. Genes in green code for periplasmic proteins. Genes in orange code for outer membrane proteins. Genes in grey code for proteins of unknown localisation. No genes code for extracellular proteins. Layout of genes is adapted from Ringel & Bruser (2018).

References

1. S. A. West, A. S. Griffin, A. Gardner, S. P. Diggle, Social evolution theory for microorganisms. *Nat Rev Microbiol* **4**, 597–607 (2006).
2. J. E. Strassmann, O. M. Gilbert, D. C. Queller, Kin discrimination and cooperation in microbes. *Annu. Rev. Microbiol.* **65**, 349–367 (2011).
3. M. Ghoul, S. B. Andersen, S. A. West, Sociomics: Using Omic Approaches to Understand Social Evolution. *Trends Genet.* **33**, 408–419 (2017).
4. S. Mitri, K. Richard Foster, The genotypic view of social interactions in microbial communities. *Annu. Rev. Genet.* **47**, 247–273 (2013).
5. S. A. West, G. A. Cooper, M. B. Ghoul, A. S. Griffin, Ten recent insights for our understanding of cooperation. *Nat. Ecol. Evol. 2021 54* **5**, 419–430 (2021).
6. 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).
7. A. S. Griffin, S. A. West, A. Buckling, Cooperation and competition in pathogenic bacteria. *Nature* **430**, 1024–1027 (2004).
8. F. Harrison, A. Buckling, Siderophore production and biofilm formation as linked social traits. *ISME J.* **3**, 632–634 (2009).
9. R. Kümmerli, A. Ross-Gillespie, Explaining the sociobiology of pyoverdin producing pseudomonas: A comment on zhang and rainey (2013). *Evolution (N. Y).* **68**, 3337–3343 (2014).
10. S. O'Brien, D. J. Hodgson, A. Buckling, Social evolution of toxic metal bioremediation in Pseudomonas aeruginosa. *Proc. R. Soc. B Biol. Sci.* **281** (2014).
11. 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).
12. A. Dragoš, *et al.*, Division of Labor during Biofilm Matrix Production. *Curr. Biol.* **28**, 1903-1913.e5 (2018).
13. Y. Chai, F. Chu, R. Kolter, R. Losick, Bistability and biofilm formation in Bacillus subtilis. *Mol. Microbiol.* **67**, 254–263 (2008).
14. 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).
15. S. B. Andersen, *et al.*, Long-term social dynamics drive loss of function in pathogenic bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10756–10761 (2015).
16. O. X. Cordero, L. -a. 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. S. Sathe, A. Mathew, K. Agnoli, L. Eberl, R. Kümmerli, Genetic architecture constrains exploitation of siderophore cooperation in the bacterium Burkholderia cenocepacia. *Evol. Lett.* **3**, 610–622 (2019).
18. 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.
19. J. Van Gestel, F. J. Weissing, O. P. Kuipers, Á. T. Kovács, Density of founder cells affects spatial pattern formation and cooperation in Bacillus subtilis biofilms. *ISME J.* **8**, 2069–2079 (2014).
20. 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.
21. 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).

22. T. Nogueira, *et al.*, Horizontal Gene Transfer of the Secretome Drives the Evolution of Bacterial Cooperation and Virulence. *Curr. Biol.* **19**, 1683–1691 (2009).
23. 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).
24. C. Hao, A. E. Dewar, S. A. West, M. Ghoul, Gene transferability and sociality do not correlate with gene connectivity. *Proc. R. Soc. B Biol. Sci.* **289**, 20221819 (2022).
25. A. Dewar, *et al.*, Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range. *Nat. Ecol. Evol.* **5**, 1624–1636 (2021).
26. L. J. Belcher, A. E. Dewar, M. Ghoul, S. A. West, Kin selection for cooperation in natural bacterial populations. *Proc. Natl. Acad. Sci. U. S. A.* **119** (2022).
27. L. J. Belcher, A. E. Dewar, C. Hao, M. Ghoul, S. A. West, Signatures of kin selection in a natural population of the bacteria Bacillus subtilis. *Evol. Lett.*, 1–21 (2023).
28. 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).
29. P. Törönen, L. Holm, PANNZER—A practical tool for protein function prediction. *Protein Sci.* **31**, 118–128 (2022).
30. M. T. Ringel, T. Brüser, The biosynthesis of pyoverdines. *Microb. Cell* **5**, 424–437 (2018).
31. 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).
32. J. S. B. Tai, *et al.*, Social evolution of shared biofilm matrix components. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2123469119 (2022).
33. S. Pollak, *et al.*, Facultative cheating supports the coexistence of diverse quorum-sensing alleles. *Proc. Natl. Acad. Sci.* **113**, 2152–2157 (2016).
34. S. A. West, A. S. Griffin, A. Gardner, Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *J. Evol. Biol.* **20**, 415–432 (2007).
35. S. A. S. A. West, A. S. Griffin, A. Gardner, Evolutionary Explanations for Cooperation. *Curr. Biol.* **17**, R661-672 (2007).
36. M. Ghoul, A. S. Griffin, S. A. West, Toward an evolutionary definition of cheating. *Evolution (N. Y).* **68**, 318–331 (2014).
37. W. D. Hamilton, The Genetical Evolution of Social Behaviour. II. *J. Theor. Biol.* **7**, 17–52 (1964).
38. G. F. Oster, E. . Wilson, *Caste and Ecology in the Social Insects* (Princeton University Press, 1978).
39. G. S. Wilkinson, Reciprocal food sharing in the vampire bat. *Nature* **308**, 181–184 (1984).
40. T. H. Clutton-Brock, *et al.*, Cooperation, Control, and Concession in Meerkat Groups. *Science (80-.).* **478**, 478–481 (2001).
41. S. A. West, A. Buckling, Cooperation, virulence and siderophore production in bacterial parasites. *Proc. R. Soc. B Biol. Sci.* **270**, 37–44 (2003).
42. M. Schuster, D. J. Sexton1, B. A. Hense, Why quorum sensing controls private goods. *Front. Microbiol.* **8**, 1–16 (2017).
43. M. Ghoul, *et al.*, Bacteriocin-mediated competition in cystic fibrosis lung infections. *Proc. R. Soc. B Biol. Sci.* **282** (2015).
44. E. T. Granato, T. A. Meiller-Legrand, K. R. Foster, T. A. Meiller-Legrand, K. R. Foster, The evolution and ecology of bacterial warfare. *Curr. Biol.* **29**, 1–39 (2019).
45. E. Kessler, M. Safrin, J. K. Gustin, D. E. Ohman, Elastase and the LasA Protease of Pseudomonas aeruginosa Are Secreted with Their Propeptides. *J. Biol. Chem.* **273**, 30225–30231 (1998).

46. S. P. Diggle, A. S. Griffin, G. S. Campbell, S. A. West, Cooperation and conflict in quorum-sensing bacterial populations. *Nature* **450**, 411–414 (2007).
47. Ö. Özkaya, R. Balbontín, I. Gordo, K. B. Xavier, Cheating on Cheaters Stabilizes Cooperation in Pseudomonas aeruginosa. *Curr. Biol.* **28**, 2070-2080.e6 (2018).
48. T. J. Scott, Cooperation loci are more pleiotropic than private loci in the bacterium Pseudomonas aeruginosa. *Proc. Natl. Acad. Sci.* **119**, e2214827119 (2022).
49. L. McNally, M. Viana, S. P. Brown, Cooperative secretions facilitate host range expansion in bacteria. *Nat. Commun.* **5** (2014).
50. T. Aramaki, *et al.*, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
51. C. P. Cantalapiedra, A. Herņandez-Plaza, I. Letunic, P. Bork, J. Huerta-Cepas, eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
52. G. J. Velicer, M. Vos, Sociobiology of the Myxobacteria. *Annu. Rev. Microbiol.* **63**, 599–623 (2009).
53. M. Vos, G. J. Velicer, Social Conflict in Centimeter-and Global-Scale Populations of the Bacterium Myxococcus xanthus. *Curr. Biol.* **19**, 1763–1767 (2009).
54. R. H. Kessin, *Dictyostelium: Evolution, Cell Biology, and the Development of Multicellularity* (Cambridge University Press, 2001).
55. J. E. Strassmann, D. C. Queller, Evolution of cooperation and control of cheating in a social microbe. *Proc. Natl. Acad. Sci.* **108**, 10855–10862 (2011).
56. J. E. Strassmann, Y. Zhu, D. C. Queller, Altruism and social cheating in the social amoeba Dictyostelium discoideum. *Nature* **408**, 965–967 (2000).
57. P. G. Madgwick, B. Stewart, L. J. Belcher, C. R. L. Thompson, J. B. Wolf, Strategic investment explains patterns of cooperation and cheating in a microbe. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4823–E4832 (2018).
58. L. J. Belcher, *et al.*, Developmental constraints enforce altruism and avert the tragedy of the commons in a social microbe. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2111233119 (2022).
59. J. L. de Oliveira, *et al.*, Conditional expression explains molecular evolution of social genes in a microbe. *Nat. Commun.* **10**, 3284 (2019).
60. A. M. Sharrar, *et al.*, Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type (2020) https:/doi.org/10.1128/mBio (October 24, 2022).
61. K. Blin, *et al.*, antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
62. S. Zhang, R. Mukherji, S. Chowdhury, L. Reimer, P. Stallforth, Lipopeptide-mediated bacterial interaction enables cooperative predator defense. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2013759118 (2021).
63. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
64. L. McNally, *et al.*, Killing by Type VI secretion drives genetic phase separation and correlates with increased cooperation. *Nat. Commun.* **8**, 14371 (2017).
65. 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).
66. 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).

67. 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).
68. 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).
69. 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).
70. 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).
71. N. Comella, A. D. Grossman, Conservation of genes and processes controlled by the quorum response in bacteria: Characterization of genes controlled by the quorum-sensing transcription factor ComA in Bacillus subtilis. *Mol. Microbiol.* **57**, 1159–1174 (2005).
72. V. Molle, *et al.*, The Spo0A regulon of Bacillus subtilis. *Mol. Microbiol.* **50**, 1683–1701 (2003).
73. K. Kobayashi, Gradual activation of the response regulator DegU controls serial expression of genes for flagellum formation and biofilm formation in Bacillus subtilis. *Mol. Microbiol.* **66**, 395–409 (2007).
74. R. A. Chong, H. Park, N. A. Moran, Genome Evolution of the Obligate Endosymbiont Buchnera aphidicola. *Mol. Biol. Evol.* **36**, 1481–1489 (2019).
75. K. Stelzner, N. Vollmuth, T. Rudel, Intracellular lifestyle of Chlamydia trachomatis and host–pathogen interactions. *Nat. Rev. Microbiol.* **21**, 448–462 (2023).
76. A. K. Hansen, N. A. Moran, Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 2849–2854 (2011).
77. C. Elwell, K. Mirrashidi, J. Engel, Chlamydia cell biology and pathogenesis. *Nat. Rev. Microbiol.* **14**, 385–400 (2016).
78. P. G. Madgwick, L. J. Belcher, J. B. Wolf, Greenbeard Genes : Theory and Reality. *Trends Ecol. Evol.*, 1–12 (2019).
79. R. Dawkins, *The Selfish Gene* (Oxford University Press, 1976).
80. T. W. Scott, A. Grafen, S. A. West, Multiple social encounters can eliminate Crozier's paradox and stabilise genetic kin recognition. *Nat. Commun.* **13** (2022).
81. A. Grafen, Do animals really recognize kin? *Anim. Behav.* **39**, 42–54 (1990).
82. V. Gligorijević, *et al.*, Structure-based protein function prediction using graph convolutional networks. *Nat. Commun. 2021 121* **12**, 1–14 (2021).
83. C. A. Ruiz-Perez, R. E. Conrad, K. T. Konstantinidis, MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. *BMC Bioinformatics* **22**, 1–16 (2021).
84. S. Lertampaiporn, *et al.*, PSO-LocBact: A Consensus Method for Optimizing Multiple Classifier Results for Predicting the Subcellular Localization of Bacterial Proteins (2019) https:/doi.org/10.1155/2019/5617153 (October 24, 2022).
85. M. A. Peabody, *et al.*, PSORTm: a bacterial and archaeal protein subcellular localization prediction tool for metagenomics data. *Bioinformatics* **36**, 3043 (2020).
86. F. Teufel, *et al.*, SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).
87. L. de Nies, *et al.*, PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome* **9**, 1–14 (2021).
88. V. Eichinger, *et al.*, EffectiveDB-updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. *Nucleic Acids Res.*

44, 669–674 (2016).
89. A. Belcour, *et al.*, Metage2metabo, microbiota-scale metabolic complementarity for the identi1cation of key species. *Elife* **9**, 1–38 (2020).
90. B. I. Cantarel, *et al.*, The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res.* **37**, 233–238 (2009).
91. S. Pollak, *et al.*, Public good exploitation in natural bacterioplankton communities. *Sci. Adv.* **7**, 1–11 (2021).
92. G. Beauclair, *et al.*, DI-tector: Defective interfering viral genomes' detector for next-generation sequencing data. *RNA* **24**, 1285–1296 (2018).
93. S. Sotcheff, *et al.*, ViReMa: A Virus Recombination Mapper of Next-Generation Sequencing data characterizes diverse recombinant viral nucleic acids. *bioRxiv*, 2022.03.12.484090 (2022).
94. Y. Sun, *et al.*, A specific sequence in the genome of respiratory syncytial virus regulates the generation of copy-back defective viral genomes. *PLOS Pathog.* **15**, e1007707 (2019).
95. L. J. Payne, *et al.*, PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. *Nucleic Acids Res.* **50**, W541–W550 (2022).
96. F. Tesson, *et al.*, Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.* **13** (2022).
97. L. Walker, Loss of altruism in the social amoeba Dictyostelium discoideum is associated with the G protein-coupled receptor grlG. *bioRxiv* **7227**, 0–3 (2022).