Social Interactions in Bacteria Mediated by Bacteriocins and Horizontal Gene Transfer

Joshua L. Thomas

St Hugh's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2023

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.

Joshua Thomas, Hilary term 2023

Acknowledgements

The irony of sacrificing my own social life to study that of bacteria is not lost on me. But these past seven years of living in Oxford have been nothing short of a dream. Ever since my first visit as a small boy on a coach from Penyrheol Comprehensive, I knew I had to come to this university. What I didn't realise then, was that three years of undergraduate study wasn't going to be enough. Thankfully, these additional four years of graduate study have just about hit the spot (for now). My love for Oxford will never fade.

There are many people that this DPhil thesis would not have been possible without. I'd like to start by thanking my incredible supervisors for everything they've done for me, academically and personally. Ashleigh, thanks for taking me in under your Celtic wing, for uncovering my love of social evolution as an undergraduate, and for teaching me what it means to be a scientist. Mel, I'll never forget the day we first met in the Missing Bean; thank you for always being there since, for teaching me how to hold a pipette, and for all the bananas and nuts. Danny, thank you for always coming up trumps when I needed you; I always loved our meetings. And an honorary mention to Stu, I never thought I'd find someone that loves Tom Jones as much as I do. A special thank you also goes to Felicity, without whom, Oxford would have remained a distant dream.

I'd like to thank the rest of the Griffin/West Group for always providing constant discussion, support, laughs, ATSs, pub trips, and for embracing my unhealthy obsession with Greggs. I'd particularly like to thank Anna for being an amazing friend and collaborator, working together has been so fun and insightful; both Anna & Becca for being the best conference, pub, and night out buddies; and Ming for all the laughs, beers, and meetings. Thanks also to the following people for providing invaluable input, feedback, and support during my DPhil: Alper, Saran, Laurie, Tom, Mati, Asher, David, Lois, Divjot, Natalia, Andrei, Jake, Connor, Elisa, Erik, James, and Sumali, and thanks to Craig MacLean and Kayla King for insightful comments throughout my DPhil.

Thank you to the ERC, CSEF, and Department of Biology for funding my DPhil, I am extremely grateful for your support.

It’s also safe to say that completing this DPhil would also not have been possible without the support of my friends and family. Thank you to the Swansea lads (Rhid, Ben, Alex, Mylsie, and Staff) for being ever-present, you all mean the world to me; JCFS (Paddy, Charlie, Theo, Will, Henry, Sasha, Ed, Ben) for the Best Day Ever, Hvar 1.0-5.0, and introducing me to the wonder that is croquet; JesuBio (Ryan, Eve, Elin) for everything during the first three years in Oxford; Dan for fighting with me side-by-side to protect the Cymraeg in this unknown land; the Housies (Reza, Serte, Annie, Leonie, Vicky) for being the best housemates I could ask for; my fellow Thirsty Swans (Amartya, JC, Florence, Ed, Alex, Matt, Avani, Alice, Caroline, Jake, André, Tom) for countless formals, TFSs, and brunches, and for generally making the past four years a joy; my fellow Darwinian Demons (Mark, Victor, Tom, Louis, Christian) for being the remarkable lads that you are; and Ben for that fateful Aperol Spritz. I feel so lucky to have all of you in my life.

I’d also like to say a special thank you to all of my incredible family. Thank you to my aunties, uncles, cousins, and grandparents for all of your love, wisdom, and kindness. Thank you to my big bro Tom, Amy, Noah, and Fern, for always putting a smile on my face and keeping me going. Thank you to my Dad, and my little sister Ava, for your constant love and support, and for making home my favourite place to be. Thank you to my Mum, for your insurmountable love and guidance; you always have been, and always will be, my inspiration. Finally, thank you to Anwar, for being by my side every step of the way; you are my rock, and you are my light.

Publications & contributions

Publications arising from this thesis include:

- **Thomas, J.L.,** Liu, M., Bray, J.E., Crook, D.W., Wilson, D.J., Griffin, A.S. and Ghoul, M. "The evolution and ecology of bacteriocin-mediated competition in *Staphylococcus aureus* colonising human hosts". In preparation for submission to The *ISME* Journal.

- Dewar, A.E.*, **Thomas, J.L.*,** Scott, T.W., Wild, G., Griffin, A.S., West, S.A. and Ghoul, M., 2021. Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range. *Nature ecology & evolution*, *5*(12), pp.1624-1636.

*** = co-first author**

Chapter 2

The original *Staphylococcus aureus* nasal carriage collection and clinical data was obtained by the Crook/Modernising Medical Microbiology group (Oxfordshire Research Ethics Committee B reference number 08/H0605/102). Source BioScience (Nottingham, UK) performed the Sanger sequencing that I used to *spa*-type my isolate collection. The rest of the work in this chapter is my own, supervised by Prof. Ashleigh Griffin, Dr. Melanie Ghoul, and Prof. Daniel Wilson, who provided input on the isolate selection criteria and comments on the chapter.

Chapter 3

The following manuscript resulted from this chapter:

- **Thomas, J.L.,** Liu, M., Bray, J.E., Crook, D.W., Wilson, D.J., Griffin, A.S. and Ghoul, M. "The evolution and ecology of bacteriocin-mediated competition in *Staphylococcus aureus* colonising human hosts". In preparation for submission to The *ISME* Journal.

The Wellcome Trust Centre for Human Genetics (WTCHG) (Oxford, UK) performed whole-genome sequencing, Dr. James Bray assembled the resulting whole-genome sequences, and fellow DPhil student Ming Liu provided input on some aspects of data analysis. The rest of the work in this chapter is my own, supervised by Prof. Ashleigh Griffin, Dr. Melanie Ghoul, and

Prof. Daniel Wilson, who provided input on experimental design, interpretation of results, and comments on the manuscript.

The work in this chapter has been combined into a single manuscript for submission as part of my integrated thesis format. However, if the work in this chapter were to be submitted as part of a traditional, non-integrated thesis, it would form four separate chapters:

1. "What is the prevalence and identity of bacteriocin producing strains in natural populations of *Staphylococcus aureus* from the human nasal cavity?". This involved screening all isolates in the collection for their ability to inhibit an indicator species for bacteriocin production, *Cellulomonas fimi,* and performing phylogenetic analysis on the resulting data.

2. "Which competitors do *Staphylococcus aureus* target with bacteriocins?". This involved screening most isolates in the collection against ecologically-relevant intraspecific and interspecific competitors and determining the prevalence of each type of inhibition.

3. "Are inhibitory strains associated with a competitive benefit in natural populations of *Staphylococcus aureus*?" This involved mapping phenotypic inhibition profiles to longitudinal strain dynamics data and testing for significant associations between inhibitory activity and short- or long-term competitive benefits.

4. "What bacteriocins do *Staphylococcus aureus* carry to perform inhibitory activity in natural populations?". This involved genomic approaches to identify bacteriocin gene clusters associated with phenotypic inhibitory activity in S. *aureus.*

Chapter 4

I completed Chapter 4 in collaboration with fellow DPhil student (at the time of collaboration), Dr. Anna Dewar. The resulting work was published as a paper in *Nature Ecology & Evolution*:

- Dewar, A.E.*, **Thomas, J.L.*,** Scott, T.W., Wild, G., Griffin, A.S., West, S.A. and Ghoul, M., 2021. Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range. *Nature ecology & evolution*, *5*(12), pp.1624-1636.

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

In the appendix of my thesis, I include Supplementary Sections 1, 2, and 3 for this paper, which my work contributed to, but were written and put together by A.E.D.

Abstract

Bacteria are highly social organisms that frequently engage in cooperative and competitive interactions to successfully survive and reproduce. Examples include cell-to-cell communication, nutrient scavenging, and chemical warfare. However, the vast majority of our understanding of bacterial sociality has come from the laboratory strains of a small number of gram-negative social evolution model organisms, such as *Pseudomonas spp.* and *Escherichia coli*. In my thesis, I aim to expand our understanding of bacterial sociality in natural populations and further across the bacterial tree of life. I do this using two different approaches. Firstly, I use laboratory experiments and sequence analysis to study the evolution and ecology of bacteriocin-mediated competition in natural *S. aureus* populations, sampled as part of a carriage study on human nasal passages. Theory and laboratory experiments to date have provided extensive evidence that bacteriocin production plays a key role in determining the competitive dynamics of bacterial strains, however evidence from natural populations to support this hypothesis is lacking. I find that inhibitory strains were associated with the propensity to displace competing strains from the nasal cavity, which occurs despite inhibitory activity not being displayed by the majority of strains and targeting interspecific over intraspecific competitors. I also provide evidence for the genetic underpinnings of bacteriocin activity, by identifying five bacteriocin gene clusters associated with inhibition. Secondly, I use a comparative approach to study the role of horizontal gene transfer in stabilising cooperation across bacteria. Bacterial cooperation is typically mediated by the secretion of extracellular public goods, which are costly molecules that provide a fitness benefit to neighbouring cells. Cooperation can be destabilised by the invasion of selfish ‘cheats’ that reap the benefit of public good production without paying a cost. It is widely accepted that horizontal gene transfer, especially *via* plasmids, can allow cooperators to ‘re-infect’ cheats with the gene for a cooperative trait, thus stabilising cooperation. Although theoretical and experimental studies have provided evidence to support this hypothesis, a comprehensive genomic study that controls for phylogenetic non-independence across species remains to be conducted. The results from our analysis of plasmid genes from 51 diverse bacterial species do not support the cooperation hypothesis across bacteria and are instead supportive of environmental variability as a determining factor in the relationship between horizontal gene transfer and extracellular proteins. Taken together, this thesis provides a body of work that emphasises the importance of testing predictions from theoretical and laboratory experiments in natural populations, and across diverse species.

Contents

Chapter 1. Introduction

Social evolution theory

Life is social. Living organisms frequently engage in social interactions with other individuals of the same and different species. A social interaction can be defined as a behaviour performed by one individual (an actor), that affects the fitness of another individual(s) (a recipient), and has evolved, at least in part, for that fitness effect (Hamilton, 1964; West *et al.,* 2007c; Bourke, 2011). Social interactions can affect the actor and recipient's lifetime direct fitness (total number of offspring produced) in different ways, allowing them to be grouped into four broad categories (Hamilton, 1964) (Table. 1). 'Mutually beneficial' interactions increase the fitness of both the actor and recipient (+/+); 'altruistic' interactions increase the fitness of the recipient, but decrease the fitness of the actor (-/+); 'selfish' interactions increase the fitness of the actor, but decrease the fitness of the recipient (+/-); 'spiteful' interactions decrease the fitness of both the actor and recipient (-/-) (Hamilton, 1964; West *et al.,* 2007c). The interactions can be grouped based on their effect on the recipient's fitness: mutual benefit and altruism are the two forms of cooperation and act to increase recipient fitness; selfishness and spite are the two forms of conflict/competition and act to decrease recipient fitness (West *et al.,* 2007c; Foster, 2010). These categories represent all types of social interactions across the entire tree of life.

		Effect on recipient	
		+	-
Effect on actor	+	Mutual benefit	Selfishness
	-	Altruism	Spite

Table 1. Hamilton's classification of social interactions. Adapted from West *et al.* (2007c).

The observation of altruism in the natural world could not be explained by Darwin's theory of natural selection and posed a challenging question for evolutionary biology: why would organisms be selected to reduce their own fitness to increase that of others? William D. Hamilton explained how altruism can be favoured in his seminal 1964 paper (Hamilton, 1964). Hamilton recognised that producing your own offspring (direct fitness) was not the only way to propagate genes into the next generation. Instead, given that individuals can share genes with other individuals to greater or lesser extents, increasing the reproductive success of individuals that share a high genetic similarity to you (indirect fitness) would also propagate your genes into the next generation. Therefore, instead of maximizing their direct fitness, Hamilton proposed that organisms are selected to maximise their 'inclusive fitness', which combines both direct and indirect components, and forms the basis of Hamilton's 'Inclusive Fitness Theory' (Hamilton, 1964).

Hamilton presented a rule for predicting the evolution of a social behaviour, known as 'Hamilton's rule': $rb - c > 0$, whereby 'r' is the coefficient of relatedness, representing the probability of two individuals sharing an allele for a social behaviour relative to the population average, 'b' is the benefit, in terms of the number of offspring gained by the recipient resulting from the social interaction, by 'c' is the cost, in terms of the number of offspring lost by the actor resulting from the social interaction (West *et al.,* 2007c). Hamilton's inclusive fitness theory (1964) was a monumental step for evolutionary biology, by extending Darwin's theory of natural selection to explain the evolution of altruism.

Sociomicrobiology

Despite social evolution theory initially focusing on the social behaviour of animals (West *et al.,* 2007b, 2007c; Bourke, 2011; Davies *et al.,* 2012), over the last two to three decades we have come to understand that microbes too are highly social organisms that engage in a variety of cooperative and competitive interactions, giving rise to the field of ‘sociomicrobiology’ (Griffin *et al.,* 2004; West *et al.,* 2006, 2007a; Nadell *et al.,* 2008, 2016; Kümmerli *et al.,* 2009a, 2009b; Foster, 2010; Mitri & Foster, 2013; Ghoul *et al.,* 2014; Ghoul & Mitri, 2016; Butaitė *et al.,* 2017; Kramer *et al.,* 2020; Figueiredo *et al.,* 2022). For example, bacteria can perform cell-to-cell communication (Brown & Johnstone, 2001; Diggle *et al.,* 2007), collectively scavenge for nutrients (Griffin *et al.,* 2004; Willsey & Wargo, 2015; Kramer *et al.,* 2020), and kill one another using mechanisms of chemical warfare (Riley & Gordon, 1996, 1999; Riley & Wertz, 2002; Kerr *et al.,* 2002; Riley & Chavan, 2007; Cornforth & Foster, 2013; Ghoul & Mitri, 2016; Granato *et al.,* 2019). Bacteria live in complex communities, whereby due to their clonal reproduction *via* binary fission, they frequently interact with other genetically identical cells from the same species (i.e., cells of the same strain). However, given the sheer abundance and diversity in most habitats, they also frequently interact with non-identical cells from the same and different species (West *et al.,* 2006; Nadell *et al.,* 2008, 2016; Mitri *et al.,* 2011; Xavier, 2011; Mitri & Foster 2013). Social behaviours, both cooperative and competitive, are known to be important for the survival and reproduction of microbes in these complex environments (West *et al.,* 2006; Nadell *et al.,* 2008, 2016; Xavier, 2011; Mitri & Foster, 2013; Ghoul & Mitri, 2016; Dragoš *et al.,* 2018; Figueiredo *et al.,* 2022).

Microbial cooperation

Microbes are known to be capable of cooperating in a number of ways. For example, microbes can cooperate by self-limiting their use of a particular resource to maximise their collective resource-use efficiency (Pfeiffer *et al.,* 2001; MacLean & Gudelj, 2006). Microbes can also cooperate during dispersal, for example, some cells of the slime mould *Dictyostelium discoideum* are known to altruistically sacrifice their ability to reproduce by forming a nonviable stalk, which facilitates the dispersal of other cells as spores (Strassmann *et al.,* 2000; Foster *et al.,* 2004). Microbes also often mediate cooperation by producing 'public goods' (West *et al.,* 2006), which represents the form of microbial cooperation that has received the largest amount of research attention to date. Public goods are molecules secreted into the environment by a cooperator cell that can provide a fitness benefit to the producing cell and neighbouring cells, but at a cost to the cooperator cell. Examples include iron-scavenging 'siderophores' that bind to iron and facilitate cellular uptake (Griffin *et al.,* 2004; Kramer *et al.,* 2020), degradative enzymes, such as proteases and lipases, that facilitate the breakdown of larger macromolecules into smaller digestible molecules (Willsey & Wargo, 2015), and signalling molecules, such as those involved in quorum-sensing systems that allow microbes to collectively upregulate the expression of many traits under conditions of high cell-density (Brown & Johnstone, 2001; Diggle *et al.,* 2007; Novick & Geisinger, 2008).

However, microbial cooperation, including public good production and other forms of cooperation, is susceptible to a selfish 'cheating' behaviour, by cells that reduce or stop their contribution towards the cost of cooperation, but continue to reap the cooperative benefit (Griffin *et al.,* 2004; West *et al.,* 2006; Foster, 2010; Ghoul *et al.,* 2014; Butaitė *et al.,* 2017). Once cheat cells emerge, they can increase in frequency, destabilise cooperation, and cause population collapse, due to their inability to perform the cooperative behaviour required for survival and reproduction (West *et al.,* 2006). A key aim in the field of social evolution is to understand how cooperation can be maintained in the face of cheating.

The stabilisation of cooperation

The stabilisation of cooperative behaviours that provide a direct fitness benefit is easier to explain and can occur either due to the shared interest of the cooperators (i.e., non-enforcement) or the enforcement of cooperation *via* one or more enforcement mechanisms, such as reward, punishment, sanctioning, reciprocity, or policing (West *et al.,* 2007b). The stabilisation of cooperative behaviours with an indirect fitness component can be more difficult to explain. This can occur when individuals cooperate with other individuals that share the gene for cooperation (West *et al.,* 2006). One mechanism by which organisms can achieve this is 'kin recognition', a form of discriminate cooperation, whereby individuals recognise relatives and preferentially cooperate with them (West *et al.,* 2007b). For example, in bacteria, strains of *Bacillus subtilis* have been shown to perform collective swarming behaviours when co-existing with kin, but not with non-kin (Stefanic *et al.,* 2015). Kin recognition and discrimination can also occur by individuals preferentially directing competitive behaviours away from relatives and towards non-relatives. This is exemplified by many species of bacteria that produce antimicrobial toxins, called 'bacteriocins', to kill and inhibit the growth of competing bacterial strains and species, but do not target clonemates from the same strain, due to them carrying a cognate gene conferring protective immunity (Riley & Gordon, 1999; Riley & Wertz, 2002) (see below for a more detailed introduction to bacteriocins).

In addition to discriminate cooperation, organisms can also perform indiscriminate cooperation, whereby they cooperate with all individuals in a given area (West *et al.,* 2007b). The production of public goods by microbes can represent an example of indiscriminate cooperation, as the public good benefit is often gained by all neighbouring cells (West *et al.,* 2006). Limited dispersal can stabilise indiscriminate cooperation, as it can create a high spatiogenetic population structure, which increases the likelihood of interactions occurring between close relatives (West *et al.,* 2006; Kümmerli *et al.,* 2009a; Mitri & Foster, 2013). This is thought to be a particularly important and widespread mechanism of stabilising cooperation in bacteria, because their asexual mode of reproduction *via* binary fission often causes bacterial cells to be positioned in close proximity to genetically identical clones, where r=1 (West *et al.,* 2006; Kümmerli *et al.,* 2009a). The spatiogenetic structure of a bacterial population is also affected by the structure of the habitat in which it lives: habitats with relatively high structure (e.g., soil-based environments) can allow bacterial cells to stay in close proximity and cooperate with their clonemates, as opposed to habitats with relatively low structure (e.g., water-based environments), which can increase mixing between kin and non-kin cells (Kümmerli *et al.*, 2009b, 2014). High habitat structure also prevents the extensive diffusion of public goods, allowing them to be retained by cooperator cells and reducing the fitness benefit associated with cheating (Kümmerli *et al.,* 2009b, 2014).

It has also been proposed that the genetic architecture of cooperative traits could help to stabilise cooperation and prevent cheating in some cases. For example, pleiotropy, whereby one gene affects multiple phenotypic traits, represents one possible mechanism (Foster *et al.,* 2004; Mitri & Foster 2016). This is because it may not be possible for a mutant cheat strain, which has lost the ability to perform a cooperative trait, to evolve if the gene for cooperation also codes for an essential cellular function. An example of the importance of pleiotropy in stabilising microbial cooperation has been identified in the slime mould *Dictyostelium discoideum*, whereby a *dimA* mutant strain, which was predicted to act as a cheat by ignoring the signalling molecule DIF-1 and therefore producing fewer non-reproductively viable stalk cells, compared to reproductive spore cells. However, due to the pleiotropic linkage between stalk and spore formation, *dimA* mutants were found to be excluded from spores in aggregates containing wild-type cells, thus removing the advantage of cheating (Foster *et al.,* 2004).

Another proposed mechanism thought to stabilise microbial cooperation is horizontal gene transfer (HGT) (Smith, 2001), whereby microbes exchange genetic material with members of the same generation (Thomas & Nielsen, 2005; Hall *et al.,* 2020). HGT is known to be extremely prevalent across microbes, such as bacteria, and can take three different forms: transformation occurs when competent bacteria take up environmental DNA and incorporate it into their genome (Griffith, 1928). Transduction involves the transfer of bacteriophages, viruses that infect bacteria by integrating into bacterial genome. Upon cell exit, bacteriophage can incorporate bacterial DNA and transfer it to other cells (Zinder & Lederberg, 1952). Conjugation occurs when bacteria inject DNA into an adjacent cell using a tube-like pilus structure (Lederberg & Tatum, 1946; Smillie *et al.,* 2010). Conjugation often involves the transfer of small, circular rings of self-replicating DNA called plasmids, many of which encode their own conjugative apparatus (Smillie *et al.,* 2010). Using HGT, particularly plasmid conjugation, cooperator cells are thought to be capable of 're-infecting' cheat cells with the gene for public good production, by inserting it into the neighbouring cell, thus increasing relatedness to r=1 at the social locus and stabilising cooperation (Smith, 2001; Nogueira *et al.,* 2009, 2012; Mc Ginty *et al.,* 2011, 2013; Dimitriu *et al.,* 2014).

Microbial competition

Social evolutionary theory predicts that while cooperation is more important within genetically identical strains, competition will generally prevail between different strains and species (West *et al.,* 2006; Foster & Bell, 2012; Mitri & Foster, 2013). The limited nature of resources and high cell densities in microbial communities results in fierce competition between genotypes (Foster & Bell, 2012; Mitri & Foster, 2013; Ghoul & Mitri, 2016). Microbial competition can be categorised into two broad types: 'exploitative' (or 'indirect') competition and 'interference' (or 'direct') competition (Ghoul & Mitri, 2016). Exploitative competition occurs when a microbe gains access to a limited resource, and in doing so, prevents access to this resource by a competitor. For example, some microbes, such as *Saccharomyces cerevisiae*, can undergo metabolic shifts to perform fermentation instead of respiration in the presence of oxygen, which increases growth rate to utilise resources faster than their competitors, despite decreasing overall yield (Pfeiffer *et al.,* 2001; MacLean & Gudelj, 2006). Microbes can also secrete extracellular molecules, such as iron-scavenging siderophores, to sequester nutrients and prevent uptake by competitors (Griffin *et al.,* 2004; Kümmerli *et al.,* 2009b; Ghoul & Mitri, 2016; Gu *et al.,* 2020; Kramer *et al.,* 2020; Figueiredo *et al.,* 2022; Kümmerli, 2022)

Interference competition occurs when microbes directly damage or disrupt the physiology of their competitors, preventing them from accessing a common resource (Ghoul & Mitri, 2016; Granato *et al.,* 2019). Microbes, such as bacteria, have evolved diverse forms of chemical weaponry to mediate interference competition, which can be broadly categorised into two types: 'contact-dependent' and 'contact-independent' killing mechanisms (Granato *et al.,* 2019). Contact-dependent mechanisms involve the delivery of inhibitory molecules by a focal cell to a directly adjacent target cell. Examples include the Type IV and VI systems in gram-negative bacteria (MacIntyre *et al.,* 2010; Basler *et al.,* 2012; Souza *et al.,* 2014) and Type VII system in gram-positive bacteria (Ulhuq *et al.,* 2020), which all use a syringe-like stabbing structure to breach the cellular membrane of their competitors and deliver toxins (Granato *et al.,* 2019). Contact-independent mechanisms involve the extracellular release of antimicrobial molecules that diffuse to target competitors in the environment and inhibit their survival and/or growth (Ghoul & Mitri, 2016). Examples include antimicrobial toxins, such as antimicrobial peptides and proteins called 'bacteriocins' (Riley & Gordon, 1999; Gardner *et al.,* 2004; Ghoul *et al.,* 2015; Bruce *et al.* 2017), 'phages' that are viruses integrated into bacterial genomes that can kill competitors upon release (Haaber *et al.,* 2016), and 'tailocins' which are derived from phages but do not contain the capsular head that carries nucleic acid (Dorosky *et al.,* 2017).

Bacteriocin-mediated competition

Bacteriocins are classically defined as ribosomally-synthesised antimicrobial peptides and proteins (Dorit *et al.*, 2016), however more recently the definition has been extended to include other groups of small molecules with antimicrobial activity not synthesised by the ribosome, such as non-ribosomally synthesised peptides (NRPs) (Heilbronner *et al.*, 2021). Bacteriocins are thought to be one of the most common types of chemical weaponry, with the vast majority of well-studied bacterial species being identified to carry one or more in their genome (Klaenhammer, 1993; Granato *et al.*, 2019). It is worth noting, however, that certain groups of understudied bacteria with distinct ecologies may not require bacteriocins. For example, the Candidate Phyla Radiation, which constitutes approximately 25% of all bacterial diversity, is a group of extremely small bacteria that live as symbiotic parasites on the outer surface of other bacteria, and may not require bacteriocins or chemical weaponry at all (Granato *et al.*, 2019). However, more generally, bacteriocins have been identified as a prominent and widespread strategy to mediate competition in bacteria. The prevalence of bacteriocins within species can also be extremely high: in natural populations of *Pseudomonas aeruginosa* and *E*scherichia *coli*, two species with particularly well-studied bacteriocins, up to ~100% and ~70% of strains can carry bacteriocins (Gordon *et al.*, 1998; Ghoul *et al.*, 2015), respectively, with some strains being capable of carrying cocktails of multiple bacteriocins (Gordon & O'Brien, 2006).

A huge diversity of bacteriocins have been identified in bacteria, each with different structural and biochemical properties (Riley & Chavan, 2007; Drider & Rebuffat, 2011; Li & Rebuffat, 2020). These bacteriocins can have diverse mechanisms of action once they enter the target cell, such as interfering with core metabolism, damaging and permeabilising cellular membranes, and inhibiting cell wall biosynthesis (Drider & Rebuffat, 2011; Cotter *et al.*, 2013; Granato *et al.*, 2019). Bacteriocins are also known to be encoded on both the bacterial chromosome and on plasmids (Giambiagi-Marval *et al.,* 1990; Riley & Wertz, 2002; Heilbronner *et al.,* 2021). In the simplest cases, bacteriocin production is coded for by a single toxin production gene, and producing bacteria avoid self-harm by encoding a corresponding immunity gene in close proximity that neutralizes the toxin (Riley & Gordon, 1999; Drider & Rebuffat, 2011). Bacteriocins are generally considered to have narrow-spectrum activity in that they target conspecific strains (Riley & Gordon, 1999; Riley & Wertz, 2002; Riley & Chavan, 2007; Cotter *et al.,* 2013), but can also have broad-spectrum activity against other species (Janek *et al.,* 2016; Li & Rebuffat, 2020; Heilbronner *et al.,* 2021).

Bacteriocins are also known to be important from an evolutionary and ecological perspective. A range of conditions are predicted to select for bacteriocin production. For example, high cell-density and abundance (Adams *et al.,* 1979; Chao & Levin, 1981; Dorit *et al.,* 2016), intermediate relatedness (Gardner *et al.,* 2004), high metabolic overlap with competitors (Bruce *et al.,* 2017), and low/intermediate nutrient conditions (Granato *et al.,* 2019) have all been shown or predicted to select for bacteriocin production. Once evolved, many bacteriocins are known to be tightly regulated depending on environmental conditions (Cornforth & Foster, 2013; Niehus *et al.,* 2021). For example, upregulation has been repeatedly detected under conditions of cellular damage and nutrient stress. This has been explained as a form of ‘competition sensing’ that allows bacteria to save on the cost of bacteriocin production by only engaging in warfare when under attack (Cornforth & Foster, 2013). Bacteriocins can also have important evolutionary and ecological consequences, by playing a key role in determining competitive strain dynamics (Riley & Gordon, 1996, 1999; Gordon & Riley, 1999; Riley & Wertz, 2002; Kerr *et al.,* 2002; Gardner *et al.,* 2004; Krikup & Riley, 2004; Majeed *et al.,* 2011; Kommineni *et al.,* 2015; Kawada-Matsuo *et al.,* 2016; Zipperer *et al.,* 2016; Lehtinen *et al.,* 2022). For example, bacteriocins can provide producers with a competitive advantage against sensitive cells, allowing them to defend niches from invasive competitors (Zipperer *et al.,* 2016), invade new niches (Kommineni *et al.,* 2015) or remove co-existing strains (Majeed *et al.,* 2011).

Studying sociomicrobiology in natural populations and across diverse bacteria

Although microbes from phylogenetically diverse groups have been shown to display social traits, our understanding of sociomicrobiology – cooperative and competitive interactions – stems from theoretical studies or studies of laboratory strains from a small number of gram-negative species, such as *Pseudomonas spp.* and *E. coli,* that have become model organisms to study social evolution. There are clear advantages to this approach: highly controlled and simplified experimental set-ups allows the causes and consequences of social traits to be isolated with relative ease, and focusing on a small number of model organisms allows the development of detailed methodologies to study the phenotypes and genotypes of social traits, with the aim of identifying broadly applicable principles.

However, if our aim is to understand the evolutionary and ecological role of social traits in nature, testing predictions from theoretical and laboratory experiments in natural populations, and across diverse species, is a crucial next step. In addition to confirming theoretical and laboratory predictions, the additional biological complexity present in natural communities means that natural studies can often also provide results that do not completely align with prediction. These studies can be particularly important to further refine theoretical and laboratory experiments, and in turn, generate further testable predictions for social traits in nature. While some natural studies to date have provided great insight into the evolution and ecology of bacterial social traits, including studies that focus on cooperation (Kümmerli *et al.,* 2014; Simonet & McNally, 2021; Belcher *et al.,* 2022), competition and cheating (Hawlena *et* *al.*, 2012; Kinkel *et al.*, 2014; Abrudan *et al.*, 2015; Ghoul *et al.*, 2015; Bruce *et al.*, 2017; Butaitė *et al.*, 2017, 2018; Kraemer *et al.*, 2017; Gu *et al.*, 2020), or both cooperation and competition (Foster & Bell, 2012; Figueiredo *et al.*, 2022), we are only just scratching the surface, and more studies of this kind are required in a wider range of species.

Thesis aims and approaches

The broad aims of my thesis are to expand our understanding of bacterial sociality in natural populations of bacteria and further across the bacterial tree of life. Firstly, I do this by studying natural populations of a species understudied from a social evolution perspective: *Staphylococcus aureus.* In particular, I focus on the evolution and ecology of bacteriocin-mediated competition and its role in determining longitudinal strain dynamics in *S. aureus*. Secondly, I use a comparative approach to study cooperation across many diverse bacterial species, and test the hypothesis that horizontal gene transfer stabilises cooperation across bacteria. Here, I provide general background on both approaches and put them into context with one another, before providing a further detailed introduction in each of their respective chapters.

Lab study system - *Staphylococcus aureus* populations living in the human nasal passage

S. aureus is a gram-positive bacterium well-known for its ability to cause diverse infections in humans and other animals, including skin and soft tissue infections (SSTIs), septicaemia, endocarditis, and toxic shock syndrome (Young *et al*., 2012; Tong *et al.*, 2015; Somerville, 2016; Turner *et al*., 2019). Despite long being identified as an opportunistic pathogen that only infected immunocompromised hosts, *S. aureus* has more recently been identified to cause infection in healthy humans (Tong *et al.*, 2015; Turner *et al*., 2019). *S. aureus* therefore poses a large healthcare problem in both hospital and community settings, which has been exacerbated by the ability of *S. aureus* to evolve resistance to numerous antibiotics, leading to the emergence of multi-drug resistant strains, such as methicillin-resistant *Staphylococcus aureus* (MRSA) (Otto, 2013; Lee *et al.*, 2018; Turner *et al*., 2019).

Despite being understudied from a social evolution perspective, *S. aureus* has been observed to display multiple social traits, including public goods from well-studied from social evolution model organisms. For example, *S. aureus* is known to produce siderophores called 'staphyloferrins' to collectively sequester and take up iron (Beasley *et al.*, 2009; Cheung *et al.*, 2009). *S. aureus* is also known to use the accessory gene regulator (*agr*) quorum sensing (QS) system by producing autoinducing peptide (AIP) signalling molecules to coordinate the collective expression of virulence factors at high cell-density (Novick & Geisinger, 2008). A study by Pollitt *et al.* (2014) demonstrated that *agr* signalling in *S. aureus* is a cooperative trait that provides a benefit to local cells in a wax moth larvae infection model. Moreover, they show that although this signalling system can be exploited by cheats that do not produce or respond to *agr* AIPs, higher levels of within-host relatedness select for cooperative QS in *S. aureus*. In addition to well-studied social traits, *S. aureus* also produces cooperative public goods more specific to its biology, such as coagulases that promote clot formation during infection (Trivedi *et al.*, 2018). To mediate competition, *S. aureus* has also been identified to have evolved multiple mechanisms of chemical warfare, including bacteriocins (Netz *et al.*, 2002; Daly *et al.*, 2010; Janek *et al.*, 2016; Kawada-Matsuo *et al.*, 2016; de Freire Bastos *et al.*, 2020; Newstead *et al.*, 2020), a type-VII secretion system (Ulhuq *et al.*, 2020), and phage (Haaber *et al.*, 2016).

S. aureus is also capable of asymptomatically colonising many parts of the human body, such as the skin, gut, and throat (Somerville, 2016), but its primary human-associated habitat is the nasal cavity (Wertheim *et al.*, 2005; Golubchik *et al.*, 2013). Within the nasal cavity, *S. aureus* has been identified in multiple nasal microenvironments (e.g., anterior naris, middle meatus, sphenoethmoidal recess), but predominantly resides in the anterior nares by attaching to nasal epithelial cells (Yan *et al*., 2013). Approximately 30% of healthy adults are colonised by *S. aureus* at a given point in time, with some individuals being persistently colonised, some intermittently colonised, and others appearing to be resistant to colonisation (Wertheim *et al.*, 2005; Miller *et al.*, 2014). Asymptomatic nasal carriage has been identified as an important risk factor for subsequent *S. aureus* infection (Wertheim *et al*., 2005), with approximately 80% of infections being caused by a strain already colonising the nasal cavity (Von Eiff *et al*., 2001). Despite this, *S. aureus* nasal colonisation has been drastically understudied compared to infection itself, and the key determinants of *S. aureus* nasal colonisation success are still unclear (Golubchik *et al.,* 2013; Krismer *et al*., 2017; Otto, 2020).

Life in the nasal cavity presents bacteria with many challenges. Nutrients are known to be particularly limiting compared to other parts of the human microbiome (Krismer *et al.*, 2014, 2017). Space is also limited and bacteria are known to produce adhesins to bind to attachment sites on nasal epithelial cells, in part to avoid removal from the nasal cavity by mucus flow (Schade & Weidenmaier, 2016; Geoghegan & Foster, 2017; Krismer *et al.,* 2017). Although it is known that human hosts respond to *S. aureus* nasal colonisation, such as by the production of protective molecules (Brown *et al.,* 2014), multiple studies have shown that host genetic factors are not the only determinants colonisation success in the nasal cavity (Andersen *et al.*, 2013; Liu *et al.,* 2015). The nasal microbiome generally has a highly variable but low cell abundance (Liu *et al.,* 2015) and an intermediate level of species diversity compared to other parts of the human microbiome, with the majority of species usually come three phyla: Actinobacteria, Firmicutes and Proteobacteria (Krismer *et al.,* 2017). Species composition in the nasal cavity can usually be categorised into one of seven different 'community state types' (CSTs), whereby one species dominates with a particularly high proportional abundance (Liu *et al.,* 2015). For example, CST1 is dominated by *S. aureus* whereas CST3 is dominated by *Staphylococcus epidermidis* (Liu *et al.,* 2015). The evolutionary and ecological causes and consequences of different CSTs not known, and we are yet to fully understand the main competitors of each species in the nasal cavity (Krismer *et al.,* 2017). Like many host-associated habitats, we also lack a clear understanding population structure in the nasal cavity (Liu *et al.,* 2015; Krismer *et al.,* 2017).

To improve our understanding of *S. aureus* asymptomatic nasal carriage, a previous study by the Crook/Modernising Medical Microbiology group, Oxford created a large-scale *S. aureus* isolate collection, by longitudinally-sampling the nasal cavities of 571 participants at two-month intervals for up to a maximum of 90-months, yielding approximately ~4,500 *S. aureus*-positive samples (Young *et al*., 2012; Golubchik *et al*., 2013; Miller *et al*., 2014; Votintseva *et al*., 2014). The study also went to extreme lengths to capture the *S. aureus* strain diversity present within populations at each time point. The strengths of this study, including its large sample size, long sampling time frame, frequent sampling, and representation of strain diversity, make it an excellent system to study how social traits can influence evolutionary and ecological dynamics in natural populations of bacteria.

Bacteriocin-mediated competition in *S. aureus*

S. aureus have been identified to carry diverse types of bacteriocins from different classes (Bastos *et al.,* 2009; Newstead *et al.,* 2020). Around ten *S. aureus* bacteriocins, which are often generally referred to as 'staphylococcins', have been fully characterised to date (Newstead *et al.,* 2020). The best-studied examples include staphylococcin C55 from class I (Navaratna *et* *al.,* 1999; Kawada-Matsuo *et al.,* 2016) and the aureocins, such as aureocin A53 (Netz *et al.,* 2002; Nascimento *et al.,* 2012) and A70 (Netz *et al.,* 2001) from class II. Most *S. aureus* bacteriocins have been identified to target the cellular membrane, and can lead to pore formation, the disruption of membrane potential, and the leakage of cellular contents (de Freire Bastos *et al.,* 2020). *S. aureus* bacteriocins are often encoded on plasmids (Netz *et al.,* 2002; de Freire Bastos *et al.,* 2020), but can also been found on the bacterial chromosome (Daly *et al.,* 2010). In addition to achieving bacteriocin immunity by coding for a cognate bacteriocin immunity gene, *S. aureus* can also achieve immunity *via* other, less easily identifiable mechanisms, such as by modifying bacteriocin targets, the action of multicomponent ABC transporters, or other unknown intrinsic cellular properties (de Freire Bastos *et al.*, 2020). In general, the underlying genetics, mode of action, and regulation of *S. aureus* bacteriocins have been understudied compared to the bacteriocins of other species (Bastos *et al.,* 2009).

The vast majority of *S. aureus* bacteriocin-related studies have predominantly been from a medical or industrial perspective (de Oliveira *et al.,* 1998; Nascimento *et al.,* 2006; Coelho *et al.*, 2007; Bastos *et al.,* 2009; Okuda *et al.,* 2013; Fagundes *et al*., 2016; Newstead *et al.,* 2020; Fernández-Fernández *et al.,* 2022), with very little being known about their evolution and ecology. Studies have demonstrated that some *S. aureus* bacteriocins can inhibit conspecific strains, and that they can also display broad activity spectra against phylogenetically distant species (Giambiagi-Marval *et al.,* 1990; de Oliveira *et al.,* 1998; Nascimento *et al.,* 2006; Coelho *et al.,* 2007; Koch *et al.,* 2014; Kawada-Matsuo *et al.,* 2016; Newstead *et al.,* 2020). Broad-spectrum activity generally appears to be more common amongst gram-positive species compared to gram-negatives (de Freire Bastos *et al.,* 2020; Heilbronner *et al.,* 2021), but the explanation for this is unclear. The small number of studies that have systematically screened natural populations of *S. aureus* for bacteriocins have typically found that between ~10-30%

of isolates display inhibitory activity, with these studies usually focusing on infectious isolates from either humans or cattle, or environmental isolates (Giambiagi-Marval *et al.*, 1990; de Oliveira *et al.*, 1998; Gamon *et al.*, 1999; Nascimento *et al.*, 2002; Ceotto *et al.*, 2009; Fernández-Fernández *et al.*, 2022). One study by Janek *et al.* (2016) found that ten out of 19 (53%) *S. aureus* strains from the human nasal cavity displayed some degree of inhibitory activity interspecifically, but no inhibitory activity against the laboratory strain *S. aureus* Newman. However, a limitation of this study, and in the field more generally, is that bacteriocin activity was not tested against other naturally-occurring intraspecific competitors from the same environment. Studies also often use a small sample size of *S. aureus* isolates and strains, as in Janek *et al.* (2016), or treat isolates likely containing the same *S. aureus* strain as independent data points causing pseudoreplication (Fernández-Fernández *et al.*, 2022). Despite identifying *S. aureus* bacteriocin activity in natural populations, and laboratory experiments demonstrating that bacteriocins can allow producing strains to inhibit and outcompete sensitive strains (Kawada-Matsuo *et al.*, 2016), we are yet to understand the role of bacteriocins in natural populations of *S. aureus.*

In my thesis, I use laboratory- and sequence-based approaches to determine the prevalence, genomic identity, activity spectra, and consequences of *S. aureus* bacteriocins in the human nasal cavity from an evolutionary and ecological perspective, in particular whether they determine competitive strain dynamics and provide a competitive benefit.

Using a comparative approach to understand microbial sociality – HGT and cooperation

In addition to laboratory study, the omics revolution has provided us with a number of computational tools to study social traits (Ghoul *et al.*, 2017). Importantly, omics studies allow the evolution and ecology of bacterial sociality to be studied in *situ*, and can avoid many of the challenges associated with culture-based approaches. Microbial genomic data can also be effectively analysed using phylogenetic comparative methods (Harvey & Pagel, 1991) to further understand the evolution of microbial sociality. Such ‘comparative genomics’ studies in microbes allow broad evolutionary and ecological predictions to be tested across diverse microbial species, while controlling for the genetic relationships between different strains and species (Ghoul *et al.,* 2017). Controlling for genetic relationships is important in multi-species and multi-strain datasets, because due to shared ancestry, closely related species, or closely related strains within a species, are more likely to share similar traits (Harvey & Pagel, 1991). Therefore, different species/strains are not phylogenetically independent from each other (Clutton-Brock & Harvey, 1977; Harvey & Pagel, 1991), and not controlling for such relationships can lead to biased results in statistical analyses and problems analogous to pseudoreplication in experimental studies (Kruskal, 1988).

In my thesis, I use a comparative approach to study cooperative genes across many bacterial species. Specifically, I test the proposed hypothesis that HGT stabilises cooperation across bacteria. While this hypothesis has received support from theoretical and experimental studies (Smith, 2001; Mc Ginty, 2011, 2013; Dimitriu *et al.,* 2014), previous genomic studies have either only included one species (Nogueira *et al.,* 2009), which may not be representative of all bacteria, or included multiple species but did not appropriately control for phylogenetic non-independence (Nogueira *et al.,* 2012; Garcia-Garcera & Rocha, 2020). We therefore performed the most comprehensive genomic test of this hypothesis to date, using 1632 genomes from 51 diverse bacterial species.

Thesis outline

My thesis is comprised of two related, but distinct, projects that aim to expand our understanding of bacterial sociality in natural populations and further across the bacterial tree of life. In Chapter 2, I provide the methodological detail for how I curated a longitudinally-sampled isolate collection of *Staphylococcus aureus* from a previous study of asymptomatic human nasal carriage. In Chapter 3, I use this collection to study the evolution and ecology of bacteriocin-mediated competition in a natural setting. In Chapter 4, I perform a comparative analysis to test the role of horizontal gene transfer in stabilising cooperation across many diverse bacterial species.

Chapter 2 outline

Chapter 2 is a methodological chapter, in which I curate a natural collection of bacterial isolates from a previous human nasal carriage study of longitudinally-sampled *Staphylococcus aureus*, to study the evolution and ecology of social traits. This involves: i) designing isolate selection criteria; ii) sub-culturing the selected isolates and screening for contaminants; iii) performing large-scale strain-typing (*spa*-typing) to separate 'co-colonised' samples containing multiple strains into their constituent strains; (iv) creating genetic distance trees to examine the genetic relationships between strains. The resulting longitudinally-sampled isolate collection forms the basis of my laboratory and genomic work in Chapter 3. Importantly, it will also form the basis of future research projects that aim to understand the importance of social traits in natural populations of *Staphylococcus aureus*.

Chapter 3 outline

In Chapter 3, I study the evolution and ecology of bacteriocin-mediated competition in the longitudinally-sampled populations of *S. aureus* from the human nasal cavity curated in

Chapter 2. Despite many theoretical and laboratory experiments suggesting that bacteriocin production plays a key role in determining the competitive dynamics of bacterial strains, there is a distinct lack of evidence supporting this in natural populations. To address this, I use laboratory assays to screen *S. aureus* for bacteriocin-mediated inhibitory activity against a range of indicator strains and ecologically-relevant competitors, to determine the prevalence and activity spectra of *S. aureus* bacteriocins. I then map inhibition profiles to strain dynamics data, to test whether inhibitory activity can explain differential strain success over time. Lastly, I use genomic approaches to determine the identity of bacteriocin gene clusters associated with inhibitory activity. I find that inhibitory activity appears to provide S. *aureus* with a competitive benefit in the human nasal cavity: inhibitory strains are more likely to displace other *S. aureus* strains during co-colonisation. This occurs despite inhibitory activity not being displayed by the majority of *S. aureus* strains and targeting interspecific over intraspecific competitors. I also provide evidence for the underlying mechanism of bacteriocins, by identifying five bacteriocin gene clusters associated with phenotypic inhibitory activity. Taken together, I provide evidence that bacteriocins play a role in determining competitive strain dynamics in natural populations of *S. aureus*, and in bacteria more generally.

Chapter 4 outline

In Chapter 4, we use a comparative approach across 51 diverse bacterial species, to test the role of horizontal gene transfer as a mechanism of stabilising cooperation in the face of cheats. Horizontal gene transfer could stabilise cooperation by allowing cooperators to ‘re-infect’ cheats with the gene for public good production, by inserting it into a neighbouring cell on a mobile genetic element, such as a plasmid. This hypothesis is widely accepted, based on previous theoretical, genomic, and experimental work. However, previous genomic studies have either only focused on one species, or have not included an appropriate phylogenetic control when studying multiple. To comprehensively test this hypothesis, we predict the sub-cellular location of genes coding for extracellular proteins, many of which act as public goods, in 1632 genomes across 51 diverse bacterial species. We then ask whether plasmids (mobile genetic elements) are overrepresented with extracellular proteins compared to chromosomes (non-mobile genetic elements), when controlling for phylogeny. Across species, we find that plasmids were not consistently overrepresented with extracellular proteins compared to chromosomes. In subsequent analyses of plasmid mobility, we find that plasmids capable of transferring at higher rates do not code for more extracellular proteins. Instead, we find that extracellular proteins involved in pathogenicity are overrepresented on the plasmids of pathogens with a broad host-range.

Chapter 2. Curating a longitudinally-sampled isolate collection of *Staphylococcus aureus* from the human nasal cavity

Abstract

In this chapter, I describe the process by which I curated an isolate collection to study the evolution and ecology of bacteriocins in natural populations of bacteria, and specifically their role in determining competitive strain dynamics, as outlined in Chapter 1. I first gained access to a longitudinally-sampled collection of *Staphylococcus aureus* from the human nasal cavity. The study involved a total of 571 participants, longitudinally-sampled for a maximum of 90-months and existed as a collection of approximately ~4,500 *S. aureus*-positive samples, collected from specific time points. In order to test my hypotheses, I required cases of both mixed and single-strain colonisation, and participants that were sampled for an extended time periods. Once I had identified and collected the samples of interest, I performed large-scale strain-typing to separate mixed samples containing multiple strains into individual strain isolates. The resulting isolate collection subset is comprised of 389 isolates, representing 64 unique strains, from 40 participants.

Introduction

The lack of studies focusing on the evolution and ecology of bacterial traits in natural populations can in large part be explained by the logistical difficulties and extensive resources required to create appropriate natural isolate collection. Firstly, to track the evolution of populations over time, studies must collect samples from the same population over longitudinal timescales. Secondly, this is compounded by the study requiring a substantial sample size, both within the same population over time, and between populations from the same habitat, to achieve sufficient power for statistical analyses. Thirdly, if the study has a target species, it will typically need to isolate that species from a complex bacterial community. Lastly, if a study wants to understand the determinants of evolutionary strain dynamics, an important requirement for studies of social interactions which often occur between strains of the same species, studies must capture the diversity of strains present, both within and between time points. While some studies to date have successfully created a collection meeting these criteria and used them to gain insights into the role of social interactions in natural populations of bacteria (Jiricny *et al.*, 2014; Ghoul *et al.*, 2015; Andersen *et al.*, 2015, 2017, 2018), our understanding is still limited, and we require further studies of this nature, especially in species understudied from a social evolution perspective.

To advance our understanding of asymptomatic nasal colonisation in *S. aureus,* the Crook/Modernising Medical Microbiology group conducted a study titled: "Bacterial and host factors associated with the disease presentation and carriage duration of *Staphylococcus aureus*" (Oxfordshire Research Ethics Committee B reference number 08/H0605/102) (Young *et al.*, 2012; Golubchik *et al.*, 2013; Everitt *et al.,* 2014; Miller *et al.*, 2014; Votintseva *et al.*, 2014; Das *et al.,* 2016; Laabei *et al.,* 2016; Gordon *et al.*, 2017). The study created a large-scale *S. aureus* nasal carriage collection in people free of *S. aureus* disease, consisting of 571 participants, sampled at approximately two-month intervals for up to a maximum of 90-

months, totaling ~9,000 nasal samples, of which ~4,500 were *S. aureus*-positive. Unlike previous studies that only strain-typed one to three *S. aureus* colonies per sample, this study aimed to better capture *S. aureus* strain diversity (Votintseva *et al.*, 2014). Importantly, contrary to the previous consensus that human hosts were only colonised by a single strain of *S. aureus* over longitudinal time period, this study found that ~18% of participants that were *S. aureus*-positive at recruitment and returned ≥12 nasal samples were co-colonised at a single time point at least once during the first 24-months of study, and that the point incidence of co-colonisation across participants was ~5% up to 24-months (Votintseva *et al.*, 2014). This collection therefore presents a brilliant opportunity to study the evolution of social interactions in natural populations of bacteria, particularly their importance in determining within-species strain dynamics.

I obtained access to the *S. aureus* carriage collection created by the Crook/Modernising Medical Microbiology group, with the aim of understanding the evolution and ecology of social interactions in natural populations of *S. aureus*. In this chapter, I describe the processes I followed for preparing a sub-set of the wider *S. aureus* carriage collection for an isolate collection suitable for testing my hypotheses in Chapter 1. Firstly, I designed selection criteria to curate a subset of the collection that would allow effective study of the role of bacteriocin-mediated competition, or any other social trait, in determining competitive strain dynamics. In brief, this involved selecting a sufficient number of participants, each with a sufficient number of time point isolates and total sampling time frames, and capturing instances where strains dominate within hosts or experience co-colonisation with other strains, to study social interactions between co-existing isolates. Secondly, I sub-cultured *S. aureus* samples in an appropriate manner to maintain strain diversity and prepared isolates for experimental use, such as by screening for contaminants. Finally, I performed large-scale strain-typing to separate 'co-colonised' samples containing multiple strains, into individual strain isolates. The resulting *S.*

aureus collection laid the foundations for Chapter 3, and will be used in future studies to understand the role of social interactions in natural populations of *S. aureus.*

Methods

Original *Staphylococcus aureus* nasal carriage collection - sampling procedure

Participants free from *S. aureus* disease were recruited from five Oxfordshire general practices (GPs) in the Thames Valley Primary Care Research Partnership between December 2008-December 2009. Eligible participants were defined as adults aged ≥ 16-years-old, attending the GP for non-*S. aureus* related reasons. Participants were recruited in age/sex strata to represent the general U.K. population (Votintseva *et al*., 2014). The first nasal swab sample from each participant was collected under research nurse supervision, who trained the participant in self-swabbing (Fig. 1 - Step 1) (Votintseva *et al.,* 2014). Participants were subsequently provided with a self-swabbing kit every 2-months during the study, which were returned *via* post in charcoal medium (less than one week) and stored at 4°C before processing (less than one week) (Votintseva *et al.,* 2014). During sample processing, culture sensitivity of the nasal swabs was increased by incubating aerobically overnight at 37°C in 5% NaCl enrichment broth (Fig. 1 - Step 2) (E&O Laboratories) (Miller *et al*., 2014). SASelect® chromogenic agar (Bio-Rad) was used to isolate *S. aureus*. Colonies are identified as pink and were further tested using DNAse, catalase, and Staphaurex tests using standard procedures (Health Protection Agency, 2007). Only samples positive in all three tests were identified as *S. aureus* and a selection of colonies were resuspended in saline and stored as a glycerol stock at -80°C (Miller *et al*. 2014).

As the study aimed to represent the full diversity of *S. aureus* strains present, glycerol stocks were made by streaking through as many pink *S. aureus* colonies as possible (Fig. 1 - Step 3), resuspending in saline, and storing at -80ºC (Fig. 1 - Step 4) (Miller *et al*., 2014; Votintseva *et* *al*., 2014). During the first ~seven months of study, multiple *S. aureus* colonies were only picked from each sample if they displayed different morphologies (Votintseva *et al*., 2014). After this point, the study switched approaches to streaking as many pink colonies per sample as possible to better capture strain diversity (Votintseva *et al*., 2014). I excluded all samples created using the initial colony morphology approach from our collection subset. To identify discrete strain types, every sample in the collection was subsequently *spa*-typed (Fig. 1 - Step 5) (see 'Methods – section 2.3'). Throughout, we define 'strain' as '*spa*-type'. Samples displaying an unambiguous *spa*-typing result (i.e., one discrete *spa*-type signal) were appropriately labelled; samples giving an ambiguous *spa*-typing result (i.e., multiple *spa*-type signals) were re-streaked and 12-24 individual colonies were picked (Fig. 1 – Step 6) for subsequent *spa*-typing to capture the diversity of *spa*-types present (Fig. 1 – Step 7) (Miller *et al*., 2014; Votintseva *et al*., 2014). After *spa*-typing, each constituent *spa*-type in each co-colonised sample was appropriately labelled, but the separated isolates were not stored for future use (Fig. 1 – Step 8).

In total, the nasal carriage study longitudinally tracked 571 participants, at approximately two-month time intervals, up to a maximum of 90-months. This yielded ~9,000 nasal samples, of which ~4,500 were *S. aureus*-positive. Specifically, of the 571 total participants in the collection, 426 participants were sampled for ≥ 24 months, 183 participants were sampled for ≥ 48 months, 60 participants were sampled for ≥ 64 months, and 50 participants were sampled for ≥ 80 months. Sampling resolution in the collection decreased after 24-months for two reasons: (i) there was a large gap in sampling for all participants between months 50-62; (ii) missing samples were more prevalent after 24-months, and especially after 64-months once sampling re-commenced consistently.

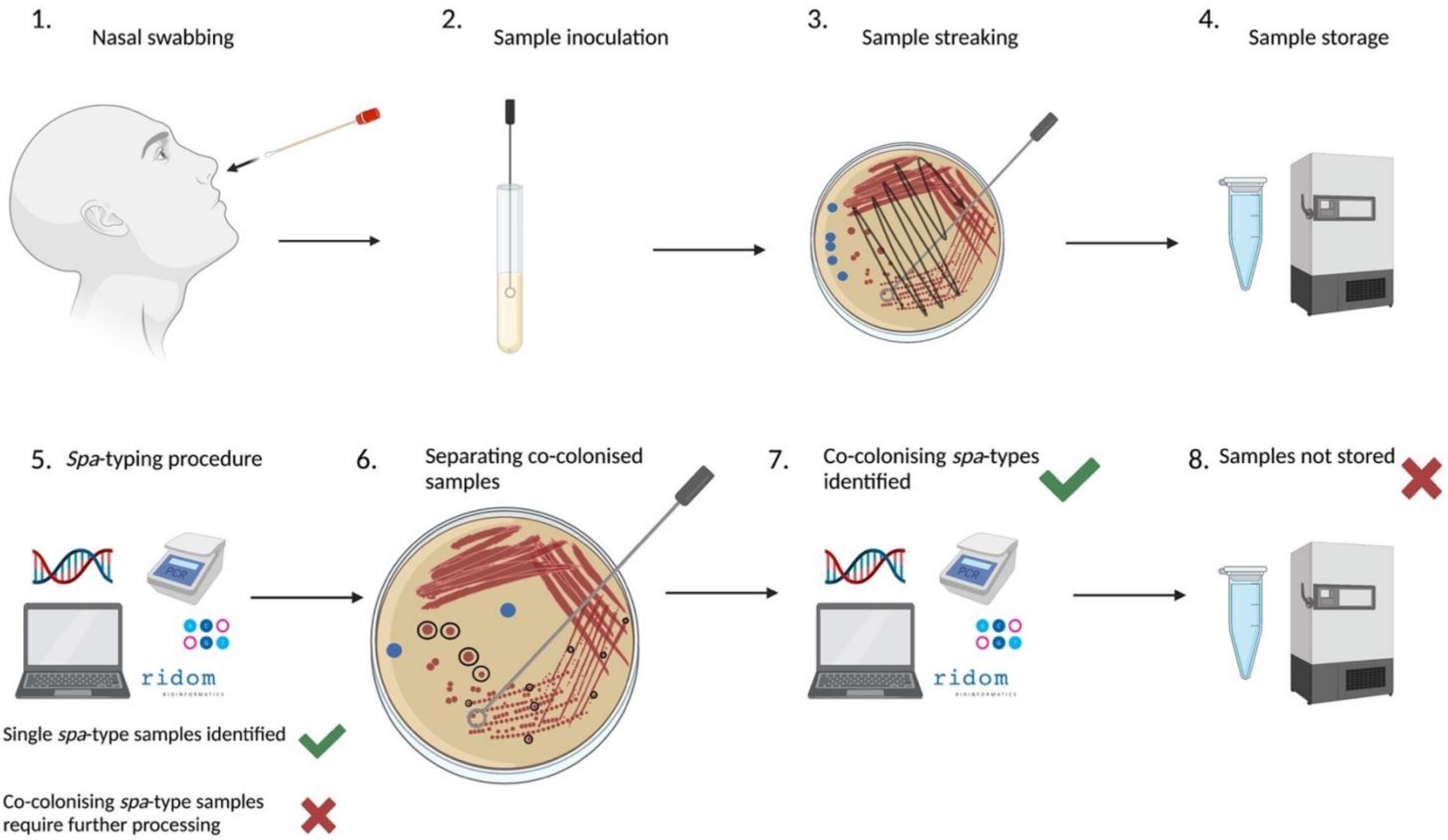

1. Nasal swabbing
2. Sample inoculation
3. Sample streaking
4. Sample storage
5. *Spa*-typing procedure
PCR
ridom
BIOINFORMATICS
Single *spa*-type samples identified ✓
Co-colonising *spa*-type samples require further processing ✗
6. Separating co-colonised samples
7. Co-colonising *spa*-types identified ✓
PCR
ridom
BIOINFORMATICS
8. Samples not stored ✗

Fig 1. Overview of the key steps involved in creating the *S. aureus* nasal carriage collection. Steps 1-8 were initially performed by the Crook/Modernising Medical Microbiology group. Following sample sub-culturing (see 'Methods – section 2.2'), I re-performed steps 6-7, and stored separated *S. aureus* isolates for future use (see 'Methods – section 2.3'). Step 1) nasal swabs were collected from each participant. Step 2) nasal samples were incubated aerobically overnight at 37°C in 5% NaCl enrichment broth (E&O Laboratories, Bonnybridge, UK). Step 3) samples were streaked onto SASelect® chromogenic agar (Bio-Rad), which identifies *S. aureus* as pink colonies. To confirm *S. aureus* identity, pink colonies were further tested using DNAse, catalase, and Staphaurex tests, using standard procedures (Health Protection Agency, 2007). To capture the diversity of *S. aureus* strains present, many pink colonies were streaked, while aiming to avoid any non-pink (depicted in blue) colonies representing other species. Step 4) streaked samples were resuspended in saline and stored as a glycerol stock at -80 ◦C. Step 5) to identify discrete strain types, every sample was *spa*-typed, involving DNA extraction, PCR amplification, Sanger sequencing, and sequence analysis of the *spa* locus (see 'Methods – section 2.3'). Samples displaying an unambiguous *spa*-typing result (i.e., one discrete *spa*-type signal) were appropriately labelled, whereas samples giving an ambiguous *spa*-typing result (i.e., multiple *spa*-type signals) required further processing to capture strain diversity. Step 6) co-colonised samples were re-streaked on SASelect® chromogenic agar (Bio-Rad) and 12-24 pink colonies were picked. Step 7) each colony was re-*spa*-typed using the approach in step 5. Step 8) each separated *spa*-type was appropriately labelled, but the separated isolates were not stored for future use at this stage. This figure was created using BioRendor.com.

Refining *S. aureus* carriage collection for hypothesis testing

Sample selection criteria

I selected a subset of samples from the original *S. aureus* nasal carriage collection to perform laboratory-based hypothesis testing. Given my aim was to understand the influence of bacteriocins on competitive strain dynamics, I required samples from participants that were sampled over relatively long time frames, and from participants that were either colonised by a single strain ('single-strain participants'), or co-colonised by multiple strains ('co-colonised participants'). I therefore selected participants based on the following criteria: co-colonised participants - (i) all participants sampled for ≥ 18 months and that; (ii) contained ≥ 3 time points containing multiple strains (overall co-colonised participants total = 20). Single-strain participants - (i) all participants sampled for ≥ 18 months and that; (ii) contained strains already selected in the co-colonised participant category, to allow us to compare the effect of co-colonisation in strains with the same genetic background (single-strain participants = 17). I then randomly selected a further 3 single-strain participants (of the 5 remaining that met criteria (i)) that contained strains not present in the co-colonised participant category (overall single-strain participants total = 20). From each co-colonised participant, we collected almost[1] all time point isolates, to track strain dynamics in full resolution. From each single *spa*-type participant, we only collected an 'early' (the first *S. aureus*-positive sample), 'middle' (the sample closest to the midpoint of the sampling time frame), and 'late' (the last *S. aureus*-positive sample) timepoint isolate, as this would be sufficient to track the within-host evolution of a single strain.

In total, these criteria corresponded to 302 time point isolates, containing 67 strains, from 40 participants. However, 81/302 isolates required further processing, as they were co-colonised

[1] To make the isolate number more manageable for laboratory work, from two co-colonised participants (participants 420 & 1231), when a single *spa*-type colonised alone for extended periods of time, I selected one in three samples (i.e., a sample every 6 months, instead of every 2 months).

with multiple strains that required separating into their individual constituent strains (see 'Methods – separating 'co-colonised' samples into constituent strains'), which increased the size of the isolate collection further. Details of the finalised isolate collection sub-set are included in the results section below (see 'Results - finalised *S. aureus* isolate collection subset').

Sub-culturing samples & screening for contaminants

I retrieved *S. aureus* freezer stocks from the Crook/Modernising Medical Microbiology laboratory's -80°C freezers located in the John Radcliffe Hospital, Oxford. I sub-cultured the available[2] samples by streaking freezer stocks onto SASelect® chromogenic agar (Bio-Rad) plates, which allow *S. aureus* to be identified as pink colonies, and incubated overnight at 37°C. For isolates containing a single strain, I picked a single *S. aureus* colony to ensure a single genotype of *S. aureus* was selected, and inoculated it overnight in 3ml TSB at 37°C, 200pm before making a glycerol stock and storing at -80°C. I noted any isolates containing any non-pink colonies (i.e., contaminants) and ensured to avoid them when sub-culturing single strain samples. For samples containing multiple strains (i.e., 'co-colonised' samples), I poured 2ml TSB onto the agar plate and mixed with a magnetic stirrer to capture all strains present, before making a glycerol stock and storing at -80°C. Again, I noted any isolates containing non-pink colonies, but they could not be removed from the co-colonised sample at this stage. A small number of isolates were contaminated to the extent that they could not be used in further laboratory work. These isolates were noted accordingly and removed from the isolate collection sub-set.

[2] Some samples meeting our selection criteria were not available for collection at the John Radcliffe Hospital. These samples were noted accordingly and were removed from the sample collection sub-set.

Separating ‘co-colonised’ samples into constituent strains

When creating the original *S. aureus* carriage collection, although the Crook/Modernising Medical Microbiology group did strain-type co-colonised sample, they used an approach that did not store the constituent strains for future use. Therefore, to finalise our isolate collection subset, I needed to separate each co-colonised isolate containing multiple strains (n = 81) into its constituent strains. Below, I outline this process, which involved: i) picking 12-24 individual colonies from each co-colonised sample; (ii) performing DNA extraction; (iii) PCR amplifying and Sanger Sequencing the Staphylococcal protein A (*spa)* locus; (iv) using the resulting *spa* gene sequence to *spa*-type (i.e., strain-type) each colony isolate and curate the finalised collection subset.

- *DNA extraction*

To perform DNA extraction, I initially streaked each co-colonised sample onto SASelect® chromogenic agar (Bio-Rad) and incubating overnight at 37°C. I then initially picked 12 pink *S. aureus* colonies from each sample (colony isolates = 972), inoculated them into 150µl TSB overnight at 37°C, and stored the cultures as glycerol stocks at -80°C. I then re-inoculated each colony isolate into 150µl TSB overnight at 37°C and extracted DNA from each isolate, following the methodology previously used to perform DNA extraction on isolates in the collection (Miller *et al*., 2014; Votintseva *et al.*, 2014). Specifically, this was a crude DNA extraction approach, whereby I centrifuged 150µl of overnight culture at 14,000 rpm for two minutes to obtain a cell pellet. I then removed excess media, re-suspended cells in 60µl of Tris-EDTA (TE) buffer (Invitrogen), and heated at 99.9 °C for ten minutes using a dry heat block. I then centrifuged samples for two minutes, removed 50µl of supernatant without disturbing the pellet, and stored the resulting crude DNA extract at -20 °C for subsequent use (Votintseva *et al.*, 2014). I repeated this DNA extraction process on a further 12 colony isolates for each co-colonised sample that was not fully separated into its constituent strains in the first round of strain-typing (additional colony isolates = 660; overall total colony isolates = 1,632) (see 'Methods – PCR amplification and sequencing of the *spa* locus').

- *PCR amplification and sequencing of the spa locus*

To determine the strain identity of each colony isolate I performed *spa*-typing (Shopsin *et al.* 1999; Harmsen *et al.*, 2003), the same approach as used on the original *S. aureus* carriage collection (Young *et al.*, 2012; Golubchik *et al.*, 2013; Miller *et al.*, 2014; Votintseva *et al.*, 2014). *Spa*-typing is a single-locus strain typing approach comparable in resolution to multilocus sequencing typing (O'Hara *et al.*, 2016) that uses sequence variation in the polymorphic Xr region of the staphylococcal protein A (*spa*) gene to identify discrete strain types (Shopsin *et al.* 1999; Harmsen *et al.*, 2003; Votintseva *et al.*, 2014). Specifically, the Xr region contains many tandem repeats, typically between 21-27 base pairs (bp) (Shopsin *et al.* 1999; Harmsen *et al.*, 2003). Differences in the identity, number, and order of these repeats within the Xr region gives rise to unique *spa*-types (Shopsin *et al.* 1999; Harmsen *et al.*, 2003).

I performed all stages of *spa*-typing following the methodology previously used to *spa*-type samples in the collection (Miller *et al.*, 2014; Votintseva *et al.*, 2014). Initially, I did this by using PCR to amplify the *spa* locus using primers 1095F, 5'-AGACGATCCTTCGGTGAGC-3', and 1517R, 5'-GCTTTTGCAATGTCATTTACTG-3' (Uhlen *et al.*, 1984; Shopsin *et al.*, 1999; Harmsen *et al.*, 2003; Votintseva *et al.*, 2014). PCR reaction mixtures included 0.25 mM deoxynucleoside triphosphates (dNTPs), 0.5 U of GoTaq Flexi DNA polymerase (Promega), GoTaq Flexi Buffer, 2.5 mM $MgCl_2$, 0.25 μM of primers, and 2 μl of DNA extract, in a total volume of 10 μl (Votintseva *et al.*, 2014). PCR conditions were as follows: 35 cycles, each consisting of 94 °C for 30 seconds, 50 °C for 30 seconds, and 72 °C for 60 seconds, with a final extension at 72 °C for 5 minutes (Votintseva *et al.*, 2014). To check the quality of PCR products, I performed gel electrophoresis on the PCR products of ten out of every batch of 96 colony isolates. Gel electrophoresis was performed using 1.2% agarose gel and applying 100v for 30 minutes. Only if all 10/10 PCR products had a clear band in the relevant region for the *spa* gene, which can vary between ~200 to ~600 bp depending on the number of 21-27 bp repeats (Shopsin *et al.,* 1999; Koreen *et al.*, 2004), would a batch be deemed to have met the quality criterion for downstream use.

I then prepared and sent PCR product samples to Source Bioscience (Nottingham, UK), who performed PCR clean-up and Sanger Sequencing. *Spa* gene sequence quality was checked using SeqSphere+ (8.4.0) software (Ridom, GmbH). *Spa* gene sequencing was performed in three stages. Firstly, 12 colony isolates from each co-colonised sample (n = 972), were initially sequenced in the forward direction (primer - 1095F, 5'-AGACGATCCTTCGGTGAGC-3'). Secondly, co-colonised samples not completely separated into constituent strains and that contained 'un-typeable' colony isolates (see 'Methods – *spa*-type identification') were re-sequenced using the same conditions but in the reverse direction (primer - 1517R, 5'-GCTTTTGCAATGTCATTTACTG-3'). The reverse sequence was combined with its corresponding forward sequence to improve sequence quality for subsequent *spa*-typing. Lastly, for co-colonised samples still not fully separated in all constituent strains, I repeated the above protocol of DNA extraction, PCR amplification, and *spa* gene sequencing in both forward and reverse directions, on a further 12 colonies per sample (n = 660). In total, DNA extraction, PCR amplification and Sanger sequencing of the *spa* locus was performed on between 12-24 colony isolates from the 81 co-colonised samples, totaling 1,632 colony isolates. Following *spa*-type identification (see 'Methods – *spa*-type identification'), I inoculated one representative of each unique *spa*-type, from each co-colonised isolate, into 3ml tryptic soy broth (TSB) for overnight growth aerobically at 37°C, 200rpm and stored the cultures as a glycerol stock at -80°C for use in the final collection subset.

- *Spa*-type identification

To determine the identity of *spa*-types in co-colonised isolates, I analysed *spa* gene sequences using SeqSphere+ (8.4.0) (Ridom, GmbH), the latest version of the software previously used to identify *spa*-types in the collection (StaphType v.2.0.3) (Ridom, GmbH), following the same methodology (Miller *et al.*, 2014; Votintseva *et al.*, 2014).

SeqSphere+ (8.4.0) (Ridom, GmbH) software provides a predicted *spa*-type identity for a given *spa* gene sequence and a quality score for this prediction: ‘unknown’ (i.e., un-typeable), ‘poor’, ‘good’, ‘excellent’, based on sequence quality (Ridom, GmbH). For each co-colonised sample, I compared my *spa*-typing results to that of the original *S. aureus* nasal carriage collection by the Crook/Modernising Medical Microbiology group, to determine the level of diversity I re-captured in each co-colonised sample, and whether I ever identified any additional, previously undetected, *spa*-types. I only accepted *spa*-type predictions of good and excellent quality.

I used SeqSphere+ (8.4.0) (Ridom, GmbH) to determine the genetic relationships between *spa*-types using the BURP (based upon repeat pattern) clustering algorithm (Mellmann *et al.*, 2007). The BURP clustering algorithm compares *spa* repeat regions to calculate a BURP distance score (Mellmann *et al.,* 2007). *Spa*-types separated by low BURP distance scores (typically four or less) can be clustered into *spa*-Clonal Complexes (*spa*-CCs) i.e., groups that share a relatively high degree of genetic similarity (Mellmann *et al.,* 2007). I used SeqSphere+ (8.4.0) (Ridom, GmbH) to create a *spa* gene minimum-spanning genetic distance tree (Fig. 2) of all *spa*-types in the collection subset (Mellmann *et al.,* 2007)**.** Due to the BURP clustering algorithm only working on *spa*-types with ≥ 5 repeats at the *spa* locus, two *spa*-types (t528 and t870) could not be included in the genetic distance tree (Mellmann *et al*., 2007).

Results

Separating co-colonised samples

To separate all 81 co-colonised samples into their constituent individual strains, previously identified by the *S. aureus* carriage study, I *spa*-typed 12-24 colonies per sample. In total, 72 % (58/81) were fully separated into all constituent strains identified by the original study, and an additional 11% (9/81) of all co-colonised samples were separated into multiple, but not all, constituent strains. Therefore, >1 strain was isolated from 83% (67/81) of co-colonised samples, while at least one strain was separated from 100% (81/81) of co-colonised samples. In total, an additional 168 separated isolates were added to the previous total of 221 isolates, giving a final total of 389 isolates.

In addition to identifying at least one strain previously found by the original *S. aureus* carriage study in all 81 of co-colonised samples, I also identified additional *spa*-types, not found by the original study, in 14% (11/81) of co-colonised samples. Of these 11 instances, nine involved strains that were identified within that participant, but in a different time point sample. Therefore, only 2% (2/81) of co-colonised samples contained strains not already identified in the participant**.** These two instances involved two strains (one per co-colonised sample) that were not identified in any of the other participants in my collection subset. All ‘unexpected’ strain isolates were added to the collection dataset at their relevant time points, and the two unique strains were added to the strain total. A total of seven co-colonised strains could not be separated from their co-colonised samples, five of which were not present in any other participant in the collection, and were therefore removed from the final isolate collection subset.

Finalised *S. aureus* isolate collection subset

In total, our finalised *S. aureus* collection subset consists of 302 *S. aureus*-positive nasal samples, which were separated into 389 *S. aureus* strain isolates, representing 64 unique *spa*-types, from 40 participants (Table. 1). The participants can be divided into two participant categories: (i) 'single-strain' participants that only contain one strain over time; (ii) 'co-colonised' participants that contain multiple strains, both within and between timepoints samples (single-strain participants: n = 20; samples = 59; isolates = 59; strains = 19; co-colonised participants: n = 20; samples = 243; isolates = 330; strains = 59) (Table. 1). We examined the genetic relationship between *S. aureus* strains in our finalised collection subset (Fig. 1). Of the 62 strains available for BURP clustering (see Fig. 1) (Mellmann *et al.,* 2007), 49/62 belong to ten *spa*-Clonal Complexes (*spa*-CCs), with the remaining 13/62 strains being 'singletons', i.e., *spa*-types that did not share high enough similarity at the *spa* locus with any other *spa*-types to be clustered into a *spa*-CC.

Participant no.	Participant ID	Strains	Time frame of *S. aureus*-containing samples (months)	No. of single strain samples	No. of co-colonised samples	Total no. samples containing *S. aureus*	No. separated *S. aureus* isolates from co-colonised samples	Total no. *S. aureus* isolates
Single-strain participants								
1	22	1 (t012)	40	3	NA	3	NA	3
2	42	1 (t1716)	40	3	NA	3	NA	3
3	162	1 (t209)	42	3	NA	3	NA	3
4	359	1 (t089)	40	3	NA	3	NA	3
5	450	1 (t382)	84	3	NA	3	NA	3
6	451	1 (t379)	66	3	NA	3	NA	3
7	454	1 (t127)	82	3	NA	3	NA	3
8	499	1 (t021)	22	3	NA	3	NA	3
9	671	1 (t032)	72	3	NA	3	NA	3
10	686	1 (t3262)	14*	3	NA	3	NA	3
11	952	1 (t084)	22	3	NA	3	NA	3

Participant no.	**Participant ID**	**Strains**	**Time frame of *S. aureus*-containing samples (months)**	**No. of single strain samples**	**No. of co-colonised samples**	**Total no. samples containing *S. aureus***	**No. separated *S. aureus* isolates from co-colonised samples**	**Total no. *S. aureus* isolates**
12	967	1 (t171)	16*	3	NA	3	NA	3
13	972	1 (t127)	18	3	NA	3	NA	3
14	997	1 (t015)	10*	2*	NA	2	NA	2
15	1209	1 (t346)	42	3	NA	3	NA	3
16	1212	1 (t056)	80	3	NA	3	NA	3
17	1307	1 (t408)	74	3	NA	3	NA	3
18	2009	1 (t230)	32	3	NA	3	NA	3
19	2030	1 (t002)	24	3	NA	3	NA	3
20	2104	1 (t571)	24	3	NA	3	NA	3
Single-strain participants – total (mean)		20 (1.00) Total unique strains = **19** (0.95)	844 (42.20)	59 (2.95)	NA	59 (2.95)	NA	**59** (2.95)

Participant no.	**Participant ID**	**Strains**	**Time frame of *S. aureus*-containing samples (months)**	**No. of single strain samples**	**No. of co-colonised samples**	**Total no. samples containing *S. aureus***	**No. separated *S. aureus* isolates from co-colonised samples**	**Total no. *S. aureus* isolates**
Co-colonised participants								
21	132	4 (t228, t3097, t1885, t2556)	18	4	4	8	11	15
22	420	5 (t379, t608, t021, t1414, t620)	82	12	4	16	7	19
23	424	3 (t3304, t7514, t7049)	84	30	4	34	10	40
24	637	2 (t870, t7050)	20	8	3	11	5	13
25	638	2 (t065, t171)	20	7	3	10	6	13
26	647	3 (t230, t6855, t019)	20	7	4	11	8	15
27	688	4 (t002, t3262, t105, t053)	14*	3	4	**7**	9	12

Participant no.	**Participant ID**	**Strains**	**Time frame of *S. aureus*-containing samples (months)**	**No. of single strain samples**	**No. of co-colonised samples**	**Total no. samples containing *S. aureus***	**No. separated *S. aureus* isolates from co-colonised samples**	**Total no. *S. aureus* isolates**
28	903	2 (t382, t2643)	14*	5	1*	6	2	7
29	926	3 (t008, t196, t127)	22	4	7	11	15	19
30	930	4 (t499, t2074, t120, t015)	22	6	4	10	7	13
31	971	4 (t230, t012, t528, t008)	22	9	3	12	4	13
32	1045	4 (t084, t209, t2119, t1510)	24	9	4	13	7	16
33	1092	3 (t012, t021, t871)	24	5	3	8	6	11
34	1231	6 (t160, t120, t4309, t040, t15780**,** t499)	80	13	3	16	7	20
35	1366	3 (t084, t089, t267)	24	7	6	13	12	19

Participant no.	**Participant ID**	**Strains**	**Time frame of *S. aureus*-containing samples (months)**	**No. of single strain samples**	**No. of co-colonised samples**	**Total no. samples containing *S. aureus***	**No. separated *S. aureus* isolates from co-colonised samples**	**Total no. *S. aureus* isolates**
36	1367	4 (t6390, t6817, t7409, t346)	24	6	7	13	15	21
37	1378	3 (t096, t528, t298)	24	8	4	12	8	16
38	2004	3 (t160, t321, t1685)	24	9	3	12	5	14
39	2060	4 (t065, t012, t267, t084)	24	4	6	10	12	16
40	2064	7 (t1239, t6825, t6826, t6814, t171, t7031, t008)	24	6	4	10	12	18
Co-colonised participants – total (mean)		73 (3.65) Total unique strains = **59** (2.95)	610 (30.50)	162 (8.10)	81 (4.05)	**243** (12.15)	**168** (8.40)	**330** (16.50)

Participant no.	**Participant ID**	**Strains**	**Time frame of *S. aureus*-containing samples (months)**	**No. of single strain samples**	**No. of co-colonised samples**	**Total no. samples containing *S. aureus***	**No. separated *S. aureus* isolates from co-colonised samples**	**Total no. *S. aureus* isolates**
All participants - summary								
All participants – total (mean)		93 (2.33) Total unique strains = **64** (1.60)	1454 (**36.35**)	**221** (5.53)	**81** (4.05)	**302** (7.55)	**168** (8.40)	**389** (9.73)

Table 1. An overview of the finalised *S. aureus* collection sub-set. Particularly important numerical values are highlighted in bold. A 'sample' refers to a nasal swab collected at a given time point. **"**Time frame of *S. aureus*-containing samples (months)" represents the time between the first and last *S. aureus*-containing sample in each participant. Note that single-strain participants only contain single-strain samples, whereas co-colonised participants contain both single-strain and co-colonised samples. The total number of single-strain samples plus co-colonised samples gives the total number of time point samples in a given participant. An 'isolate' refers to an individual strain of *S. aureus* taken from each time point sample; each co-colonised sample required separating into its constituent strain isolates. The total number of *S. aureus* isolates is the total of single-strain isolates plus the total number of separated strain isolates from co-colonised samples. "Total unique strains" in 'all participants' is calculated by summing the number of unique strains in the whole collection i.e., by ignoring the distinction between single-strain and co-colonised participants. Samples and isolates that were unavailable for collection, contaminated, or selectively not collected for logistical reasons, were removed from the final collection subset displayed in Table 1. Cases where this led to participants containing a sampling time frame of <18 months, fewer than three samples in single-strain participants, or fewer than three co-colonised samples in co-colonised participants, as set out in the sample selection criteria ('Methods – Sample selection criteria'), are denoted by an asterisk *.

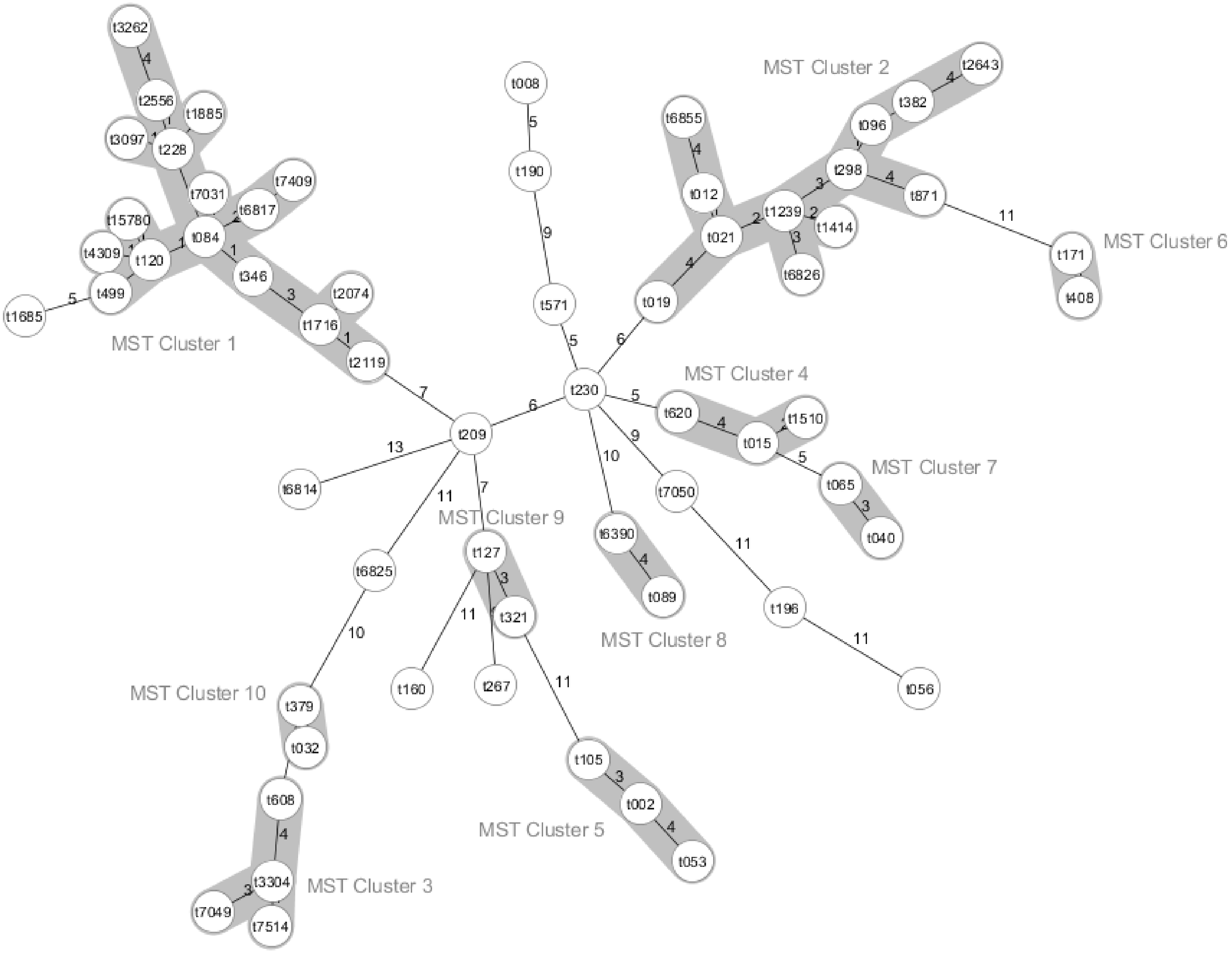


Fig 2. A *spa* gene minimum-spanning genetic distance tree displaying the relationships between *S. aureus* strains in the collection. The genetic distance tree is based on the *spa* gene sequence of each strain in the collection and produced using the BURP (based-upon repeat pattern) clustering algorithm (Mellmann *et al*., 2007), integrated within SeqSphere+ (8.4.0) (Ridom, GmbH). Each strain represents an individual '*spa*-type'. Two strains (t528 and t870) were removed from the tree, as they did not contain the required number of repeats (≥ 5) for BURP clustering (Mellmann *et al*., 2007). The remaining strains (n = 62) are denoted as circles and are clustered into minimum-spanning tree (MST) clusters, which represent *spa*-Clonal Complexes (*spa*-CCs), and are denoted by grey shading. Each MST cluster contains *spa*-types with a BURP distance of four or less, the default for BURP clustering (Mellmann *et al*., 2007). Branch length corresponds to the BURP distance between strains, which is also labelled as a numerical value on each branch.

Discussion

In this chapter, to allow the future study of social traits in natural populations of bacteria, I curated a longitudinally-sampled collection of *S. aureus* isolates, from a previous study on asymptomatic nasal carriage in human participants (Young *et al*., 2012; Golubchik *et al*., 2013; Miller *et al*., 2014; Votintseva *et al*., 2014). The finalised collection consists of 40 participants (including 20 single-strain participants and 20 co-colonised participants) and 308 time point samples, which were separated into 389 isolates, representing 64 unique *S. aureus* strains.

A key strength of this isolate collection subset is that it tracks natural bacterial populations within hosts over time, and stands in comparison or improves upon the number of participants, samples, sampling interval length, and overall timeframe of previous longitudinal sociomicrobiology studies, providing sufficient power for statistical analyses (Jiricny *et al*., 2014; Ghoul *et al*., 2015; Andersen *et al*., 2015, 2017, 2018). Another strength is that we successfully re-captured the vast majority of within-participant strain diversity originally identified in the nasal carriage collection, allowing us to effectively test whether social traits play a role in determining natural strain dynamics in *S. aureus.* A further strength of the collection is that it tracks *S. aureus* in its asymptomatic commensal state during nasal carriage in healthy human hosts. Most previous studies of bacterial sociality in longitudinally-sampled natural populations have collected clinical samples from infected patients (Jiricny *et al*., 2014; Ghoul *et al*., 2015; Andersen *et al*., 2015, 2017, 2018). While these collections are highly medically relevant and are associated with logistical benefits, they also have added complications, caused by extensive antibiotic exposure and severe host immune responses. While the human immune system and *S. aureus* are known to interact during nasal colonisation (Krismer *et al*., 2017), previous studies have shown that host genetics are unlikely to be the only determinant of *S. aureus* nasal colonisation (Andersen *et al.*, 2013). Studying social traits in non-clinical settings could help to further isolate their role in determining the evolutionary and ecological success of bacteria.

There are many logistical challenges associated with creating a natural bacterial isolate collection, which often cause collections to be limited in some respects. A limitation of the overall nasal carriage collection is that sampling intervals increase and become more inconsistent after month 24. This partly reflected staff turnover and prioritization of other projects in the Modernising Medical Microbiology group. However, appropriate statistical methods can be implemented to reduce the effect of issues caused by missing data (Laird, 1988; Twisk & de Vente, 2002; Sainani, 2015). Datasets can also be subset to perform variations of each analysis to test the robustness of results, for example by performing analyses on samples only collected during a period of high sampling resolution (e.g., 0-24 months), in addition to the whole collection. Another potential limitation is caused by the nature of nasal self-swabbing, which could lead to false-negatives in some sampling time points, either for a particular strain, or for the general presence of *S. aureus*. It is possible to at least partially alleviate the effect of false-negative samples by implementing data modifications when set criteria are met. For example, previous studies on the *S. aureus* nasal carriage collection have defined ‘strain loss’ as the absence of *S. aureus* in two consecutive nasal samples, instead of absence in just one nasal sample (Miller *et al.*, 2014). Ideally, to test the robustness of results, studies should perform variations of analyses with and without such criteria. Lastly, while we successfully re-captured all of the strains present in 72% of co-colonised samples, and more than one strain in 83% of co-colonised samples, we could not separate all strains, and all unseparated strains were removed from the collection and could not be examined for social traits. However, capturing all strains from all samples was not necessary for us to effectively study the role of social traits in affecting longitudinal strain dynamics.

Chapter 3. The evolution and ecology of bacteriocin-mediated competition in *Staphylococcus aureus* colonising human hosts

J. L. Thomas[1], M. Liu[1], J. E. Bray[1], D. W. Crook[2], D. J. Wilson[3*], A. S. Griffin[1*] & M. Ghoul[1*]

1. Department of Biology, University of Oxford, UK

2. Nuffield Department of Medicine, Experimental Medicine Division, University of Oxford, John Radcliffe Hospital, Oxford, UK

3. Big Data Institute, Nuffield Department of Population Health, University of Oxford, UK

* = co-last author

Abstract

Bacteriocins are antimicrobial toxins produced by bacteria to kill competing strains and species. Theory and laboratory experiments suggest that bacteriocin production plays a key role in determining the competitive dynamics of bacterial strains. However, there is a lack of evidence supporting this in natural populations where population structures are often unknown. Here, we examined the role of bacteriocin-mediated competition in longitudinally-sampled populations of *Staphylococcus aureus* from the human nasal cavity. Specifically, we tested whether *S. aureus* strains capable of inhibiting the growth of indicator strains and ecologically-relevant competitors were more successful colonisers over time. We found that bacteriocin-producing strains were associated with the propensity to displace competing strains from the nasal cavity. This association was evident despite bacteriocin production not being observed in the majority of strains and targeting interspecific over intraspecific competitors. Whole-

genome sequencing revealed five bacteriocin gene clusters associated with inhibitory activity. Taken together, we provide evidence that bacteriocins play a role in determining competitive strain dynamics in natural populations of *S. aureus*. More generally, our study demonstrates the value of using natural populations to test predictions from theoretical and laboratory experiments, which can often produce unexpected results.

Introduction

Resources such as nutrients and space are often limited in microbial communities (West *et al.,* 2006; Foster, 2010; Ghoul & Mitri, 2016). Bacteria frequently mediate competition for resources through the production and secretion of antimicrobial toxins called 'bacteriocins', which target competing strains and/or species (Riley & Gordon, 1999; Riley & Wertz, 2002; Ghoul & Mitri, 2016; Janek *et al.,* 2016; Nadell *et al.*, 2016; Garcia-Bayona & Comstock, 2018; Granato *et al.*, 2019; Heilbronner *et al.*, 2021). Bacteriocins are widespread, with the vast majority of well-studied bacterial species being identified to carry one or more (Klaenhammer, 1993; Granato *et al.*, 2019). Theoretical and laboratory experiments support a key role for bacteriocins in determining competitive strain dynamics (Levin, 1988; Frank, 1994; Riley & Gordon, 1996, 1999; Gordon & Riley, 1999; Kerr *et al.,* 2002; Gardner *et al.,* 2004; Krikup & Riley, 2004; Waite & Curtis, 2009; Kommineni *et al.*, 2015; Libberton *et al.,* 2015; Kawada-Matsuo *et al.,* 2016; Lehtinen *et al.,* 2022). For example, bacteriocins have repeatedly been shown to provide producing strains with a competitive advantage over sensitive strains in laboratory experiments, allowing bacteriocin producers to invade or defend established bacterial populations (Riley & Gordon, 1999; Waite & Curtis, 2009; Libberton *et al.,* 2015; Kawada-Matsuo *et al.,* 2016). Bacteriocin-mediated interactions can also select for competing strains to evolve bacteriocin immunity, removing the competitive advantage associated with bacteriocin production (Kerr *et al.,* 2002; Krikup & Riley, 2004). However, the extent to which bacteriocins mediate competition in natural populations of bacteria is not well known. In particular, there is a distinct lack of evidence regarding whether bacteriocin-producing strains gain a competitive benefit in natural populations over time.

Studies of bacteriocins in natural populations typically provide insight by quantifying the prevalence of bacteriocins in natural environments and elucidating underlying mechanisms, using either culture-based (Janek *et al.,* 2016; Coyne *et al.,* 2019; Fernández-Fernández *et al.,* 2022) or purely genomic approaches (Donia *et al.,* 2014; Aleti *et al.,* 2019). Many of these studies are conducted with a medical or industrial focus, such as to identify novel antimicrobials or probiotics to treat infections or preserve food (Nascimento *et al.,* 2002, 2006; Cotter *et al.,* 2013; Okuda *et al.,* 2013; de Freire Bastos *et al.,* 2020). Studies with more of an evolutionary and/or ecological focus have provided insight by mapping bacteriocin profiles to microbial community presence/absence and abundance data, to identify negative associations between species-level competitors (Yan *et al.,* 2013; Liu *et al.,* 2015; Zipperer *et al.,* 2016). For example, Zipperer *et al.* (2016) found that strains of *Staphylococcus lugdunensis* isolated from the human nasal cavity produce an antimicrobial toxin called lugdunin, which can inhibit laboratory strains of *S. aureus*, and that nasal colonisation with *S. lugdunensis* is associated with a reduced *S. aureus* nasal carriage. However, the majority of natural bacteriocin studies, such as those discussed above (e.g., Janek *et al.,* 2016; Zipperer *et al.,* 2016; Fernández-Fernández *et al.,* 2022), only screen natural isolates against a set of indicator strains and species instead of testing for bacteriocin activity against co-occurring strains from the same environment, which represent the most ecologically-relevant competitors of bacteriocin-producing strains.

Studies screening natural strains for bacteriocin activity against other naturally-occurring strains have provided insight into evolutionary causes of bacteriocin production (Cordero *et al.,* 2012; Hawlena *et al.,* 2012; Perez-Gutierrez *et al.,* 2013; Kinkel *et al.,* 2014; Abrudan *et al.,* 2015; Bruce *et al.,* 2017). For example, natural studies have provided evidence to support the prediction that bacteriocin-mediated inhibitory interactions are more common under conditions of strong resource competition (Kinkel *et al.,* 2014; Bruce *et al.,* 2017). When screening natural soil isolates of *Pseudomonas fluorescens* for bacteriocin activity, Bruce *et al.* (2017) found that bacteriocin-mediated inhibition was more common between strains with a relatively high degree of overlap in their metabolic requirements i.e., between relatively strong competitors. This result supported that of a previous study which identified a positive correlation between the level of inhibitory interactions and the degree of metabolic niche overlap in *Streptomyces spp.* soil isolates (Kinkel *et al.,* 2014). High levels of resource competition are also thought to explain the general observation that many species produce bacteriocins with a narrow-spectrum of activity that either solely or more frequently target conspecifics instead of heterospecifics, as observed in natural populations of *Xenorhabdus spp.* isolated from soil environments (Hawlena *et al.,* 2012).

Once evolved, many bacteriocins are known to be tightly regulated to reduce the cost of their expression, *via* a process called ‘competition sensing’ (Cornforth & Foster, 2013). Competition sensing occurs when bacteria only produce bacteriocins in response to cues of competition, including nutrient limitation, which indicates exploitative competition, or cellular damage, which indicates interference competition. Natural bacteriocin studies have provided evidence to support the role of competition sensing in natural populations, for example, a further study of *Streptomyces spp.* soil isolates found that the prevalence of inhibitory interactions increased when strains were co-cultured with a competitor, compared to being cultured alone (Abrudan *et al.,* 2015). Interestingly, this study also found that some strains from *Streptomyces spp.* were capable of reducing the inhibitory activity of other strains when grown in co-culture, suggesting a role for the suppression of bacteriocin activity in bacteria.

Natural studies have also provided insight into the evolutionary consequences of bacteriocins, such as how bacteriocin production can affect the assembly of bacterial populations (Hawlena *et al.,* 2010a, 2010b; Cordero *et al.,* 2012; Perez-Gutierrez *et al.,* 2013; Bruce *et al.,* 2017). For example, Bruce *et al.* (2017) observed that while the majority *P. fluorescens* isolates (~63%) could inhibit one or more other isolates, relatively few pairwise interactions between isolates resulted in inhibition (~7%), a pattern also observed in other species, including *Bacillus spp*. isolated from the soil (Perez-Gutierrez *et al.,* 2013) and *Vibrio spp.* isolated from the ocean (Cordero *et al.,* 2012). Interestingly, despite the low prevalence of pairwise inhibitory interactions, Bruce *et al.* found that *P. fluorescens* isolates were more likely to inhibit isolates from different patches, which they are unlikely to directly interact with, compared to isolates from their own patch. Similar findings have been identified in natural populations of *Bacillus spp.* (Perez-Gutierrez *et al.,* 2013) and *Xenorhabdus spp.* (Hawlena *et al.,* 2010a, 2010b) isolated from soil environments. This suggests that immunity to bacteriocins can reduce the prevalence of inhibitory pairwise interactions between genotypes and can play an important role in the assembly of natural bacterial communities

However, the majority of natural bacteriocin studies, such as those described above, only screen a single time point snapshot from a population for bacteriocin activity (Cordero *et al.,* 2012; Hawlena *et al.,* 2012; Perez-Gutierrez *et al.,* 2013; Kinkel et al., 2014; Abrudan *et al.,* 2015; Bruce *et al.,* 2017). If we are to fully understand the evolutionary consequences of bacteriocin production, and specifically the role of bacteriocins in determining competitive strain dynamics in natural populations over time, we require studies that sample populations longitudinally, such that evolutionary and ecological dynamics can be tracked. Some studies have sampled natural populations over time and tested for bacteriocin activity, however most of these studies either do not sample at different time points from within the same host/local habitat, accurately capture strain-level diversity, or investigate whether bacteriocin production can determine differential strain success (Gordon *et al.,* 1998; Wilson *et al.,* 1998; Kraemer *et al.,* 2017). For example, Kraemer *et al.* (2017) sampled *Pseudomonas spp.* from a forest soil environment for over 2-years and tested isolates for bacteriocin activity. They found that bacteriocin activity appears to be structured in time, as interactions between isolates sampled from the same time point were more likely to be inhibitory compared to isolates from different time points, which is suggested to be caused by temporally co-occurring isolates representing stronger competitors. However, this study did not aim to explicitly test whether bacteriocin production drives differential strain success in natural bacterial populations, and is unlikely to have accurately captured the strain-level diversity present in the population, due to the partial 16S sequencing approach used to distinguish between *Pseudomonas spp.* isolates.

To identify the role of bacteriocins in determining competitive strain dynamics in natural populations over time, we require studies that do three things: (i) longitudinally sample natural bacterial populations over extended time periods and capture strain diversity; (ii) experimentally screen samples for bacteriocin activity, and test whether it determines strain success; (iii) combine phenotypic data with genomic data to understand the underlying mechanisms of bacteriocin activity. Interestingly, a previous study adopting this proposed approach in natural populations of *Pseudomonas aeruginosa*, a gram-negative species with well-studied bacteriocins, was unable to identify a clear competitive benefit associated with bacteriocin activity (Ghoul *et al.*, 2015). Other studies using similar approaches, such as early studies of *Escherichia coli* from the human gut, also found conflicting evidence for the role of bacteriocins (Sears *et al.*, 1950; Branche *et al.*, 1963). More recently, a study by Holt *et al.* (2013) that sampled *Shigella sonnei* across human hosts in local endemic populations over 15-years found that dominant strains contained a signature of positive selection for a plasmid (pDPT1) encoding a bacteriocin (E5-type colicin). Note that this study did not conduct longitudinal sampling of the same individuals. Further studies of bacteriocin-mediated interactions between strains in natural populations are now required, particularly in species whose bacteriocins have been underexplored.

In this study, to further understand the role of bacteriocins in determining competitive strain dynamics in natural populations, we used a collection of *Staphylococcus aureus,* a gram-positive opportunistic pathogen from the human nasal cavity, consisting of 389 nasal isolates of 64 strains, from 40 human participants, tracked at two-month intervals for up to 88-months per participant (Miller *et al.*, 2014; Votintseva *et al.*, 2014). Although *S. aureus* is known to be capable of producing bacteriocins (de Oliveira *et al.*, 1998; Netz *et al.*, 2002; de Freire Bastos *et al.*, 2020), including in the nasal cavity (Janek *et al.*, 2016), they, and those from other gram-positive species, have been understudied from an evolutionary and ecological perspective. To address this, we firstly screened all *S. aureus* isolates for their ability to inhibit the growth of an indicator strain for bacteriocin production, other *S. aureus* strains also isolated from the nasal cavity, and three interspecific competitors that also commonly colonise the nasal cavity, to characterise inhibitory activity in terms of prevalence, the identity of inhibitory strains, and activity spectra. Secondly, we mapped the resulting inhibition profiles to longitudinal strain dynamics data, to test whether inhibitory activity can explain differential strain success over time. Finally, we performed whole-genome sequencing on a subset of genomes to identify bacteriocin genes responsible for any observed inhibitory activity.

Methods

Overview

To determine the prevalence and ecological importance of bacteriocins in *Staphylococcus aureus*, we accessed a longitudinally-sampled natural sample collection of *S. aureus* from the human nasal cavity (Young *et al.,* 2012; Golubchik *et al.,* 2013; Miller *et al.,* 2014; Votintseva *et al.,* 2014). We used laboratory assays to screen bacterial isolates for their ability to inhibit the growth of: (i) a positive indicator for *S. aureus* bacteriocin production: *Cellulomonas fimi* (de Oliveira *et al.,* 1998; Coelho *et al.,* 2007); (ii) ecologically-relevant intraspecific competitors and a positive indicator strain for intraspecific inhibition: Newman *ΔdltA* (Peschel *et al.,* 1999; Janek *et al.,* 2016); (iii) ecologically-relevant interspecific competitors: *Moraxella catarrhalis*, *Corynebacterium pseudodiphtheriticum*, and *Staphylococcus epidermidis* (Liu *et al.,* 2015; Janek *et al.,* 2016; Krismer *et al.,* 2017). We subsequently mapped inhibition profiles to strain dynamics data to identify any competitive benefits associated with inhibitory activity. Finally, we performed whole-genome sequencing on a selection of isolates, based on their inhibitory activity profiles to mine genomes for the presence of bacteriocin gene clusters (BGCs).

Staphylococcus aureus isolate collection

For details of the *Staphylococcus aureus* isolate collection used in this chapter, see "Chapter 2 - curating a longitudinally-sampled isolate collection of *Staphylococcus aureus* from the human nasal cavity".

Bacterial strains and growth conditions

We sub-cultured *S. aureus* nasal samples and isolated individual strains of *S. aureus* as described in Chapter 2. *S. aureus* Newman is a commonly used laboratory strain and Newman *ΔdltA* is an isogenic knock-out mutant that lacks D-alanine modification of teichoic acids in the cell wall, increasing the strain's susceptibility to antimicrobial activity and making it an effective indicator strain for *S. aureus* intraspecific inhibition (Peschel *et al.*, 1999; Janek *et al.*, 2016). *C. fimi* (ATCC® 484™), *M. catarrhalis* (ATCC® 25240™), *C. pseudodiphtheriticum* (ATCC® 10700™), and *S. epidermidis* (ATCC® 14990™) are all commonly used laboratory strains and were purchased from American Type Culture Collection (ATCC). *C. fimi* is known to be highly susceptible to *S. aureus* bacteriocins and is therefore an effective positive indicator for bacteriocin production (de Oliveira *et al.*, 1998; Coelho *et al.*, 2007). We cultured all species aerobically at 37°C, 200rpm in the following media: *S. aureus* = tryptic soy broth (TSB); *S. epidermidis* = nutrient broth; *C. fimi* and *C. pseudodiphtheriticum* = King's broth (KB), and *M. catarrhalis* = brain heart infusion broth (BHI). In laboratory screens, we grew all species in or on tryptic soy agar (TSA), aerobically at 37°C.

Phenotypic characterisation of inhibitory activity

Preparation of bacterial cultures & lawns

We screened *S. aureus* isolates for bacteriocin-mediated inhibitory activity using the approach described by Janek *et al.* (2016). Initially, we prepared *S. aureus* cultures to be tested for inhibitory activity by inoculating a freezer stock for each isolate into 150µl tryptic soy broth (TSB) in 96-well plates and incubated aerobically overnight at 37°C. We prepared strains to be seeded into agar lawns by inoculating a freezer stock in 3ml of the relevant media (stated above) and incubated aerobically overnight at 37°C, 200rpm. Next, we prepared agar test plates by cooling autoclaved TSA to 45°C and inoculating with the relevant lawn strain. Lawn strains were standardised (*S. aureus* and *S. epidermidis* OD_{600} = 0.01; *C. fimi* and *C. pseudodiphtheriticum* OD_{600} = 0.04; *M. catarrhalis* = OD_{600} = 0.1) and further diluted 1000-fold by mixing 40µl of bacterial culture into 40ml of TSA. After mixing, the 40ml plates were poured and left to dry.

Agar spot assay

We used a pin replicator to stamp pure overnight cultures of *S. aureus* onto the seeded TSA plates, which were subsequently incubated aerobically overnight at 37°C. Once a uniform lawn had formed, we checked for clear zones of inhibition surrounding the *S. aureus* spots, a characteristic phenotype of bacteriocin activity (see Fig. S10 for a representative zone of inhibition image). Each species' lawn was grown for the following incubation period: *S. aureus* and *S. epidermidis* = 12 hours, *M. catarrhalis* and *C. pseudodiphtheriticum* = 15 hours, *C. fimi* = 18 hours. If uniform lawns were yet to form after this period, plates were re-incubated and continuously checked until uniform lawns appeared. Within each screen, we performed three replicates of every *S. aureus* spot on every species lawn. Isolates were defined as displaying inhibitory activity if they displayed a positive result in all three replicates. To confirm positive results, we re-screened all isolates displaying inhibitory activity, using the same approach. We included a positive and negative control strain for inhibitory activity on every test plate. Where stated, following the methodology of Janek *et al.* (2016), we stress-induced bacterial strains by adding a final concentration of 200μM 2,2'-bipyridine for iron limitation and 0.01% hydrogen peroxide (H_2O_2) for oxidative stress, both of which represent ecologically-relevant stressors experienced by bacteria colonising the nasal cavity (Janek *et al.*, 2016). To measure the size of inhibition zones, *S. aureus* spots were standardised to $OD_{600} = 0.4$ and screened as described above. Inhibition zones were then measured in cm from colony edge to inhibition zone edge. We imaged plates using a G:box EF2 camera box (Syngene Ltd) and a Canon EOS 5D Mark IV (Canon Inc.).

Genome selection & sequencing

Genome selection

To identify genes associated with inhibitory activity, we whole-genome sequenced a selection of 95 inhibitory and non-inhibitory isolates from our laboratory screens. More detailed information about the isolate selection criteria can be found in 'Supplementary Methods – genomic analyses'. In brief, we selected: (i) inhibitory isolates – a representative isolate for every inhibitory strain, in each participant it was present. If multiple time point isolates were available, the earliest and latest were selected; (ii) non-inhibitory isolates – all isolates that display a loss of inhibitory activity over time, in strains that otherwise display inhibitory activity; (iii) non-inhibitory isolates – two random isolates, for each non-inhibitory strain present in both single-strain participant and co-colonised participant categories, one from each category; (iv) non-inhibitory isolates - one random isolate from each participant yet to have an isolate selected from the above criteria; (v) non-inhibitory isolates – four random isolates from four strains yet to be selected from the above criteria.

Genome sequencing & assembly

We performed DNA extraction using the Qiagen DNeasy® Blood & Tissue kit (Qiagen) with a Qiacube® Connect (Sigma), following manufacturer's instructions (Qiagen). In addition, we pre-treated *S. aureus* isolates with lysostaphin (0.2 mg/ml) (Sigma) (Schindler & Schuhardt, 1964) at 37°C for two hours to ensure cells were effectively lysed before extracting DNA. We quantified the concentration of DNA extracts using the Quant-iT™ PicoGreen™ dsDNA assay kit (Invitrogen), as per manufacturer's instructions. We sent isolates to the Wellcome Trust Centre for Human Genetics, University of Oxford for whole-genome sequencing. Multiplexed genomic DNA libraries were prepared, quality checked, and sequenced over one unit of a flow cell. Sequencing was performed using an Illumina NovaSeq 6000 machine generating 151-bp paired-end reads.

We assembled the sequencing reads using the Velvet assembly software (v1.2.10) (Zerbino & Birney, 2008). We then sampled multiple K-mer lengths and automatically searched for an optimum coverage cut-off value using the VelvetOptimiser software (v2.2.4) (https://github.com/tseemann/VelvetOptimiser). We sampled all odd numbered K-mers between 71 and the sequencing read length of 151. We used the default Velvet and VelvetOptimiser parameters, except that assemblies were not scaffolded and instead we applied a minimum contig length of 200 bases. We uploaded FASTQ files to the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena), to the project PRJEB53461 (ERP138263). Run identifiers are ERR9834854 to ERR9834949.

Data analysis

Statistical analysis - phenotypic inhibition profiles and longitudinal strains dynamics

We initially characterised phenotypic inhibition data to determine the inhibitory activity of each strain against each competitor type: positive indicator (*C. fimi*), intraspecific competitors, and interspecific competitors (see 'Supplementary methods - data analysis - section 1' for more details). To test whether there was a significant difference in the prevalence of strains displaying intraspecific versus interspecific inhibition, we used McNemar's test for paired, binary data **(**McNemar, 1947**;** Pembury Smith & Ruxton, 2020).

Next, we characterised longitudinal strain dynamics data to identify patterns of differential strain success. Specifically, we calculated a measure for: i) colonisation persistence ('long-term success'); ii) the ability of each strain to displace other strains ('short-term success') (see 'Supplementary methods – data analysis – section 2' for more details'). We then mapped the ability of each strain to inhibit the positive indicator to its longitudinal strain dynamics, to test whether inhibitory activity was associated with differential strain success. To test whether inhibitory strains were significantly associated with more persistent colonisation, we used a generalised linear model (GLM) with 'inhibitory activity' (inhibitor/non-inhibitor) as the explanatory variable and the 'mean proportion of host time points colonised' as the response variable. Proportion data was arcsine square root transformed to improve normality. To test whether inhibitory strains were significantly associated with successfully displacing other strains, we used a binomial GLM with 'inhibitory activity' (inhibitor/non-inhibitor) as an explanatory variable, and 'displacement potential' (displacer/non-displacer) as a binary response variable. See 'Supplementary methods - data analysis - section 2' for more details on all analyses performed.

For each analysis, we performed multiple variations to check the robustness of results. Specifically, we varied the: i) number of time point samples analysed; ii) criterion for the number of consecutive time points required for a strain to be considered 'lost' from the nasal cavity; iii) participant types (co-colonised and single-strain) analysed; iv) criterion for what constitutes an independent data point (each unique strain in the collection, versus each 'strain x participant' interaction); v) inclusion of a phylogenetic non-independence control; vi) competitor being used to determine 'inhibitory activity'. See 'Supplementary methods – data analysis - section 3' for more details on all analysis variations.

Genomic analysis – mining for bacteriocin gene clusters (BGCs)

To identify BGCs, we used BAGEL4 (van Heel *et al.*, 2018) and antiSMASH 6.0 (Blin *et al.*, 2021), two of the most widely used and effective computational tools at predicting BGC presence and identity (Russell & Truman, 2020) (see 'Supplementary Methods - genomic analyses' for further software details). We extracted bacteriocin gene sequence data from both software for subsequent analysis in Geneious Prime 2022.1.1 and EMBOSS Transeq (Rice *et al.*, 2000; Goujon *et al.*, 2010). For further details about the genomic analyses conducted in this study, see 'Supplementary Methods - genomic analyses'.

Phylogenetic analysis

- *Creating phylogenies and genetic distance trees*

We used SeqSphere+ (8.4) (Ridom, GmbH) to determine the genetic relationships between *spa*-types using the BURP (based upon repeat pattern) clustering algorithm (Mellmann *et al.*, 2007). We then used SeqSphere+ (8.4) (Ridom, GmbH) to create a minimum-spanning genetic distance tree (Fig. 1A) and a neighbour-joining tree using *spa* gene sequences (Fig. S2). We used FigTree (v1.4.4) to midpoint root the neighbour-joining *spa* gene phylogeny and convert it to nexus format for use in the phylogenetic analyses described below. We confirmed the accuracy of the neighbour-joining *spa* gene tree, by generating two further phylogenies using the whole-genome sequence data obtained from a subset of 89 genomes in the collection. Firstly, we used SeqSphere+ (v8.4) (Ridom, GmbH) to create a core-genome MLST (cgMLST) neighbour-joining tree based on the genome-wide allelic profiles of 1861 loci (Fig. S8). Secondly, we used REALPHY (v1.13) (Bertels *et al.*, 2014) to align our assembled *S. aureus* contigs to the complete reference genome MSSA 476 using Bowtie2 (v2.5.1) (Langmead & Salzberg, 2012) and estimate a maximum likelihood (ML) tree using PhyML (v3.0) (Fig. S9) (Guindon *et al.,* 2010). The neighbour-joining *spa* gene phylogeny was consistent with both whole-genome phylogeny approaches.

- *Phylogenetic signal analysis*

To test whether inhibitory activity displayed a phylogenetic signal, we used the function ‘phylo.d’ (v1.0.1) in R from the package ‘caper’ (Fritz & Purvis, 2010). Phylo.d calculates a measure of phylogenetic signal (*D* statistic) for binary traits. A *D* value equal to 1 represents a trait with relatively low phylogenetic signal, i.e., a trait that is randomly distributed across the phylogeny with no conserved phylogenetic signal; a *D* value equal to 0 represents a trait with relatively high phylogenetic signal, i.e., a trait that displays phylogenetic clumping (Fritz & Purvis, 2010). *D* values can also be <0 when the binary trait is extremely clumped, and >1 when the binary trait is overdispersed (Fritz & Purvis, 2010). Therefore, the more phylogenetically clumped a binary trait is, the lower its *D* value. In addition to calculating the *D* statistic, phylo.d tests whether it significantly differs from 1 (i.e. random association) and 0 (i.e. phylogenetic clumping). In our analysis, we included phylogeny as the explanatory variable and the presence/absence of either: i) inhibitory activity; ii) a BGC signature, as the response variable.

- *Controlling for phylogenetic non-independence*

Statistical models used in biology usually require data points to be independent from one another, however shared ancestry means that closely related bacterial strains are more likely to share similar traits than distantly related strains (Harvey & Pagel, 1991). To control for phylogenetic non-independence between strains in certain analyses, we used a Markov chain Monte Carlo generalized linear mixed-effects model (MCMCglmm) in R, with phylogeny controlled as a random effect. The MCMCglmm approach has previously been described in detail (Hadfield, 2010; Hadfield, 2019) (See 'Supplementary methods - data analysis - section 3' for more details).

Results

Inhibitory activity is not displayed by the majority of *S. aureus* strains and is phylogenetically dispersed

To test for the prevalence of inhibitory activity in *S. aureus*, we first screened all 389 longitudinally-sampled natural isolates (64 strains) from the nasal cavity of 40 human participants against a known positive indicator for *S. aureus* bacteriocin production: *C. fimi* (de Oliveira *et al.,* 1998; Coelho *et al.,* 2007). Overall, the majority of *S. aureus* strains did not display inhibitory activity against *C. fimi*: 26.6 % of strains (17/64) and 20.6 % of isolates (80/389) were able to inhibit the growth of the indicator (Fig. 1A; Fig. S1). At a whole-participant level, 37.5 % (15/40) carried at least one inhibitory strain (Fig. 1A). Inhibitory activity did not increase in prevalence under nasally-relevant stress conditions (Table. S1).

S. aureus strains did not display a significant phylogenetic signal for inhibitory activity (Fig. 1B; Fig. S2). Specifically, the estimated *D* statistic (*D* = 0.689, n = 62) did not significantly differ from 1 (i.e., expectation if a trait is randomly distributed) (p = 0.096) and did significantly differ from 0 (i.e., expectation if a trait displays phylogenetic clumping) (p = 0.016) (see 'Methods – phylogenetic signal analysis' for further information about how to interpret the *D* statistic) (Fritz & Purvis, 2010). Therefore, inhibitory strains were relatively unclustered in the *S. aureus* phylogeny, i.e., closely related strains were not more likely to be inhibitory than distantly related strains. Instead, evidence for inhibitory activity was widespread across the *S. aureus* phylogeny: 5/10 *spa*-clonal complexes (*spa*-CCs) and 5/11 unrelated *spa* singletons displayed inhibitory activity (Fig. 1B).

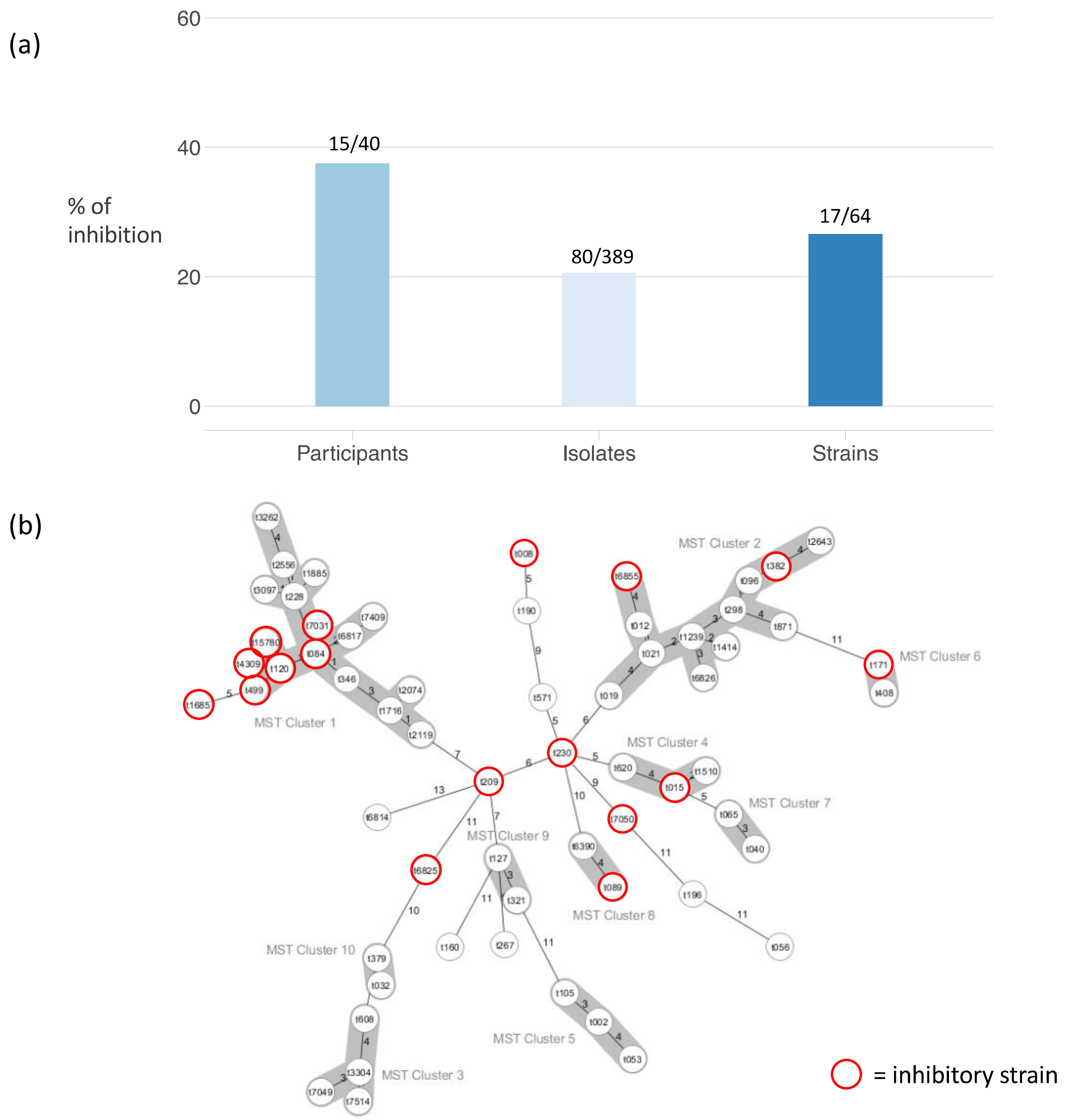


Fig 1. Inhibitory activity is not displayed by the majority *S. aureus* strains and is phylogenetically dispersed. (A) An overview of the prevalence of inhibitory activity at the participant, isolate, and strain level in *S. aureus*. 'Participants' refers to the % of participants with at least one strain displaying inhibitory activity. 'Isolates' refer to the total number of bacterial isolates in the collection, which often contains the same strain sampled over longitudinal time periods within participants. 'Strains' are defined as *spa*-types (see 'Chapter 2 – methods – *spa*-type identification'). Sample sizes are depicted above each bar. (B) A *spa* gene minimum-spanning genetic distance tree displaying the distribution of inhibitory activity amongst *S. aureus* strains. Strains in the bacterial collection are denoted as circles and are clustered into their minimum-spanning tree (MST) clusters using the BURP clustering algorithm (Mellmann *et al.*, 2007), which represent *spa*-Clonal Complexes (*spa*-CCs) of closely related strains, and are denoted by grey shading. Two *spa*-types could not be clustered, giving a total of 62 *spa*-types (see 'Chapter 2 - methods - *spa*-type identification'). Branch length corresponds to the BURP genetic distance between strains, which is also numerically labelled on each branch. Strains displaying inhibitory activity are circled in red. Throughout, inhibitory activity is defined as the ability to inhibit *C. fimi*, a known positive indicator for *S. aureus* bacteriocin production (de Oliveira *et al.*, 1998; Coelho *et al.*, 2007).

Intraspecific inhibition is rare in *S. aureus* nasal populations

To measure inhibitory activity, we first screened the 17 strains (isolates = 80; participants = 15) that inhibited growth of the positive indicator for bacteriocin production (*C. fimi*), against a positive-indicator for *S. aureus* intraspecific inhibition: Newman *ΔdltA* (Janek *et al.*, 2016). Only 2/17 strains displayed inhibitory activity against the intraspecific indicator (isolates = 8/92; participants = 2/15; strain identities = participant 637/strain t7050 & participant 2064/strain t171) (Fig. S1). To confirm that these strains were also unable to inhibit *S. aureus*, we screened a selection of 20 non-inhibitory strains (isolates = 24; participants = 19) against the intraspecific indicator (see 'Supplementary Methods – laboratory screens'). Of these strains, 0/20 performed inhibition (Fig. S3). The prevalence of intraspecific inhibition did not increase under stress-induced conditions (Table. S2).

Next, to further understand the level of inhibition between naturally-occurring strains, we screened the two strains that inhibited the intraspecific indicator, against 20 natural strains of *S. aureus* (see 'Supplementary Methods - laboratory screens' for selection criteria). Only one of the two strains could perform intraspecific inhibition against any natural strain (strain identity = participant 2064/strain t171) (Fig. 2; Table. S2). This strain successfully inhibited 19/20 natural strains (Fig. S4). The only strain it could not inhibit was the only other strain belonging to the same *spa*-CC (strain identity = participant 1307/strain t408) (Fig. 1B; Table. S3).

To confirm the reliability of the intraspecific indicator strain, we performed further screens against the same 20 natural strains of *S. aureus*. Specifically, we screened i) one representative per participant from each of the 15 strains that inhibited *C. fimi*, but did not inhibit the intraspecific indicator (isolates = 22; participants = 14); (ii) the same selection of 20 non-inhibitor strains against *C. fimi* mentioned above. As expected, we found that 0/15 and 0/20 strains could inhibit any natural *S. aureus* lawn, respectively (Table. S3). To confirm this result, we repeated this screen against three randomly selected natural *S. aureus* lawns under stress-induced conditions (see 'Supplementary methods – laboratory screens'), and found no increase in the prevalence of intraspecific inhibition (Table. S3). Therefore, we estimate that only ~3% (2/64) of *S. aureus* strains from the human nasal cavity have the capacity to perform any intraspecific inhibition, with only ~1.6% (1/64) performing inhibition against naturally-occurring intraspecific competitors.

S. aureus **more frequently inhibits interspecific competitors**

We next determined whether *S. aureus* uses bacteriocins to target interspecific competitors by screening all *S. aureus* isolates (isolates = 389; strains = 64; participants = 40) against three different species, from three major phyla that commonly inhabit the nasal cavity (Proteobacteria – *Moraxella catarrhalis*; Actinobacteria – *Corynebacterium pseudodiphtheriticum*; Firmicutes – *Staphylococcus epidermidis*) (Liu *et al.*, 2015; Janek *et al.*, 2016; Krismer *et al.*, 2017).

Of the 17 strains that inhibited the growth of *C. fimi*, seven (41.2%) successfully inhibited at least one of the three interspecific species (Fig. 2). More specifically, inhibitory activity was displayed against *M. catarrhalis* (7/17 strains) and *C. pseudodiphtheriticum* (2/17 strains). None of the 17 strains showed any inhibitory activity against *S. epidermidis* (Table. S1). Interspecific inhibitory activity was, therefore, found to be more common than intraspecific inhibitory activity (McNemar's test: $\chi^2 = 4.17$, df = 1, p-value = 0.04) (Fig. 2B).

Of the 47 strains that did not inhibit *C. fimi,* none were able to inhibit any of the three species (Fig. 2). Therefore, *S. aureus* inhibitory activity generally follows a ‘nested’ pattern: only strains that inhibit *C. fimi* can inhibit *M. catarrhalis*; only strains that inhibit *M. catarrhalis* can inhibit *C. pseudodiphtheriticum*; only strains that can inhibit *C. pseudodiphtheriticum* can inhibit natural *S. aureus*. The only possible exception is strain t7050, which only inhibits *C. fimi* and the intraspecific positive indicator (Table. S1; Table. S2). Altogether, this shows *C. fimi* is a good indicator for inhibitory activity, and in the subsequent text, we refer to strains that inhibit *C. fimi* as ‘inhibitors’ and strains that do not inhibit *C. fimi* as ‘non-inhibitors’.

(a)

	Type of inhibiton assay		
Strain identity	Indicator strain	Intraspecific	Interspecific
t002			
t008	🟩		🟩
t012			
t015	🟩		
t019			
t021			
t032			
t040			
t053			
t056			
t065			
t084	🟩		🟩
t089	🟩		🟩
t096			
t105			
t120	🟩		
t127			
t160			
t171	🟩	🟩	🟩
t190			
t196			
t209	🟩		
t228			
t230	🟩		
t267			
t298			
t321			
t346			
t379			
t382	🟩		🟩
t408			
t499	🟩		
t528			
t571			
t608			
t620			
t870			
t871			
t1239			
t1414			
t1510			
t1685	🟩		🟩
t1716			
t1885			
t2074			
t2119			
t2556			
t2643			
t3097			
t3262			
t3304			
t4309	🟩		
t6390			
t6814			
t6817			
t6825	🟩		🟩
t6826			
t6855	🟩		
t7031	🟩		
t7049			
t7050	🟩		
t7409			
t7514			
t15780	🟩		

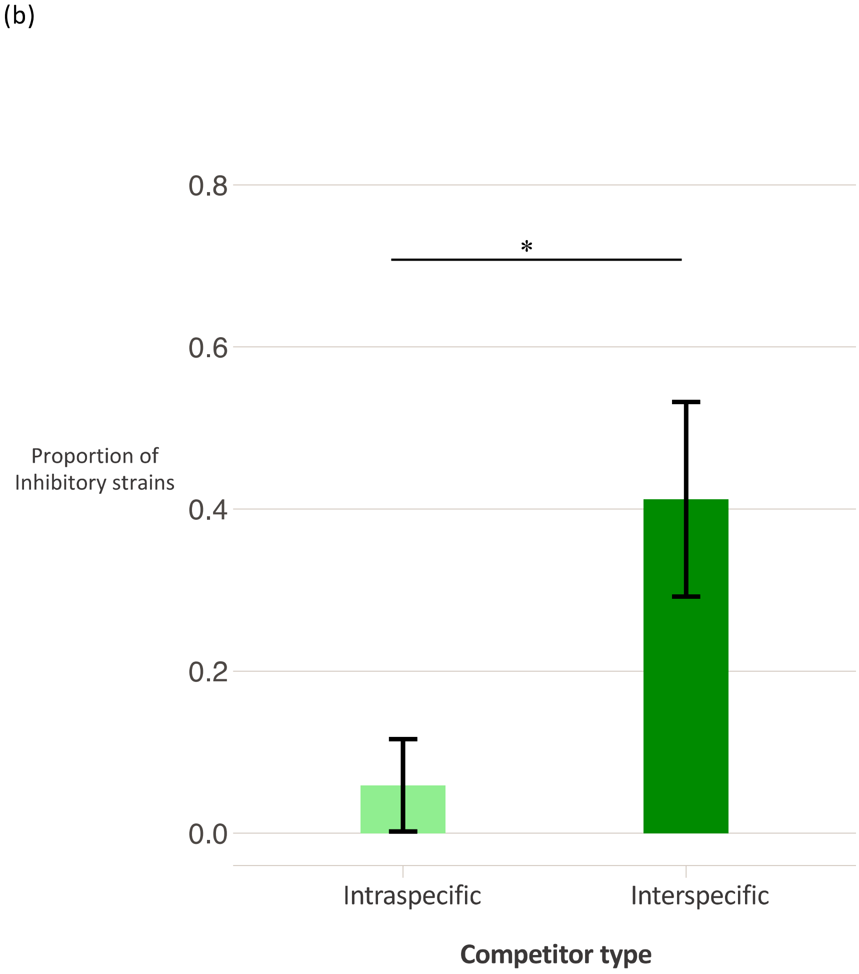


Fig 2. *S. aureus* bacteriocins have a relatively broad-spectrum of activity. (A) An overview of the inhibitory activity of 64 *S. aureus* strains in three different types of inhibition assay. ‘Indicator strain’ refers to the ability to inhibit a positive indicator for bacteriocin production (*C. fimi*); ‘intraspecific’ refers to the ability to inhibit any other natural *S. aureus* strains; ‘interspecific’ refers to the ability to inhibit any of the other three nasal commensal species. Green cells represent inhibitory activity, beige cells represent no inhibitory activity. (B) The proportion of strains that displayed inhibitory activity against *C. fimi* that can also inhibit intraspecific or interspecific competitors. ‘Intraspecific’ and ‘interspecific’ are defined as in panel A. Error bars represent the standard error of the proportion. Asterisks denote statistical significance: * = $p < 0.05$.

Inhibitory activity is stable within-hosts over time, but is not always consistent between-hosts

To characterise how inhibitory activity changes over time, we compared inhibitory activity within strains present across multiple time points in a given participant (n = 40 participants). Inhibitory activity was highly stable within strains over time (Fig. S2). When considering all 'strain x participant' interactions, where strains are present for >1 time point, only 4/75 (5.3% ± 2.6% (SE)) displayed any variation in inhibitory activity against *C. fimi*, intraspecific, or interspecific competitors over time (Fig. S2).

Next, we determined whether strains displayed variation in inhibitory activity when colonising multiple different participants. Inhibitory activity was not always consistent between-participants (Fig. S2). Of the 23 strains present in multiple participants, seven (30.4% ± 9.6% (SE)) displayed variation in inhibitory activity in at least one of the three screens against *C. fimi*, intraspecific, or interspecific competitors (Fig. S2). This result shows that the ability of a strain to display inhibitory activity in one host does not mean it will do so in all hosts.

Competitive benefit of bacteriocins

Inhibitory strains are not associated with more persistent colonisation

To examine whether inhibitory strains were associated with long-term competitive benefits, we first tested whether they were associated with more persistent colonisation (see 'Supplementary Methods - section 2 & 3' for analysis details). Inhibitory strains were not found to be more persistent colonisers of the nasal cavity (Fig. 3A). Specifically, we found that inhibitory strains did not colonise a greater mean proportion of time points compared to non-inhibitory strains in co-colonised participants (GLM: t-value = 1.148, n = 57, p = 0.256; Table. S1) (Fig. 3A). This result was consistent across multiple forms of analysis: (i) controlling for phylogenetic non-independence between strains using Bayesian statistics (MCMCglmm: posterior mean = 0.176; n = 57; p = 0.238; Table. S2); (ii) using all co-colonised time points up to 88-months (GLM: t-value = 0.884; n = 63; p = 0.38; Table. S1); (iii) removing the 'strain loss' data modification (GLM: t-value = 1.149; n = 57; p = 0.256; Table. S1); (iv) including single-strain participant data (GLM: t-value = 0.929; n = 61; p = 0.357; Table. S1); (v) using each 'strain x participant' interaction as an independent data point instead of each unique strain (see 'Supplementary Methods – section 3') (GLM: t-value = 1.034; n = 69; p = 0.305; Table. S1); (vi) focusing only on strains that also inhibit ecologically-relevant intraspecific and interspecific competitors (GLM: t-value = 1.017; n = 57; p = 0.314; Table. S1).

Second, we found no evidence that competitive exclusion explains the prevalence of single strain colonisation. We compared inhibition ability in strains colonising participants alone with co-colonising strains and found that strains colonising alone were not more likely to be inhibitory (Binomial GLM: z-value = 0.838, n = 78, p = 0.402) (Fig. S9). This result was consistent across multiple forms of analysis, including controlling for phylogenetic non-independence between strains (MCMCglmm: posterior mean = 0.966; n = 78; p = 0.338), and only including strains present in both participant categories (McNemar's test: $\chi^2 = 0.800$, df = 1, p-value = 0.371). Taken together, these results suggest that bacteriocin-mediated inhibitory activity does not improve long-term colonisation success in *S. aureus.*

Inhibitory strains are associated with successful strain displacement

To examine whether inhibitory strains were instead associated with short-term competitive benefits, we tested whether they were associated with the ability to successfully displace other co-colonising strains. Inhibitory strains were significantly more likely to perform strain displacement than non-inhibitory strains (Binomial GLM: z-value = 2.482, n = 57, p = 0.013; Table. S6) (Fig. 3b) (see 'Supplementary Methods - section 2 & 3' for analysis details). This result was consistent across multiple forms of analysis, specifically: (i) controlling for phylogenetic non-independence between strains using Bayesian statistics (MCMCglmm: posterior mean = 2.88; n = 57; p = 0.008; Table. S7); (ii) using all co-colonised time points up to 88-months (Binomial GLM: z-value = 2.325, n = 63, p = 0.020-; Table. S6); (iii) removing the 'strain loss' data manipulation (Binomial GLM: z-value = 2.267, n = 57, p = 0.0234; Table. S6) (iv) using each 'strain x participant' interaction as an independent data point instead of each unique strain (Binomial GLM: z-value = 2.689, n = 69, p = 0.0072; Table. S6) (v) focusing only on strains that also inhibit ecologically-relevant intraspecific and interspecific competitors (Binomial GLM: z-value = 2.007, n = 57, p = 0.0447; Table. S6). In addition to strain-level displacement analyses, we have performed isolate-level displacement analyses and found a similar pattern (see 'Supplementary Methods – section 2' and 'Supplementary results – Table. S8'). Taken together, inhibitory activity appears to provide a short-term competitive benefit in *S. aureus*, by allowing inhibitory strains to successfully displace co-occurring strains in the nasal cavity.

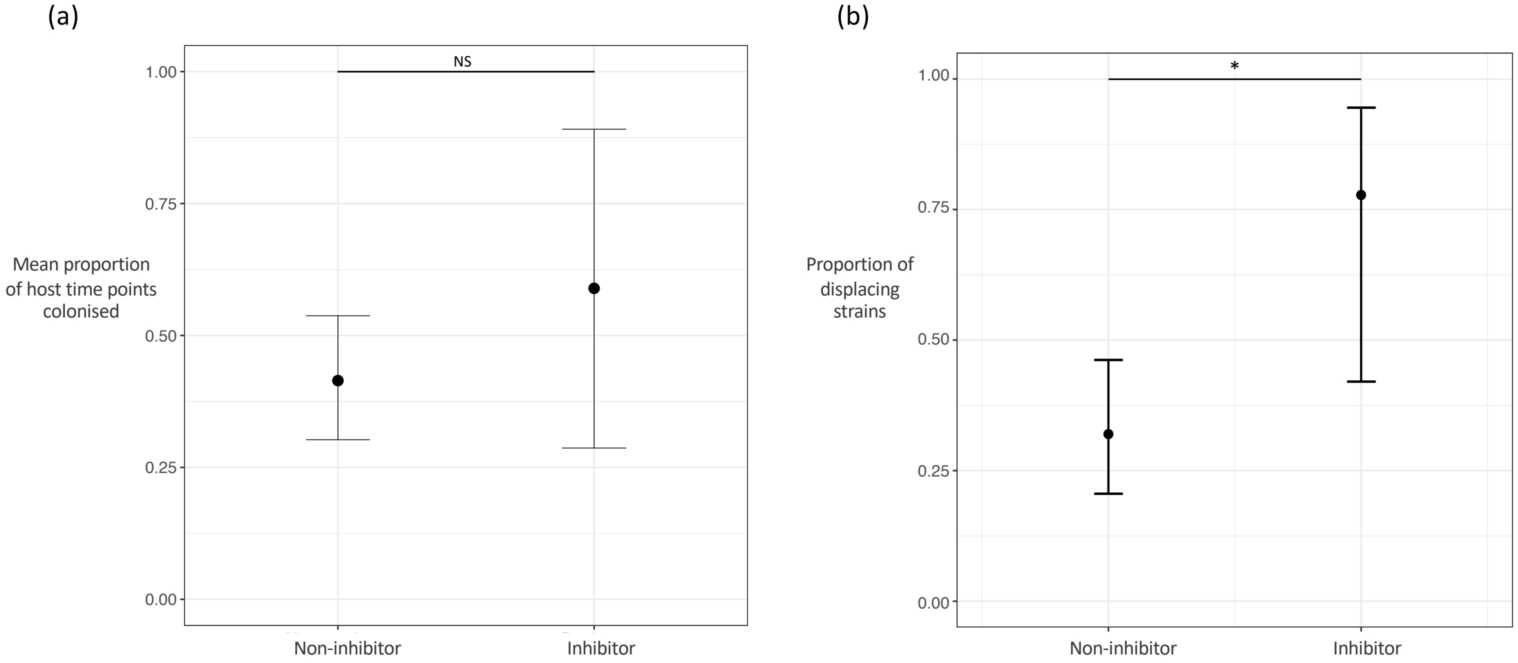


Fig 3. Inhibitory strains are not associated with more persistent colonisation, but are associated with successful strain displacement. In both panels, the x-axis categorizes strains into those that display ('inhibitor') and do not display ('non-inhibitor') inhibitory activity against the positive indicator: *C. fimi.* (A) The y-axis represents colonisation persistence, measured as the mean proportion of host time points colonised. Proportion data was arcsine square root transformed to improve normality and back-transformed for data visualization. (B) The y-axis represents strain displacement ability, measured as the proportion of displacing strains amongst inhibitor and non-inhibitor categories. In both panels, only data points from co-colonised participants, up to 24-months of sampling were included (see 'Supplementary results – Table 4-8' for alternative forms of analysis). Error bars represent 95% confidence intervals. Asterisks denote statistical significance: * = p < 0.05; NS = p > 0.05.

Five Bacteriocin Gene Clusters (BGCs) were associated with inhibitory activity

To determine the genetic underpinnings of our observed inhibitory phenotypes, we mined a selection of whole-genome sequences for BGCs associated with inhibitory activity in *S. aureus.* We identified BGC hits in 30% of inhibitory strains (40% of inhibitory genomes) (Table. S9). BGC hits associated with inhibitory activity can be divided into four distinct BGC types: Lanthipeptide class I, Lanthipeptide class II, Linear Azole containing Peptide (LAP), and a ribosomally synthesised and post-translationally modified peptide (RiPP)-like putative bacteriocin (Fig. 4; Table. S9). All four BGC types can be defined as ‘Class I’ bacteriocins, as they contain post-translational modifications (Bastos *et al.*, 2009). The four BGC types are predicted to code for five different core bacteriocins, which displayed inhibitory activity against different sets of competitors (Fig. 4; Table S9). This suggests that *S. aureus* uses different bacteriocin strategies in the nasal cavities of different hosts.

To test the reliability of our laboratory screens to detect known BGCs, we also mined the genomes of a selection of non-inhibitory strains. We found BGCs predicted to be active in only 6% of non-inhibitory strains (6% of non-inhibitory genomes), all of which were Lanthipeptide class I bacteriocins (Fig. 4; Table. S9) (See ‘Supplementary Methods – genomic analyses’ for more details). This shows that our laboratory screens were generally accurate in identifying strains containing known BGCs.

We next determined the phylogenetic distribution of BGCs in *S. aureus.* As seen with phenotypic inhibitory activity, we found that identifiable BGCs did not display a significant phylogenetic signal in *S. aureus* (Fig. S7). Specifically, the estimated *D* statistic (*D* = 0.694, n = 62) did not significantly differ from 1 (i.e., expectation if a trait is randomly distributed) (p

= 0.093) and did significantly differ from 0 (i.e., expectation if a trait displays phylogenetic clumping) (p = 0.008) (see 'Methods – phylogenetic signal analysis' for further information about how to interpret the *D* statistic) (Fritz & Purvis, 2010). We also found that each individual genome, strain, and *spa*-Clonal Complex (*spa*-CC) only ever carried one type of BGC associated with inhibition (Table S9; Fig. S7).

We next examined the stability of BGCs, by determining whether entire BGCs, or genes within BGCs, were gained or lost within strains over the course of sampling. As expected from the high stability of bacteriocin phenotypes within strains over time (Fig. S2), we found that BGCs were highly conserved within strains (Table S9). Specifically, we found no differences in BGCs within strains over time when examining the: i) presence/absence of whole BGCs; ii) presence/absence of bacteriocin-related genes within BGCs (Table S9). Taken together, the high stability of phenotypic inhibitory activity within strains over time is in line with the lack of genetic changes in BGCs over time, at least in strains with identifiable bacteriocins.

(a)

Bacteriocin Gene Cluster (BGC) Type	Prediction of bacteriocin identity	Bacteriocin class	% strain prevalence	Mapping to experimental data *C. fimi* Inhibitors	 Non-inhibitors	 Interspecific inhibitors	 Intraspecific inhibitors
Lanthipeptide (class I)	Bsa	Class I	6.7% (3/45)	✓	✓	✓	✗
Lanthipeptide (class II)	Bicereucin BsjA2-like	Class I	2.2% (1/45)	✓	✗	✓	✗
	Staphylococcin C55	Class I	2.2% (1/45)	✓	✗	✓	✓
LAP	Listerolysin S-like	Class I	2.2% (1/45)	✓	✗	✓	✗
RiPP-like bacteriocin	Putative bacteriocin	Class I	2.2% (1/45)	✓	✗	✗	✗

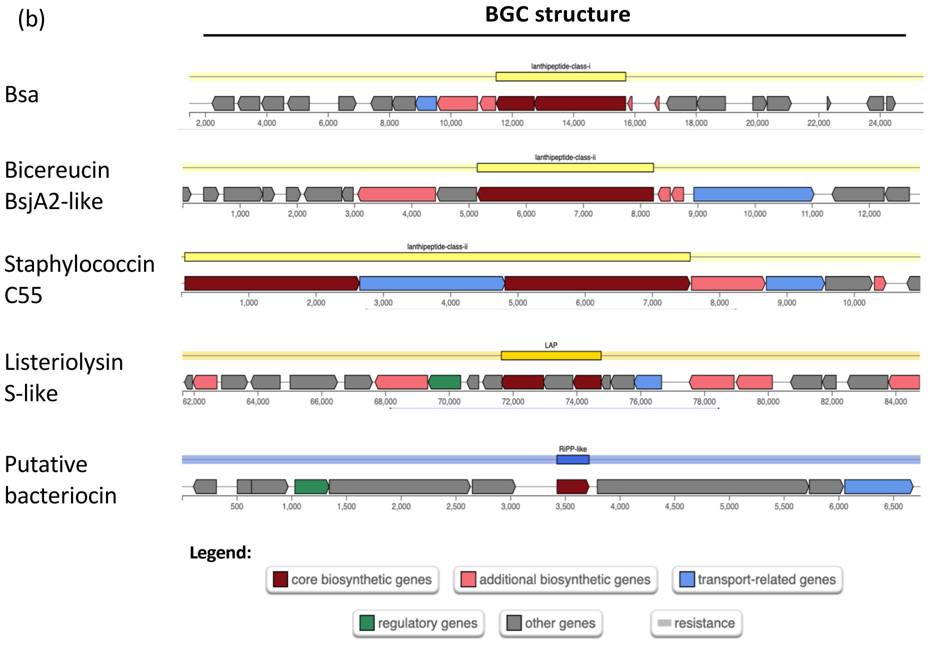


Fig 4. Five Bacteriocin Gene Clusters (BGCs) were associated with inhibitory activity in *S. aureus*. (A) An overview of the BGCs identified in *S. aureus* associated with inhibitory activity. Each row represents a unique BGC type. For each BGC type, we provide a prediction of the core bacteriocin, it's bacteriocin class, and its prevalence amongst strains that were whole-genome sequenced from the collection. All data was obtained from BAGEL4 (Van Heel *et al.,* 2018) and/or antiSMASH 6.0 (Blin *et al.,* 2021) software (see 'Supplementary methods – genomic analyses' for more details). Each BGC type is then mapped to our laboratory screen data to determine its activity spectra. A BGC is assigned a tick if it is ever associated with the described strain category, and a cross if it is never associated with the strain category. (B) A schematic representing the genetic structure of each BGC type following analysis in antiSMASH 6.0 (Blin *et al.,* 2021). Genes are colour coded based on their functional annotation, including: core biosynthetic genes, additional biosynthetic genes, transport-related genes, regulatory genes, other genes, and resistance genes. For more details about the genomic analyses, see 'Supplementary Methods – genomic analyses' and Table S9.

Discussion

Overview

We provide evidence for a role of bacteriocins in mediating competitive strain dynamics in natural populations of bacteria – but perhaps not in the simple way we might expect from lab experiments or theoretical models. We show that inhibitory activity appears to provide S. *aureus* with a short-term competitive benefit in the human nasal cavity: inhibitory strains are more likely to displace other *S. aureus* strains during co-colonisation (Fig. 3B), despite inhibitory activity not being displayed by the majority of strains (Fig. 1) and targeting interspecific over intraspecific competitors (Fig. 2). We have also provided evidence for the likely underlying mechanism of inhibition, by identifying five bacteriocin gene clusters (BGCs) associated with phenotypic inhibitory activity (Fig. 4).

Evidence for the role of bacteriocins in determining competitive strain dynamics

Despite many theoretical and laboratory experiments supporting a key role for bacteriocins in determining competitive strain dynamics (Levin, 1988; Frank, 1994; Riley & Gordon, 1996, 1999; Gordon & Riley, 1999; Kerr *et al.,* 2002; Gardner *et al.,* 2004; Krikup & Riley, 2004; Waite & Curtis, 2009; Kommineni *et al.,* 2015; Libberton *et al.,* 2015; Kawada-Matsuo *et al.,* 2016; Lehtinen *et al.,* 2022), our study is one of the first to provide evidence of this in longitudinally-sampled natural populations of bacteria. The lack of evidence for the role bacteriocins play in natural populations can in part be explained by the difficulty associated with collecting large natural bacterial isolate collections that track strain dynamics over time. Indeed, the majority of studies providing insight into the role of bacteriocins in natural populations only sample a single time point snapshot from a population (Cordero *et al.,* 2012; Hawlena *et al.,* 2012; Perez-Gutierrez *et al.,* 2013; Kinkel et al., 2014; Abrudan *et al.,* 2015;

Bruce *et al.,* 2017). However, even studies that have tracked natural populations over time have been unable to identify a clear competitive benefit associated with bacteriocin production (Sears *et al.,* 1950; Ghoul *et al.,* 2015; Kraemer *et al.,* 2017). For example, early studies of *E. coli* provided conflicting evidence on the role of bacteriocins in determining differential strain success in the human gut, which is yet to be resolved (Sears *et al.,* 1950; Branche *et al.,* 1963). More recently, Ghoul *et al.* (2015) mapped the phenotypic and genomic bacteriocin profiles of *P. aeruginosa* from the cystic fibrosis lung infections to longitudinal strain dynamics data, but found no evidence for the role of bacteriocins in determining competitive strain dynamics. Interestingly, evidence for a competitive benefit associated with bacteriocin production was provided by Holt *et al.* (2013) in a less well-studied species, *Shigella sonnei*. Holt *et al.* sampled *S. sonnei* across endemic populations over a 15-year period and found that the most dominant strains contained a signature of positive selection for a plasmid (pDPT1) encoding a bacteriocin (E5-type colicin). They subsequently performed phenotypic assays to show this bacteriocin was active against ecologically-relevant competitors (Holt *et al.,* 2013). While they did not longitudinally track the same populations of bacteria over time to obtain high-resolution strain dynamics, as we did in this study, the work on *S. sonnei* is one of the only other studies to provide evidence for the role of bacteriocins in determining differential strain success, by tracking bacteriocin producers in natural populations over time.

Why are bacteriocins associated with a short-term, rather than a long-term, benefit?

We find that inhibitory activity is not associated with a long-term benefit of allowing inhibitory strains to more persistently colonise the nasal cavity, but is associated with a short-term benefit of being more likely to displace other strains over time (Fig. 3). The absence of a long-term benefit associated with bacteriocin production could possibly be explained by it generally being considered to be a costly trait (Cornforth & Foster, 2013; Maldonado-Barragán & West, 2020).

This cost has led to many theoretical and laboratory experiments predicting that instead of always allowing producers to outcompete non-producers, bacteriocin production can give rise to different social evolutionary dynamics, such as cyclical dynamics analogous to the 'rock-paper-scissor' game between bacteriocin 'producing', 'immune', and 'sensitive' strains (Kerr *et al.*, 2002; Gardner *et al.,* 2004; Kirkup & Riley, 2004). This occurs when a bacteriocin producing strain initially outcompetes a sensitive strain, however due to the metabolic cost of bacteriocin production, the producing strain is subsequently outcompeted by a strain that does not produce the bacteriocin but is immune to it. Finally, the immune strain is outcompeted by a sensitive strain that does not pay the cost of producing either the bacteriocin or immunity protein (Kerr *et al.,* 2002; Riley & Chavan, 2007). While, like all previous studies of bacteriocins in natural populations, we do not find direct evidence for rock-paper-scissor dynamics, we do find that inhibitory activity is only associated with short-term displacement benefits, instead of long-term competitive benefits. Bacteriocins are also not expected to be beneficial under all conditions (Cornforth & Foster, 2013), and potential fluctuations in environmental conditions in the nasal cavity over time, known to be possible in host-associated habitats exposed to the external environment (Krismer *et al.* 2017), could cause bacteriocins to be beneficial at some time points but not others in some hosts. More generally, this finding highlights the importance of screening longitudinally-sampled natural populations to identify the exact ecological and evolutionary role of bacteriocins, as simply screening the most persistent strains would have failed to identify the associated competitive benefit in this study.

Why is bacteriocin-mediated intraspecific inhibition rare in S. aureus?

A caveat to the competitive benefit observed in this study is that we find strains capable of inhibiting intraspecific competitors are rare (1/64 strains) (Fig. 2), and that intraspecific inhibition was always associated with isolates carrying the staphylococcin C55 (Navaratna *et al.,* 1999) bacteriocin gene cluster (BGC) (Fig. 4; see 'Discussion – mapping bacteriocin genotype to phenotype' for further discussion). This low prevalence of intraspecific inhibition is in line with previous studies screening natural populations of *S. aureus* for bacteriocin activity (Giambiagi-deMarval *et al.,* 1990; de Oliveira *et al.,* 1998; Bastos *et al.,* 2009; Janek *et al.,* 2016). For example, Janek *et al.* (2016) found that 0/19 (0%) *S. aureus* strains isolated from the human nasal cavity could inhibit the laboratory strains *S. aureus* Newman, and that only 2/19 (11%) could inhibit *S. aureus* Newman *ΔdltA*, the same intraspecific inhibition positive indicator strain used in our study. Also in line with the results of our study, this low prevalence of intraspecific inhibition was observed despite interspecific inhibition being relatively more prevalent, as 10/19 (53%) *S. aureus* strains were observed to inhibit at least one interspecific competitor. Studies of other species have also observed that intraspecific inhibition occurs at low frequencies, for example, a study of *Pseudomonas fluorescens* soil isolates found that only ~7% of pairwise isolate interactions were inhibitory (Bruce *et al.,* 2017). Low levels of intraspecific inhibition have also been observed in *Bacillus spp.* isolated from soil environments (Perez-Gutierrez *et al.,* 2013) and in *Vibrio spp.* isolated from ocean environments (Cordero *et al.,* 2012). However, this low prevalence of intraspecific inhibition stands in contrast with studies of other well-studied species from a bacteriocin perspective, such as *P. aeruginosa* and *E. coli*, where up to ~100% and ~70% of strains can display intraspecific inhibitory activity, respectively (Gordon *et al.,* 1998; Ghoul *et al.,* 2015). It also contrasts with high levels of intraspecific inhibition identified in some gram-positive species, such as in *Lactococcus spp*. and other lactic acid bacteria (LAB) (Klaenhammer 1988; Cintas *et al.,* 2001; Mokoena, 2017), and *Streptomyces spp.* (Westhoff *et al.,* 2021).

Several reasons could account for the rarity of bacteriocin-mediated intraspecific inhibition in *S. aureus*. Firstly, bacteriocin activity may be neutralised by intraspecific bacteriocin immunity. Bacteriocin immunity is generally thought to be less costly than bacteriocin production that often also requires immunity (Riley & Gordon, 1999; Kerr *et al.,* 2002), potentially making it easier to evolve than production itself. Indeed, our results provide evidence that bacteriocin immunity is prevalent in natural *S. aureus* populations, as although ~27% of strains display inhibitory activity, only ~1% can inhibit other naturally-occurring intraspecific competitors (Fig. 2). Secondly, *S. aureus* strains may not interact within hosts frequently enough to select for intraspecific inhibition: only ~18% of participants that were *S. aureus*-positive at recruitment and returned ≥12 nasal samples were co-colonised with multiple *S. aureus* strains at a single time point during the first 24-months of the original nasal carriage study (Votintseva *et al.,* 2014). Thirdly, if *S. aureus* strains do interact, they may mediate intraspecific competition using other types of chemical weaponry. *S. aureus* has been shown to use phage (Haaber *et al.,* 2016), the type VII secretion system (T7SS) (Cao *et al.,* 2016; Ulhuq *et al.,* 2020), and even quorum-sensing signalling peptides to inhibit the growth of intraspecific competitors (Novick & Geisnger, 2008). Further research is required to understand the relative importance of different types of chemical weaponry in natural populations of *S. aureus,* and bacteria more generally.

Do environmental conditions in the nasal cavity select against bacteriocins in general?

Environmental conditions in the nasal cavity may also be unfavourable for the use of bacteriocins by *S. aureus* in general. It has been suggested that bacteriocins are likely most strongly selected under intermediate nutrient conditions, due to their associated cost, rather than high or low nutrient conditions (Granato *et al.,* 2019), yet the nasal cavity is known to be a nutrient scarce environment (Krismer *et al.,* 2014). The lack of nutrients also means that population cell-densities in the nasal cavity are generally lower than in other environments, such as the human gut (Krismer *et al.,* 2017). Previous studies have shown that bacteriocin production is less prevalent in low cell-density environments compared to high cell-density environments, likely caused by bacteriocins being more effective at higher concentrations (Adams *et al.,* 1979; Chao & Levin, 1981). We are also yet to understand how microbial populations are spatially structured in the nasal cavity, and the proximity with which different strains and species interact (Yan *et al.* 2013; Krismer *et al.,* 2017). These are important factors in determining the level of competition between strains or species (Mitri & Foster, 2013; Granato *et al.,* 2019), as recently shown for the human skin microbiome (Conwill *et al.* 2022), and therefore the potential benefit to be gained by expressing bacteriocins.

Despite potential for selection against bacteriocin activity in the nasal cavity, some studies have identified a high prevalence of bacteriocin activity in other species in this habitat, such as *Staphylococcus epidermidis* (Janek *et al.,* 2016). Studies have also shown that other species produce antimicrobials that inhibit the ability of *S. aureus* to colonise the nasal cavity, such as the production of lugdunin by *Staphylococcus lugdunensis*, suggesting an important ecological role for diffusible toxins in this habitat (Zipperer *et al.,* 2016). It is, therefore, possible that *S. aureus* bacteriocin activity is more common in the nasal cavity under natural conditions, but we were unable to detect this activity *in vitro.* We lack a detailed understanding of how bacteriocins are regulated in *S. aureus* (de Freire Bastos *et al.*, 2020) but many bacteriocins are known to be tightly regulated in response to environmental conditions, such as nutrient stress and cellular damage (Cornforth & Foster, 2013). Despite finding that different types of nasally-relevant stress, including iron-limitation and oxidative stress, did not increase the prevalence of bacteriocin activity, it is possible that the combination of environmental factors in the nasal cavity in natural populations could be required for this effect.

Why do S. aureus produce relatively broad-spectrum bacteriocins?

S. aureus strains that displayed inhibitory activity more frequently produced broad-spectrum bacteriocins that inhibited interspecific competitors. The evolution of bacteriocin activity spectrum has received very little attention from an evolutionary perspective, but certain conditions have been proposed to select for broad-spectrum over narrow-spectrum bacteriocins. For example, when multiple phylogenetically-diverse species with largely non-overlapping metabolic requirements compete for a common resource, such as space or an essential nutrient in low supply (Krismer *et al.,* 2017). This could be particularly relevant in the nasal cavity, as competition for space *via* mucosal attachment sites and essential scarce nutrients, such as iron, is expected to be fierce (Krismer *et al.,* 2014, 2017). Bacterial strains at high abundances relative to other strains in the environment are also expected to be under selection for broad-spectrum bacteriocin activity (Palmer & Foster, 2022). This is because highly abundant strains can afford to make bacteriocins that broadly kill many competitors, whereas a less abundant strains must focus its resources on targeting its most significant competitor (Palmer & Foster, 2022). Despite bacterial cell abundances being relatively low in the nasal cavity compared to other parts of the human microbiome (Yan *et al.,* 2013; Liu *et al.,* 2015), *S. aureus* is known to be present at a relatively high abundance compared to other species in the nasal cavity of some hosts (Frank *et al.,* 2010; Yan *et al.* 2013; Liu *et al.,* 2015), which could in part explain the broad-spectrum nature of their bacteriocins.

How could broad-spectrum bacteriocins provide an intraspecific competitive benefit?

How, then, can interspecific inhibition provide the short-term competitive benefit of successful intraspecific strain displacement observed in this study? One possible explanation is that interspecific inhibition would allow *S. aureus* strains to outcompete and clear interspecific competitors in the nasal cavity, altering the nasal biotic environment and creating empty niches/patches for them to exploit. It has also been suggested that bacteriocin production could provide a mechanism for strains to spatially segregate themselves within their environment, given that inhibitory concentrations of bacteriocins are more likely to be reached in the close vicinity of the producer in structured environments (Chao & Levin, 1981; Heilbronner *et al.,* 2021). Once interspecific inhibitor strains begin to dominate the nasal cavity, it's possible they could outcompete and displace other *S. aureus* strains through other mechanisms of competition, for example exploitative competition for nutrients or space. It is widely accepted that bacteria utilise multiple mechanisms to mediate competition (Ghoul & Mitri, 2016; Stubbendieck & Straight, 2016; Granato *et al.,* 2019), however further study is required to understand how these mechanisms can be used synergistically in natural populations. Another possible explanation, as previously discussed, is that *S. aureus* intraspecific inhibition is more common in the nasal cavity than it is *in vitro*, due to it requiring complex environmental conditions to upregulate its activity, which would explain successful strain displacement *via* direct bacteriocin inhibition.

Mapping bacteriocin genotype to phenotype

We identified four distinct bacteriocin gene cluster (BGC) classes in *S. aureus*, corresponding to five different bacteriocin core gene predictions (Fig. 4). Each bacteriocin displayed inhibitory activity against a different set of competitors, suggesting each represents a different competitive strategy (Fig. 4). Interestingly, even though we found bacteriocin activity and their associated BGCs were phylogenetically dispersed (Fig. 1; Fig. S7), each *S. aureus* strain and *spa*-Clonal Complex (*spa*-CC) was restricted to only carrying one type of bacteriocin associated with inhibition, even when colonising different participants. Each individual genome also only ever contained one BGC associated with inhibition, which was always highly stable, never being gained or lost within a strain during the sampling time frame (Table. S9).

These results contrast to those in many other species, which have been found to carry a 'cocktail' of bacteriocins types, both within individual genomes and within the same clonal complex (Gordon & O'Brien, 2006; Ghoul *et al.*, 2015). Having said this, we only identified BGCs predicted to be active in 30% of strains (40% of genomes), and we detected no other secondary metabolite gene cluster types known to cause antimicrobial effects (Table. S9). Further work is required to fully elucidate the genetic underpinnings of bacteriocin production in natural populations of *S. aureus*, as it appears likely that many bacteriocins remain undetected and uncharacterised. This emphasizes the benefit of using phenotypic data, in combination with genomic data, when trying to determine the prevalence and evolutionary importance of bacteriocins in a given species.

Mapping our genomic data to the observed phenotypic inhibition profiles also allows us to determine the observed activity spectrum of each predicted bacteriocin in our study (Fig. 4a). Moreover, for BGC hits that have been previously identified and tested for inhibitory activity, it allows us to determine whether the activity spectra match that observed in previous studies. Of particular note, the only strain capable of performing intraspecific inhibition against *S. aureus* strains from the nasal cavity in our study was found to carry a BGC predicted to be staphylococcin C55 (Fig. 4a; Table S9). Staphylococcin C55 has previously been identified to be produced by *S. aureus* and to inhibit many other *S. aureus* strains (Navaratna *et al.,* 1999; Kawada-Matsuo *et al.,* 2016). The only other BGC hit in our study that has previously been identified as a known *S. aureus* bacteriocin and has been screened for its inhibition activity is Bsa (bacteriocin of *Staphylococcus aureus*). However, the activity spectrum of Bsa based on previous studies remains unclear, as while it has been shown to inhibit other *S. aureus* laboratory strains, particularly when produced by a strain with hyperactivated expression of the *agr* quorum-sensing system (Koch *et al.,* 2014), it has also been shown to display little intraspecific inhibitory activity when screened against other naturally-occurring *S. aureus* strains (Fagundes *et al.*, 2017), as observed in our study. Further molecular and genetic work is required to definitively characterise the identity of the bacteriocins produced by the BGCs identified in our study, particularly those bacteriocins without a clear core peptide prediction or those from BGCs showing similarity to those typically found in other bacterial species. Nonetheless, mapping BGC predictions to phenotypic inhibition profiles is an important first step in determining the activity spectra of bacteriocins in natural populations.

Limitations & future directions

Our study provides a general framework that can be used by future studies to map bacteriocin activity, or any other competitive trait, to strain dynamics data to determine its evolutionary and ecological role in natural populations of microbes, from any species or habitat. There are, however, improvements which could be made in future work. Firstly, identifying correlations between bacteriocin activity and measures of success, such as the ability to persist, or displace competitors, is an important first step to understand the role of bacteriocins determining competitive strain dynamics, particularly when such correlations are found by replicating across many different genetic backgrounds. However, other factors, such as other mechanisms of competition or the ability to withstand abiotic environmental variability and host immune responses, are likely to play a role in determining differential strain success. To further understand the role of bacteriocins in determining the outcome of competition, future experimental studies should perform competition assays using naturally-occurring strains, with and without knockout mutations in bacteriocin genes (Kommineni *et al.*, 2015; Janek *et al.*, 2016; Quereda *et al.*, 2016). Secondly, this study focused on natural populations of a single species, *S. aureus*, to provide an in-depth understanding of its bacteriocin activity in many strains over longitudinal time periods. Future studies could conduct whole-community sampling, to strain- and species-level, to understand how bacteriocins can shape their composition over time. This is particularly relevant when bacteriocins are expected to have a broad-spectrum of activity. Collecting data on the relative abundance of strains/species, in addition to presence/absence data, will allow us to determine whether bacteriocin activity allows strains/species to reach relatively high abundances compared to competitors and dominate their habitat.

Conclusion & impact

Overall, given their prevalence and predicted importance in mediating microbial competition, we must now shift emphasis to understand the evolutionary and ecological roles of bacteriocins in natural populations. This study is one of the first to provide evidence for a role of bacteriocins in determining natural competitive strain dynamics over time. Importantly, we do this in an opportunistic pathogen isolated from the human microbiome. Many recent studies have shown how compositional changes in the human microbiome are a major factor in causing disease (Durack *et al.*, 2019; Heilbronner *et al.*, 2021), and other studies have begun to associate bacteriocin activity with outbreaks of infection (Holt *et al.,* 2013; Quereda *et al.,* 2016). Therefore, understanding how bacterial pathogens use bacteriocins to gain an ecological or evolutionary benefit and cause disease is of paramount clinical importance, especially if we are to effectively manipulate microbial communities using probiotic treatments (Dobson *et al.,* 2012; Cotter *et al.*, 2013; Krismer *et al.*, 2017).

Supplementary Methods – data analysis

In this section, we provide detailed methods for data collection, processing, and analysis. In section 1, we define inhibitory and non-inhibitory strains. In section 2, we explain how we map inhibition profiles to longitudinal strain dynamics to test whether inhibitory activity is associated with strain success. This involves measuring the colonisation time and displacement ability of each strain. We also measure the variation in inhibitory activity of each strain. In section 3, we explain the alternative forms of statistical analysis used for testing the robustness of our main results.

1. Characterising phenotypic bacteriocin inhibition profiles

1.1. Inhibition against a positive indicator strain - C. fimi

We labelled an *S. aureus* strain as 'inhibitory' or 'non-inhibitory' according to its ability to inhibit a positive indicator for bacteriocin production: *Cellulomonas fimi.* We defined an 'isolate' as *S. aureus* present in a single time point sample, a 'strain' as all *S. aureus* isolates from the same *spa*-type (see 'Methods – *spa*-typing'), a 'participant' as an individual in the study, and 'whole collection' as the entire collection of all participants (Fig. SM1). Given we were interested in identifying strains that were capable of inhibiting the indicator strain, we labelled a strain as being inhibitory within a participant if any of its isolates were inhibitory (Fig. SM1; strain A in participant 1 & 2). In cases where a given strain was present in multiple different participants, we labelled the strain as inhibitory if it was inhibitory in any participant (Fig. SM1; strain A in participant 1-3). The inhibitory activity of a strain can therefore be considered as a binary trait (1/0) at three levels: i) individual-isolate; ii) within-participant; iii) within-collection (Fig. SM1).

1.2. Inhibition against intraspecific and interspecific competitors

We then tested for the ability of *S. aureus* isolates to inhibit intraspecific competitors and interspecific competitors. We did this by screening *S. aureus* isolates against a selection of other naturally-occurring nasal *S. aureus* isolates representing intraspecific competitors (see 'Supplementary Methods – laboratory screens') and three ecologically-relevant species, *Moraxella catarrhalis*, *Corynebacterium pseudodiphtheriticum*, and *Staphylococcus epidermidis*, representing interspecific competitors from the nasal cavity (Liu *et al.,* 2015; Janek *et al.*, 2016; Krismer *et al.,* 2017). We defined an isolate as an 'intraspecific inhibitor' if it inhibited any other naturally-occurring strain of *S. aureus* and an 'interspecific inhibitor' if it inhibited at least one of the three ecologically-relevant interspecific competitors. After determining whether each isolate could inhibit competitors, we collapsed the data to: i) within-participant and; ii) within-collection levels, as described for the *C. fimi* screens in section 1.1. To test whether there was a significant difference in the prevalence of intraspecific inhibitor versus interspecific inhibitor strains, we used McNemar's test for paired, binary data (McNemar, 1947; Pembury Smith & Ruxton, 2020).

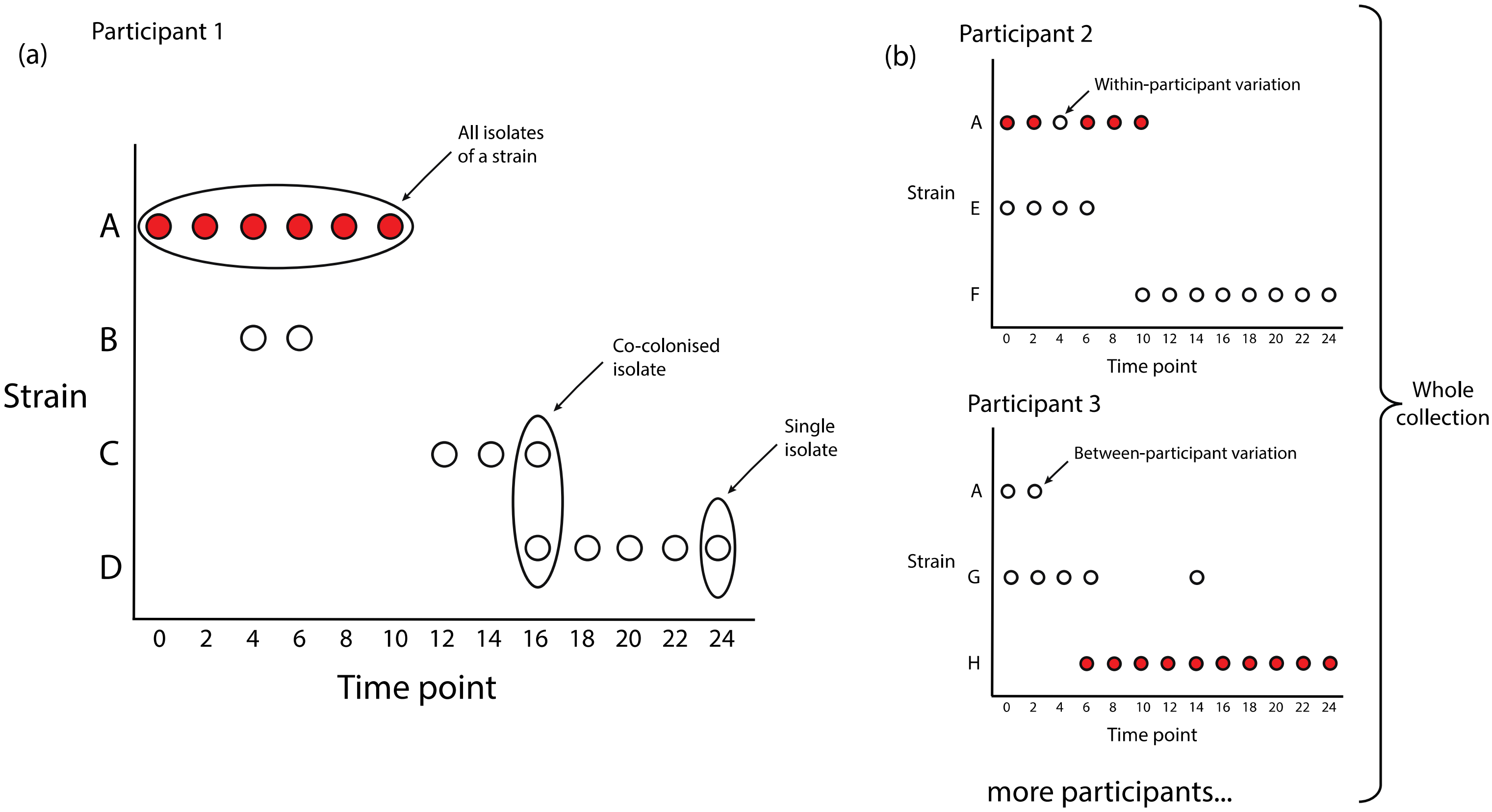


Fig SM1. Characterising phenotypic inhibition profiles. Each plot represents the strain dynamics present within an individual participant. The y-axes represent strains and the x-axes represent the time point each sample was collected. Circles represent the presence of a given strain in a time point sample. Circle colour represents the inhibition profile for an isolate against a given competitor: red = inhibitor; white = non-inhibitor. Panel (a) provides examples of co-colonised isolates, where a single time point sample contains two or more strains; and single isolates, where a single time point sample contains one strain. It also shows how we often observe the same strain at multiple time point isolates. Panel (b) shows how a given strain, e.g., Strain A, can be found in multiple participants. In participant 2, strain A displays within-strain variation in inhibitory activity, which would still qualify strain A as inhibitory within this participant, as one or more of its isolates are inhibitory. In participant 3, strain A displays between-strain variation in inhibitory activity: it displays no inhibitory activity within participant 3, but does display inhibitory activity with participants 1 and 2. Strain A would therefore be defined as an inhibitor at the whole-collection level, as it displays inhibition in one or more participants, but a non-inhibitor if focusing within participant 3.

2. Mapping phenotypic bacteriocin inhibition profiles to longitudinal strain dynamics

In this section, we map phenotypic inhibition profiles to longitudinal strain dynamics to determine whether inhibitory activity is associated with differential strain success in natural populations. Firstly, we do this by calculating measures of 'long-term' success (section 2.1), including: i) length of colonisation time; (ii) colonisation type, i.e., whether a strain colonises alone, defined as 'single-strain participants', or co-colonises with other strains, defined as 'co-colonised participants'. Secondly, we calculate measures of 'short-term' success, including strain displacement at the isolate-level and at participant-level (section 2.2). And thirdly, we quantify any observed variation in inhibitory activity within participants over time and between participants (section 2.3).

Throughout section 2, we use the strain dynamics data obtained in the John Radcliffe (JR) Hospital's S. *aureus* carriage study (see 'Methods – Oxford nasal carriage collection') for the 40 participants in our refined laboratory collection (see 'Methods – refined isolate collection for laboratory use'). For each time point in the study, the data includes: (i) whether *S. aureus* is present, and if so, the identity of all detected strains present; (ii) whether *S. aureus* was absent, denoted as 'NG' (no growth); (iii) whether a time point sample was not collected, denoted as 'NA'.

2.1. 'Long-term success' analysis

2.1.1. Colonisation persistence analysis

We measure each strain's persistence in the nasal cavity as the 'proportion of participant time points colonised' (explained in Fig. SM2a). We calculate a proportion instead of using absolute time, due to some participants being sampled for a different total number of time points. For strains present in multiple participants, we calculated the ‘mean proportion of participant time points colonised’.

To test whether inhibitory strains persisted in the nasal cavity for a greater proportion of time than those displaying no inhibitory activity, we used a generalized linear model (GLM) with ‘inhibitory activity’ (inhibitor/non-inhibitor) as the explanatory variable and the ‘mean proportion of participant time points colonised’ as the response variable. Proportion data was arcsine square root transformed to improve normality. We performed multiple variations of this analysis to check the robustness of results (see ‘Supplementary Methods – data analysis - section 3’ for explanation and Tables S4 & S5 for results).

2.1.2. Participant type analysis

We define a ‘single-strain participant’ as a participant colonised by one strain during the sampling time frame and a ‘co-colonised participant’ as a participant colonised by two or more strains during the sampling time frame (Fig. SM2b). Strains from single-strain participants can be considered to have a higher long-term success than strains from co-colonised participants, as they colonise alone and for longer time periods.

To test whether single-strain participants had a greater proportion of inhibitory strains compared to co-colonised participants, we used a generalized linear model (GLM) with ‘participant type’ (single/co-colonised) as the explanatory variable and the ‘inhibitory activity’ (inhibitor/non-inhibitor) as the response variable. We performed multiple variations of this analysis to check the robustness of results (see ‘Supplementary Methods – data analysis - section 3’ for explanation and Tables S6 & S7 for results).

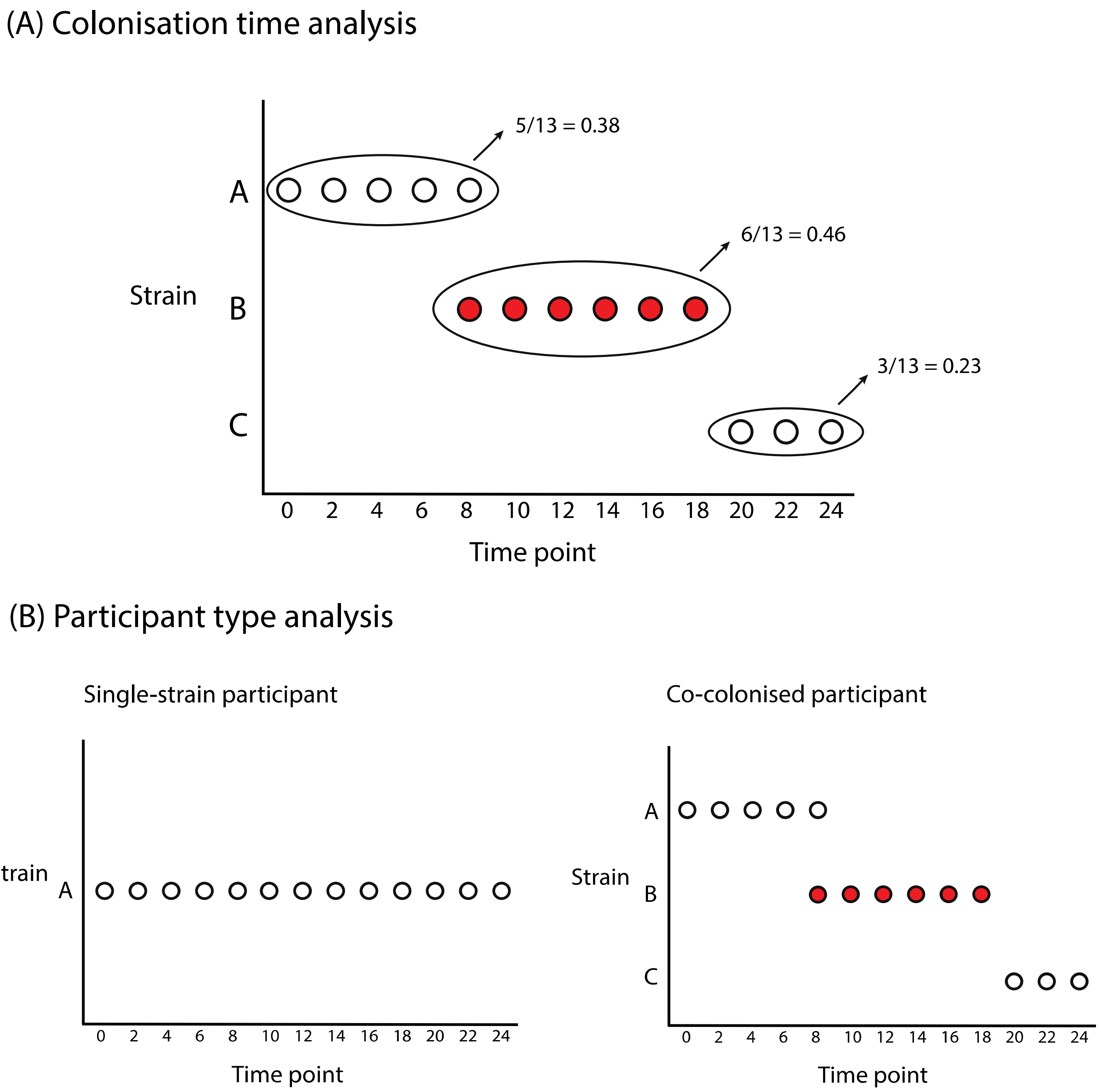


Fig. SM2. Characterising 'long-term success' from strain dynamics data. Definitions of axes, dot presence/absence, and dot colour are the same as in Fig. SM1. Panel (a) explains how we calculated the proportion of participant time points colonised for each strain. For strains present in multiple hosts, we calculated the mean proportion of host time points colonised. Panel (b) explains how we defined each participant as either a 'single-strain participant' or 'co-colonised participant'. Single-strain participants are colonised by a single strain; co-colonised participants are colonised by two or more strains.

2.2. *'Short-term success' analysis*

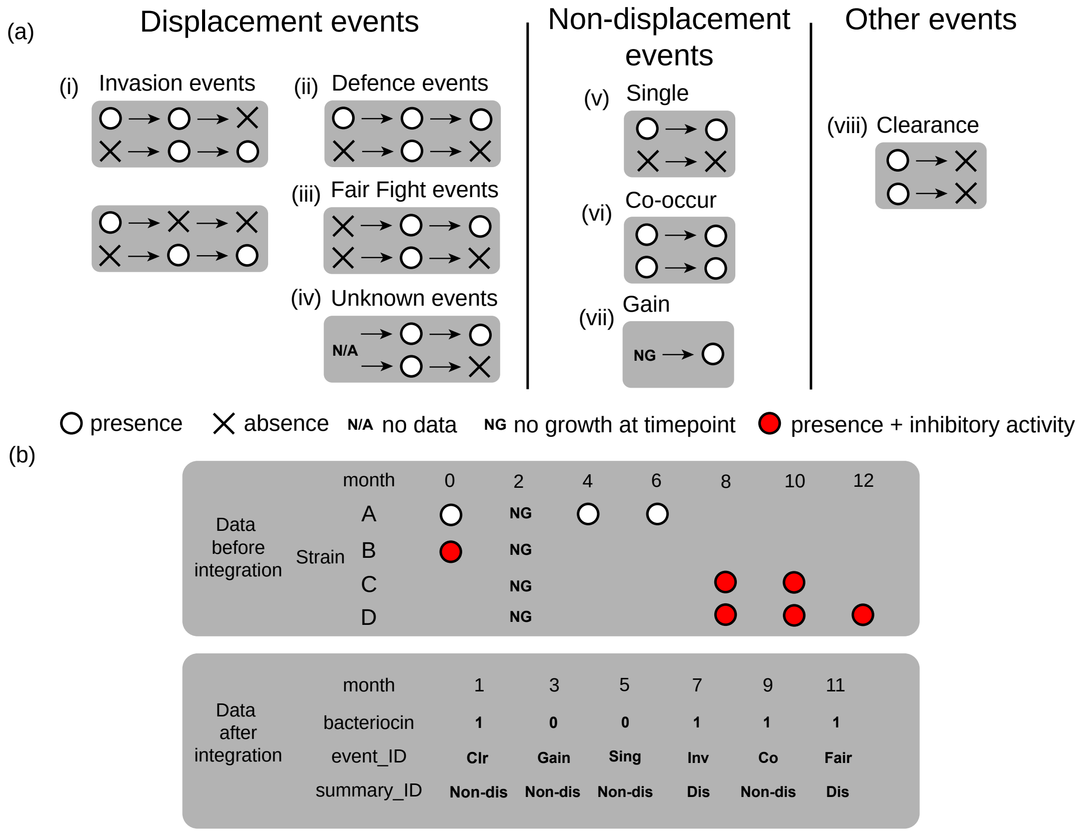
2.2.1. *Displacement analysis – isolate-level*

We identified interesting competitive strain dynamics within co-colonised participants, in that some strains appear to successfully displace (i.e. remove) other strains and continue to colonise the nasal cavity. Displacement events can occur in many different ways, including i) 'invasion', ii) 'defence', iii) 'fair fights', and iv) 'unknown' (Fig. SM3). Non-displacement events can also occur in many different ways, also outlined in Fig. SM3v-vii.

To analyse displacement and non-displacement events, we created a hypothetical time point in between each pair of consecutive time point isolates. For example, in any given participant between time points 2 and 4, we created a hypothetical time point 3 (Fig. SM3b). We assign this hypothetical time point to either contain a 'displacement' or a 'non-displacement' event (Fig. SM3b; event_ID and summary ID rows). We also labelled each hypothetical time point with whether they contained any inhibitory isolates (=1) or whether they only contained non-inhibitory isolates (=0) (Fig. SM3; bacteriocin row).

First, to test whether there was a statistically significant association between 'displacement event type' and inhibitory isolates, we used a binomial GLM with 'displacement event type' (i.e., event_ID row in Fig. 3Mb) as an explanatory variable and 'inhibitory activity' (i.e., bacteriocin row in Fig. 3Mb) as the binary response variable. Second, due to the small sample size of each displacement event sub-category, we also re-performed this analysis but with the modification of pooling all displacement event types and non-displacement event types together to give two categories (i.e., summary ID row in Fig. 3Mb). We also performed multiple variations of both analyses to check the robustness of results (see ‘Supplementary Methods - Section 3’ for explanation and Table. S8 for results).

The above analysis approach allows us to determine whether inhibitory isolates are generally associated with displacement or non-displacement events. However, this general association includes cases where inhibitors both ‘win’ and ‘lose’ displacement events. Therefore, for each displacement event involving an inhibitor, we determined whether the inhibitor ‘won’ or ‘lost’ the displacement event (i.e., whether the inhibitor persists to the next time point). To test whether inhibitory isolates were significantly associated with winning displacement events, we used a binomial GLM with ‘displacement winner identity’ (inhibitor/non-inhibitor) as the response variable and tested whether this significantly differed from a null model assuming a 50% probability of either displacement outcome (see Table. S8 for results).



	month	0	2	4	6	8	10	12
Strain	A	○	NG	○	○			
	B	●	NG					
	C		NG			●	●	
	D		NG			●	●	●

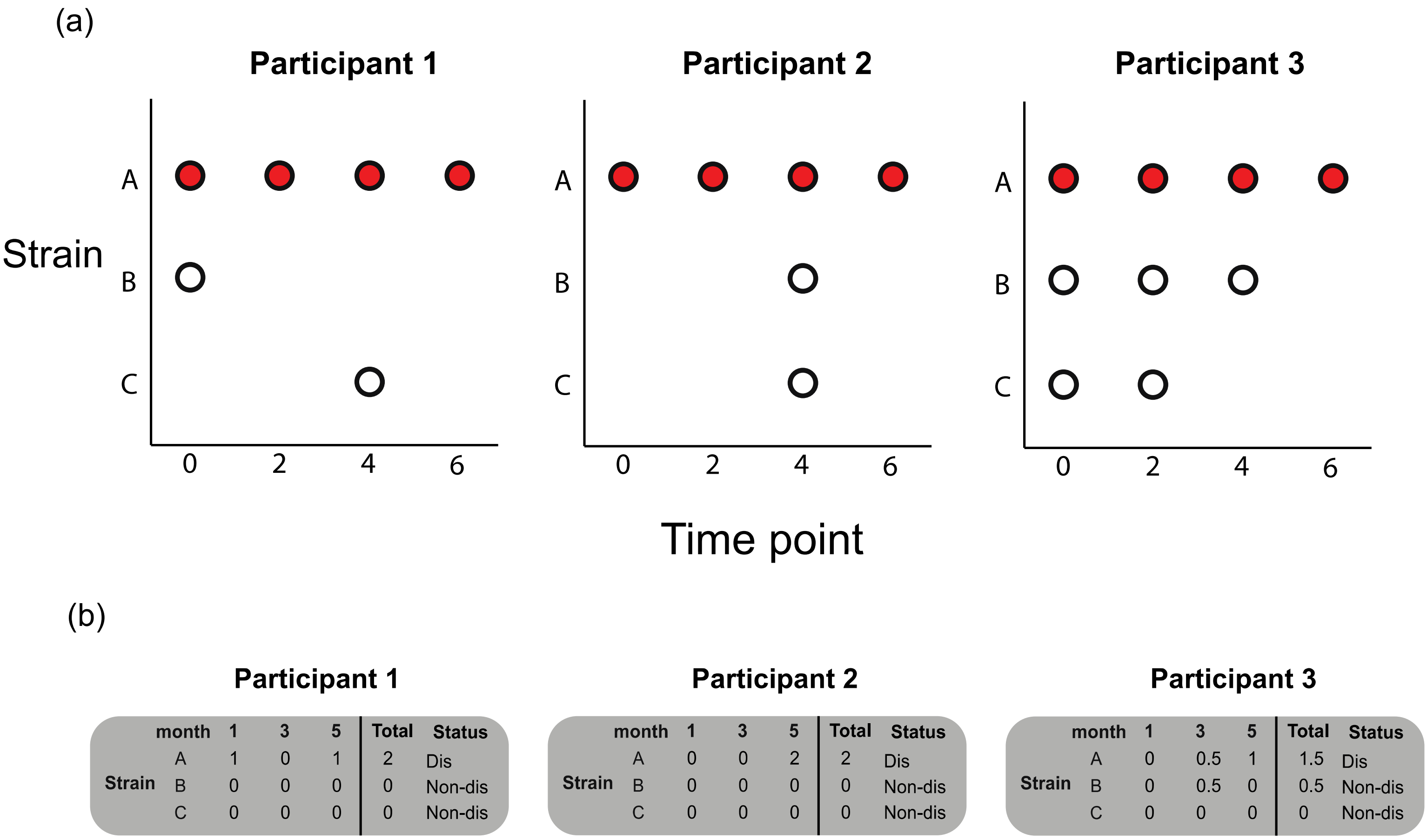
month	1	3	5	7	9	11
bacteriocin	1	0	0	1	1	1
event_ID	Clr	Gain	Sing	Inv	Co	Fair
summary_ID	Non-dis	Non-dis	Non-dis	Dis	Non-dis	Dis

Fig. SM3 – Characterisation of displacement events. Panel (a) provides examples of all displacement and non-displacement event types. Displacement events can be sub-categorized into: i) ‘invasion’ events, when an invading strain successfully displaces a resident strain, which can either occur after co-colonisation (top) or without co-colonisation (bottom); ii) ‘defence’ events, when a resident strain prevents the successful invasion of an attacking strain; iii) ‘fair fight’ events, when two (or more) strains invade the nasal cavity at the same time point, after a period of no colonisation, and one (or more) of these strains get displaced in a future time point, without the invasion of any additional strains; iv) ‘unknown’ events , in all other cases where a displacement event occurs, but we cannot define it as an attack, defence, or fair fight event. Non-displacement events can be sub-categorized into: v) ‘single’ events, when a single strain persists alone from one time point to the next; vi) ‘co-occurrence’ events, when two (or more) strains persist from one time point to the next; vii) ‘gain’ events, when one (or more) strains colonise the nasal cavity after a period of ‘no growth’ (NG). Finally, the only events that cannot be categorized as displacement or non-displacement events are categorized as: viii) ‘clearance’ events, when one or more strains are lost from the nasal cavity, leaving no persisting strains. The bottom panel outlines how we created a hypothetical time point between each pair of time point samples. We used strain dynamics data to label which displacement or non-displacement event it was associated with and whether it was associated with inhibitory activity, i.e. whether either time point in the pair contained an inhibitory isolate (=1) or not (=0). We subsequently used this data to test whether inhibitory isolates were significantly associated with displacement events.

2.2.2. Displacement analysis – strain-level

We next tested whether inhibitory strains were more likely to perform displacement compared to non-inhibitory strains. This differs from the previous section 2.2.1, which focused on the isolate-level and did not analyse strain identity.

We calculated a 'displacement score' for each strain, within each participant, which corresponds to the number of strains it displaced during the sampling time frame (Fig. SM4). We define a strain as a 'displacer' if it obtains a displacement score of ≥1 in any participant, and as a 'non-displacer' if it only ever obtains a displacement score of <1 in a participant. To test whether inhibitor strains were significantly more likely to displace other strains, we used a binomial GLM with 'inhibitory activity' (inhibitor/non-inhibitor) as an explanatory variable, and 'displacement potential' (displacer/non-displacer) as a binary response variable. We performed multiple variations of this analysis to check the robustness of results (see 'Supplementary Methods - section 3' for explanation and Tables S6 & S7 for results).



Participant 1

	month	1	3	5	Total	Status
	A	1	0	1	2	Dis
Strain	B	0	0	0	0	Non-dis
	C	0	0	0	0	Non-dis

Participant 2

	month	1	3	5	Total	Status
	A	0	0	2	2	Dis
Strain	B	0	0	0	0	Non-dis
	C	0	0	0	0	Non-dis

Participant 3

	month	1	3	5	Total	Status
	A	0	0.5	1	1.5	Dis
Strain	B	0	0.5	0	0.5	Non-dis
	C	0	0	0	0	Non-dis

Fig SM4. Calculating displacement scores at the strain-level. Definitions of axes, dot presence/absence, and dot colour are the same as in Fig. SM1. Panel (a) provides examples of different displacement events associated with three strains in three different participants. Panel (b) provides the corresponding displacement scores for each strain in each participant. In participant 1 and participant 2, strain A displaces two strains, obtaining a total displacement score = 2 in each participant. Note that displacement events count regardless of whether they occur at different time points (e.g., in participant 1), or the same time point (e.g., in participant 2). In participant 3, strains A and B are both associated with the displacement of strain C between time points 2 and 4. In cases where more than one strain is associated with winning a given displacement event, then the winning strains are awarded a displacement score = 'number of strains displaced/number of strains that win the displacement'. For example, in participant 3, strain A and B are both awarded 0.5 for displacing strain C between time points 2 and 4. In participant 3, strain A subsequently displaces strain B, giving it a total displacement score = 1.5. A strain is labelled as a 'displacer' within a participant if it obtains a displacement score ≥1 in that participant, and as a 'displacer' within the whole-collection if it is labelled as a displacer in one or more participants. A strain is labelled as a 'non-displacer' within a participant if it obtains a displacement score <1 in that participant, and a 'non-displacer' within the whole-collection if it is not labelled as a displacer in any participant.

2.3. Variation in inhibitory activity

2.3.1. Within-strain, within-participant, over time

To determine the level of variation in inhibitory activity within strains over time, we identified whether inhibitory activity switched between being present or absent (1/0) in consecutive time point isolates (Fig. SM5a). We calculated the proportion of 'strain x participant' interactions displaying variation in inhibitory activity over time, and the standard error of this proportion, for all strains present in two or more time point isolates. We did this for all types of inhibition (e.g., inhibition of *C. fimi,* interspecific competitors, and intraspecific competitors).

2.3.2. Within-strain, between-participants

We labelled strains that displayed between-participant variation in inhibitory activity (Fig. SM5b). We calculated the proportion of stains displaying between-participant variation in inhibitory activity, and the standard error of this proportion. We did this for all types of inhibition (e.g., inhibition of *C. fimi,* interspecific competitors, and intraspecific competitors).

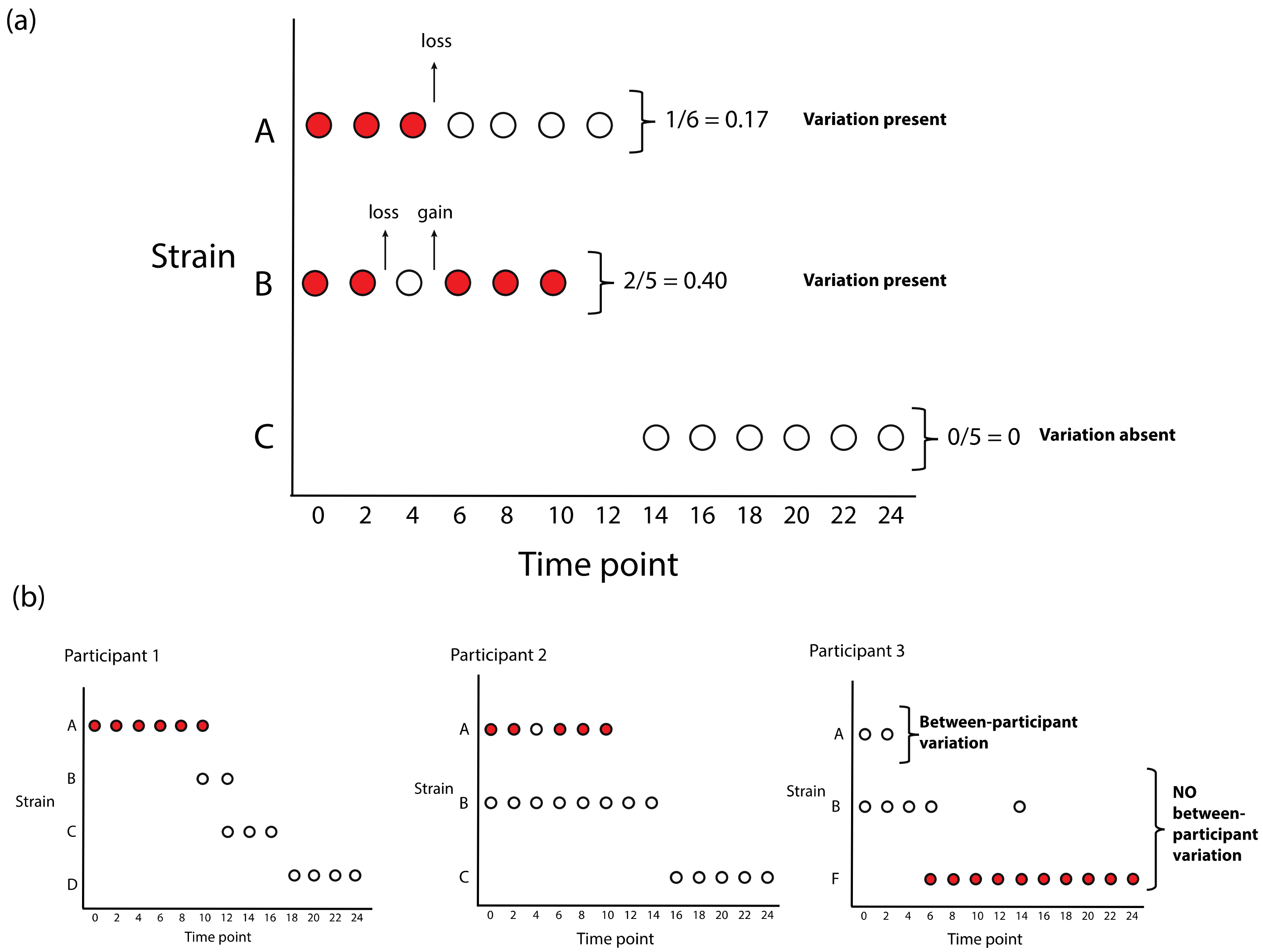


Fig SM5. Characterising variation in inhibitory activity. Definitions of axes, dot presence/absence, and dot colour are the same as in the above figures. Panel (a) explains how we calculate a measure for within-strain, within-participant variation in inhibition. This involves counting the number of 'switches' in inhibition/non-inhibition between consecutive time point isolates and dividing this by the total number of consecutive timepoint pairs to calculate the 'proportion of variation' for each strain. When a strain is present in more than one participant, we calculate its 'mean proportion of variation' across participants. Panel (b) explains how we define whether a strain displays between-participant variation in inhibition.

3. Variations of each analysis

Given the nature of our dataset, there were many ways we could subset our data to perform variations of each analysis, allowing us to check the robustness of results. Here, we outline the variables we vary for each analysis, reasoning for each, and which variable state we use in our main analyses. The full list of main analyses and additional forms of analysis for each results section can be found in Tables S4-S8.

Firstly, we varied the ‘number of time points’, or time frame used in different analyses. This is important because the resolution of time points within participants in the dataset decreases after 24-months (‘Chapter 2 – methods – original *Staphylococcus aureus* nasal carriage collection – sampling procedure’). Our main analyses therefore all use strain dynamics data capped at 24-months, but we also perform additional analyses using all strain dynamics data up to 88-months.

Secondly, due to the possibility of false-negatives from nasal swabs, previous studies working on the strain dynamics in this isolate collection have defined ‘strain loss’ as the absence of a strain in two consecutive time points, instead of in a single time point (Miller *et al.,* 2014; Votintseva *et al.,* 2014). Our main analyses therefore also include this ‘strain loss’ data modification, where we convert ‘absence’ to ‘presence’ for a strain at a given time point if it was present at each time point either side of this absence (Fig. SM6). This control is particularly important in displacement analyses as false-negatives could inflate the successful displacement count. We also perform additional analyses without the ‘strain loss’ data modification.

Thirdly, we varied certain analyses to either focus on one participant category (co-colonised) or both participant categories (co-colonised and single-strain participant). Displacement analyses only contain co-colonised participants as single-strain participants do not contain displacement events (Fig. 3b). Our main colonisation time analysis (Fig. 3a) only contains co-colonised participants, but we perform variations of this analysis with single-strain participant data added. Our main 'variation in inhibitory activity' analyses (Fig. S2) contain both participant types.

Fourthly, we varied analyses to either contain: i) each strain, within each participant, as an independent data point; ii) each unique strain in the collection as an independent data point, by collapsing data from strains present in multiple participants into a single data point, as described in section 1.1. To avoid the pseudoreplication of strains, our main analyses focused at the unique strain level. But we also performed additional analyses with data at the 'strain x participant' level, as this allowed us to control for 'Participant ID' as a random effect in generalized linear mixed-effect models (GLMMs) and Markov chain Monte Carlo GLMMs (MCMCglmms) (see paragraph below).

Fifthly, we performed additional analyses to control for the phylogenetic relationships between strains (Harvey & Pagel, 1992). We did this using a Markov chain Monte Carlo GLMM (MCMCglmm) in R, with phylogeny as a random effect (Hadfield, 2010) (see 'Methods – controlling for phylogenetic non-independence'). MCMCglmms have been described in detail previously (Hadfield, 2010; Hadfield, 2019). We did not control for phylogeny in our main analyses as doing so did not increase model fit in MCMCglmms, based on DIC comparisons.

Sixthly, we varied how inhibitory activity was defined in different analyses: i) inhibition of positive indicator *C. fimi*; ii) inhibition of intraspecific or interspecific competitors. We use the inhibition of *C. fimi* as the response variable in our main analyses because it is likely to best represent the total number of strains capable of ecologically-relevant inhibition, given we only screened for interspecific inhibition against three other species from the nasal cavity.

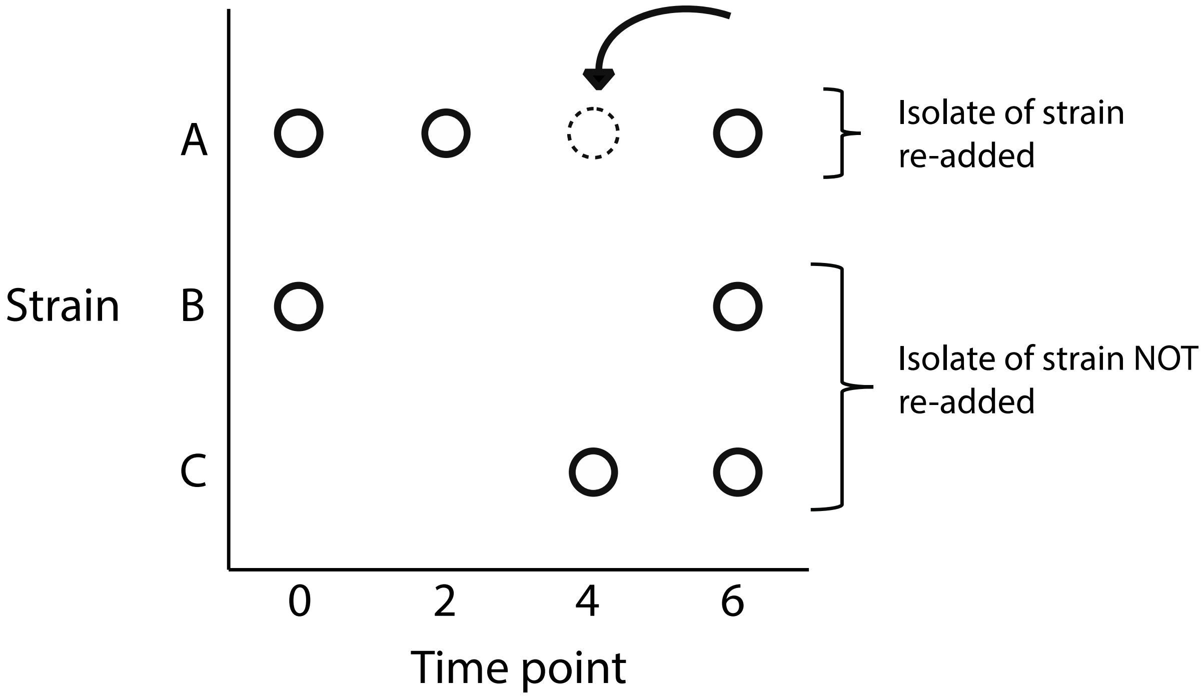


Fig SM6. 'Strain loss' criteria to account for false negative samples. Due to the possibility of false negatives for the presence of a given strain at a sample time point, following previous studies on the collection (Miller *et al.*, 2014), we only define a strain to be 'lost' from the nasal cavity if it is absent in two consecutive time point samples (e.g., strain B). If a strain was only absent in one time point sample, but present in both adjacent time points (e.g., strain A), then the strain is re-added to this missing sample time point (dashed circle). The inhibition profile for all re-added isolates was labelled as 'NA'. We performed analyses with and without this data modification to check the robustness of results.

Supplementary Methods – laboratory screens

Zone of inhibition measurement – isolate selection

To measure the zone of inhibition of isolates, we randomly selected isolate from each ‘strain x participant’ interaction displaying inhibitory activity against the positive indicator for bacteriocin production (*C. fimi*) (isolates = 22, participants = 15, strains = 17) (Table. S1). Selected isolates were screened against *C. fimi*, *M. catarrhalis*, *C. pseudodiphtheriticum*, *S. epidermidis* (Table. S1), and against Newman *ΔdltA* (*S. aureus* positive indicator for intraspecific inhibition) and three *S. aureus* strains from the nasal cavity (Table. S2). Two of the nasal strain isolates were randomly selected (isolate IDs = 132-22 t228; 499-24 t021). The other strain isolate (isolate ID = 637-12 t870) was selected as it co-existed in participant 637 with strain t7050, which displayed intraspecific inhibition against the *S. aureus* positive indicator, but not against natural *S. aureus* strains, and we therefore wanted to further test its ability to perform intraspecific inhibition (Table. S2). *S. aureus* isolates were not re-screened against *S. epidermidis* as no *S. aureus* isolates were observed to inhibit *S. epidermidis* in the initial round of screening.

Stress-induction experiment – isolate selection

To test whether nasally-relevant environmental stress affected inhibitory activity, we screened the same 22 inhibitory isolates described above against *C. fimi*, *S. aureus* positive indicator for intraspecific inhibition, and the three selected *S. aureus* nasal strains described above, under conditions nasally-relevant stress conditions, including oxidative stress and iron-limitation (Table. S2).

In addition to testing isolates identified to be inhibitory against *C. fimi* in previous screens, we also tested whether non-inhibitory strains became inhibitory under stress induced conditions (Table. S2). We selected 24 non-inhibitory isolates (participants = 19, strains = 20) based on the following criteria: i) 15 randomly selected non-inhibitory strain isolates (n = 15); ii) all isolates that, following genomic analysis, were determined to carry a bacteriocin gene cluster (BGC), but did not display inhibitory activity (n = 6); iii) all remaining isolates that did not display inhibitory activity, despite other isolates of the same strain displaying inhibitory activity in the same participant (n = 3) (total n = 24) (Table S2).

Intraspecific inhibition screens – lawn isolate selection criteria

To perform the screens testing the level of intraspecific inhibition between natural *S. aureus* isolates, we screened the 22 inhibitory isolates and 24 non-inhibitory isolates described above against 20 strains of *S. aureus* from the nasal cavity (Table. S3). As we particularly wanted to test the inhibitory activity of strains identified to inhibit the *S. aureus* positive indicator for intraspecific inhibition (strain identities = participant 637/strain t7050 & participant 2064/strain t171) (Table. S2), we selected lawn isolates using the following criteria: : i) a randomly selected isolate from the two most closely related to each of the two inhibitory strains (n=4); ii) a randomly selected isolate from all strains that co-colonise participants 637 and 2064 with the two inhibitory strains (n=9); iii) a randomly selected strain isolate from each of the remaining *spa*-Clonal Complexes (*spa*-CCs) that were not covered by the above criteria (n=5); iv) two additional randomly selected strain isolates to increase sample size (n=2). In total, 20 lawns were selected for intraspecific screening (Table. S3).

Supplementary Methods - genomic analyses

Genome selection & quality checks

As a control, we duplicated the sequencing of one isolate, to ensure we got the same result across two separately sequenced genomes. Following genome assembly using, we checked assembly quality and identified any sample contamination. One genome could not be assembled to an appropriate standard and was removed from the genomic analysis (contigs = >2000; N50 = 968). A further genome was identified to be a different species (*Staphylococcus argenteus*) by performing rMLST analysis using the ‘species ID’ tool on PubMLST (Jolley *et al.,* 2012, 2018), and was removed from the whole isolate collection. Four additional genomes provided a conflicting *spa*-type identity compared to previous *spa*-typing, and were therefore removed from the collection and any subsequent analyses, as we could not be certain of the identity of strains in each isolate. Therefore, a final total of 89 genomes were used for further genomic analyses (Table. S9). However, we also repeated all analyses including genomes and isolates, and always obtained the same result.

Genomic analysis – mining genomes bacteriocin gene clusters (BGCs)

Following whole-genome sequencing, we genomically screened 89 *S. aureus* genomes from our collection for bacteriocin gene clusters (BGCs) using antiSMASH (6.0) (Blin *et al.,* 2021) and BAGEL4 (van Heel *et al.*, 2018) (Table S9). antiSMASH and BAGEL4 identify BGCs by comparing gene products in the input sequence to large databases of previously characterised genes to detect conserved molecular signatures and sequence similarity associated with all secondary metabolite (antiSMASH) or bacteriocin genes (BAGEL). Genes for the production, processing, and transport of a particular bacteriocin are usually encoded in close proximity within the genome (Heilbronner *et al.*, 2021), which allows for effective detection of BGCs.

Once BGCs are identified, both software provide the BGC class, a prediction of the core bacteriocin, and a functional annotation of genes in the BGC (van Heel *et al.,* 2018; Blin *et al.,* 2021).

We used both antiSMASH and BAGEL4 because each software has different strengths and weaknesses, meaning a combination of the two provided more accurate identification and BGC predictions than using either alone. All BGC type predictions (Fig. 4; Table. S9) are from antiSMASH. Bacteriocin identity predictions (Fig. 4; Table S9) are from a combination of antiSMASH and BAGEL4 outputs, depending on which software gave the highest quality prediction. In addition, to ensure that each prediction was plausible, we consulted the relevant literature and compared the structure of the BGC in our input sequence to the structure of the previously characterised BGC. All schematics of the genetic organisation of BGCs were taken from antiSMASH (Fig. 4). Once antiSMASH and BAGEL outputs were obtained, we mapped BGC hits to the phenotypic inhibition profile of each isolate. We also used antiSMASH output to check for the presence of any other secondary metabolite biosynthetic gene cluster types, in addition to bacteriocins, that have known antimicrobial effects.

To determine the stability of BGCs within strains over time, we downloaded all genomic data from antiSMASH for each BGC, to determine whether: i) whole BGCs were gained or lost within strains over time; ii) bacteriocin-related genes within BGCs (see Fig. 4) were gained or lost within strains over time (Table. S9). In cases where BGCs were identified in the genome of an isolate displaying no inhibitory activity, we used EMBOSS Transeq (Rice *et al.,* 2000; Goujon *et al.*, 2010) to identify intragenic stop codons in bacteriocin-related genes, which could cause loss-of-function.

Chapter 3 – Supplementary Results

No.	Isolate ID	*C. fimi*	*C. fimi* + oxidative stress	*C. fimi* + iron limitation	*M. catarrhalis*	*C. pseudodiphtheriticum*	*S. epidermidis*
				C. fimi inhibitors			
1	162-48-t209	🟨	🟨	🟨	🟦	🟦	🟦
2	359-6-t089	🟠	🟧	🟠	🟠	🟨	🟦
3	450-4-t382	🟨	🟨	🟨	🟨	🟦	🟦
4	637-12-t7050	🟨	🟨	🟨	🟦	🟦	🟦
5	638-24-t171	🟨	🟨	🟨	🟨	🟦	🟦
6	647-8-t6855	🟨	🟨	🟨	🟦	🟦	🟦
7	926-24-t008	🟨	🟨	🟨	🟦	🟦	🟦
8	967-18-t171	🟨	🟨	🟠	🟨	🟦	🟦
9	971-8-t008	🟨	🟨	🟨	🟨	🟦	🟦
10	997-2-t015	🟨	🟨	🟨	🟦	🟦	🟦
11	1231-46-t120	🟨	🟨	🟨	🟦	🟦	🟦
12	1231-46-t4309	🟨	🟨	🟨	🟦	🟦	🟦
13	1231-68-t15780	🟨	🟨	🟨	🟦	🟦	🟦
14	1231-80-t499	🟨	🟨	🟨	🟦	🟦	🟦
15	1366-2-t089	🟠	🟧	🟠	🟠	🟨	🟦
16	1366-14-t084	🟨	🟨	🟨	🟨	🟦	🟦
17	2004-4-t1685	🟨	🟨	🟨	🟨	🟦	🟦
18	2009-32-t230	🟨	🟨	🟨	🟦	🟦	🟦
19	2064-0-t6825	🟨	🟨	🟨	🟨	🟦	🟦
20	2064-2-t171	🟥	🟫	🟫	🟠	🟠	🟦
21	2064-20-t7031	🟨	🟨	🟨	🟦	🟦	🟦
22	2064-20-t008	🟨	🟨	🟨	🟦	🟦	🟦
				Non-inhibitors			
23	022-24-t012	🟦	🟦	🟦	🟦	🟦	🟦
24	132-6-t228	🟦	🟦	🟦	🟦	🟦	🟦
25	420-76-t1414	🟦	🟦	🟦	🟦	🟦	🟦
26	499-24-t021	🟦	🟦	🟦	🟦	🟦	🟦
27	637-16-t870	🟦	🟦	🟦	🟦	🟦	🟦
28	647-4-t019	🟦	🟦	🟦	🟦	🟦	🟦
29	926-12-t190	🟦	🟦	🟦	🟦	🟦	🟦
30	971-10-t230	🟦	🟦	🟦	🟦	🟦	🟦
31	1212-24-t056	🟦	🟦	🟦	🟦	🟦	🟦
32	1307-24-t408	🟦	🟦	🟦	🟦	🟦	🟦
33	2004-0-t160	🟦	🟦	🟦	🟦	🟦	🟦
34	2030-24-t002	🟦	🟦	🟦	🟦	🟦	🟦
35	2060-6-t065	🟦	🟦	🟦	🟦	🟦	🟦
36	2060-24-t084	🟦	🟦	🟦	🟦	🟦	🟦
37	2064-4-t6814	🟦	🟦	🟦	🟦	🟦	🟦
38	454-24-t127	🟦	🟦	🟦	🟦	🟦	🟦
39	972-6-t127	🟦	🟦	🟦	🟦	🟦	🟦
40	926-8-t008	🟦	🟦	🟦	🟦	🟦	🟦
41	926-14-t127	🟦	🟦	🟦	🟦	🟦	🟦
42	2004-4-t321	🟦	🟦	🟦	🟦	🟦	🟦
43	2004-6-t321	🟦	🟦	🟦	🟦	🟦	🟦
44	162-24-t209	🟦	🟦	🟦	🟦	🟦	🟦
45	1231-68-t120	🟦	🟦	🟦	🟦	🟦	🟦
46	1366-12-t084	🟦	🟦	🟦	🟦	🟦	🟦

Zone size key:

🟦	no inhibition
🟨	≤ 1
🟠	> 1 to ≤ 2 mm
🟧	> 2 to ≤ 3 mm
🟥	> 4 to ≤ 5 mm
🟫	> 5 to ≤ 6 mm

Table S1. Measuring inhibition zone size and the effect of stress induction. The y-axis represents isolates that were tested for inhibitory activity. Each 'isolate ID' is constituted of: participant ID, sample time point, and *spa*-type. The x-axis represents isolates inoculated into lawns to be tested

against, in addition to whether screens were performed under standard, oxidative stress, or iron-limited conditions. The inhibition zone size for each isolate is represented as a heat map, as defined in the ‘zone size key’ (top right). Inhibition zone size was measured from the colony edge to the inhibition zone border. In total, we screened 22 isolates determined to inhibit *C. fimi* (a positive indicator for bacteriocin production) in previous screens (participants = 15, strains = 17) and 24 non-inhibitory isolates (bottom) (participants = 19, strains = 20) (see Supplementary Methods – laboratory screens’ for selection criteria). Inhibitory isolates (top) are ordered according to participant number, followed by sampling time point. Non-inhibitory isolates (bottom) are ordered according to their selection criteria (see Supplementary Methods – laboratory screens’): i) isolates 23-37 represent 15 randomly selected strains; ii) isolates 38-43 represent 6 isolates that, following genomic analysis, were identified to carry a bacteriocin gene cluster (BGC) but did not display inhibitory activity; iii) isolates 44-46 represent the remaining 3 strain isolates that did not display inhibitory activity, despite displaying inhibitory activity at other time points within the same participant. Test isolates were screened against *C. fimi* under standard, oxidative stress, and iron-limited conditions, and against *M. catarrhalis* and *C. pseudodiphtheriticum*, and *S. epidermidis* under standard conditions.

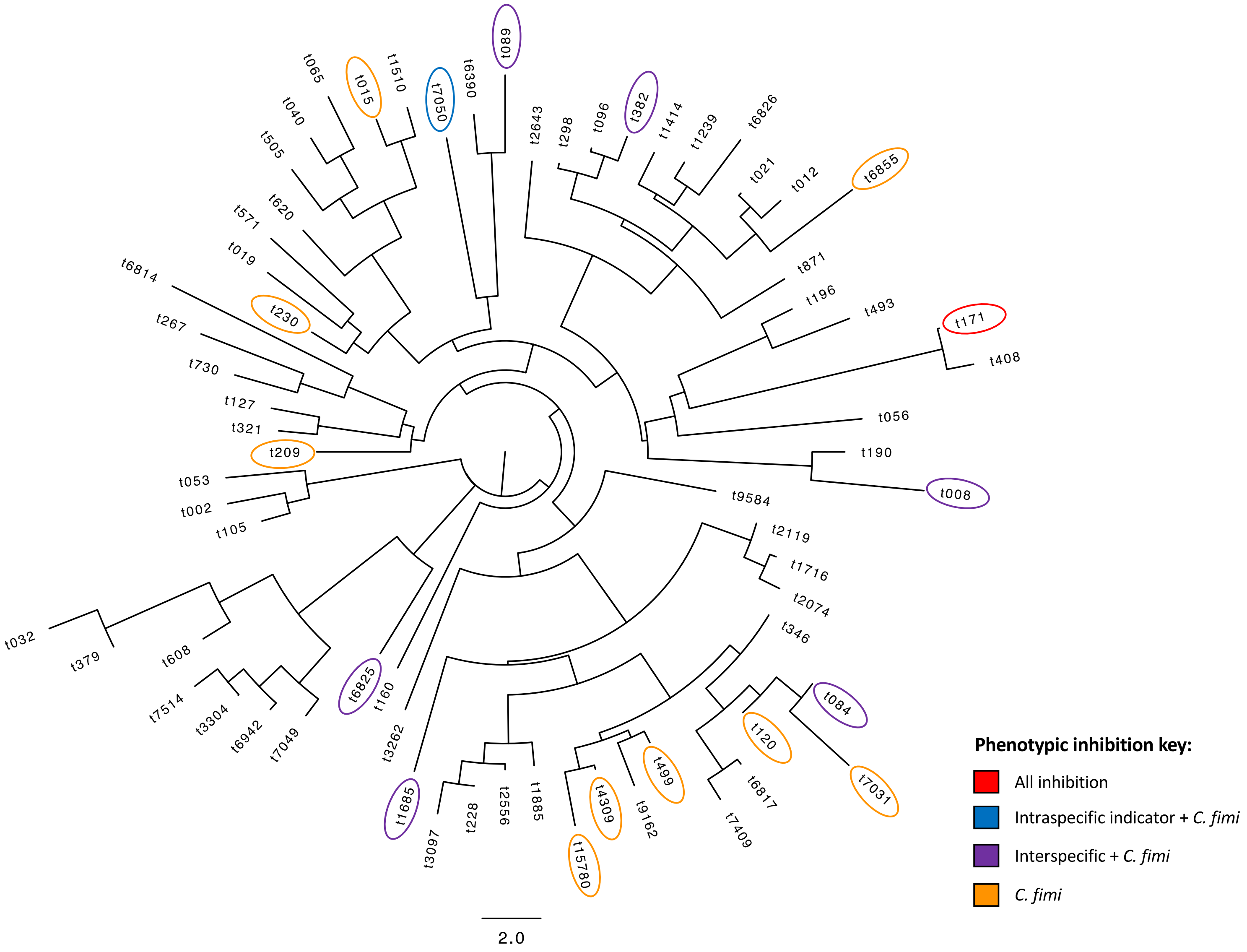


Fig S1. A *spa* gene neighbour joining tree representing an overview of phenotypic inhibition profiles. The *spa* neighbour joining tree is based on *spa* gene sequence of all strains in the collection and produced using the BURP (based-upon repeat pattern) clustering algorithm (Mellmann *et al*., 2007), integrated within SeqSphere+ (8.4.0) (Ridom, GmbH). Each *spa*-type represents a unique strain in the collection. Two *spa*-types (t528 and t870) were removed from the tree, as they did not contain the required number of repeats (≥ 5) for BURP clustering (Mellmann *et al*., 2007) (total strains = 62). All *spa*-types that display inhibitory activity are circled with a ring. Ring colour represents different phenotypic inhibition profiles, outlined in the 'phenotypic inhibition key' (bottom right). The tree was midpoint rooted using FigTree (v1.4.4).

Zone size key:

🟦 no inhibition
🟨 ≤ 1
🟧 > 1 to ≤ 2 mm

No.	Isolate ID	*S. aureus* indicator	*S. aureus* indicator + oxidative stress	*S. aureus* indicator + iron limitation	*S. aureus* nasal	*S. aureus* nasal + oxidative stress	*S. aureus* nasal + iron limitation
				C. fimi inhibitors			
1	162-48-t209	🟦	🟦	🟦	🟦	🟦	🟦
2	359-6-t089	🟦	🟦	🟦	🟦	🟦	🟦
3	450-4-t382	🟦	🟦	🟦	🟦	🟦	🟦
4	637-12-t7050	🟨	🟨	🟨	🟦	🟦	🟦
5	638-24-t171	🟦	🟦	🟦	🟦	🟦	🟦
6	647-8-t6855	🟦	🟦	🟦	🟦	🟦	🟦
7	926-24-t008	🟦	🟦	🟦	🟦	🟦	🟦
8	967-18-t171	🟦	🟦	🟦	🟦	🟦	🟦
9	971-8-t008	🟦	🟦	🟦	🟦	🟦	🟦
10	997-2-t015	🟦	🟦	🟦	🟦	🟦	🟦
11	1231-46-t120	🟦	🟦	🟦	🟦	🟦	🟦
12	1231-46-t4309	🟦	🟦	🟦	🟦	🟦	🟦
13	1231-68-t15780	🟦	🟦	🟦	🟦	🟦	🟦
14	1231-80-t499	🟦	🟦	🟦	🟦	🟦	🟦
15	1366-2-t089	🟦	🟦	🟦	🟦	🟦	🟦
16	1366-14-t084	🟦	🟦	🟦	🟦	🟦	🟦
17	2004-4-t1685	🟦	🟦	🟦	🟦	🟦	🟦
18	2009-32-t230	🟦	🟦	🟦	🟦	🟦	🟦
19	2064-0-t6825	🟦	🟦	🟦	🟦	🟦	🟦
20	2064-2-t171	🟧	🟧	🟧	🟧	🟧	🟧
21	2064-20-t7031	🟦	🟦	🟦	🟦	🟦	🟦
22	2064-20-t008	🟦	🟦	🟦	🟦	🟦	🟦
				Non-inhibitors			
23	022-24-t012	🟦	🟦	🟦	🟦	🟦	🟦
24	132-6-t228	🟦	🟦	🟦	🟦	🟦	🟦
25	420-76-t1414	🟦	🟦	🟦	🟦	🟦	🟦
26	499-24-t021	🟦	🟦	🟦	🟦	🟦	🟦
27	637-16-t870	🟦	🟦	🟦	🟦	🟦	🟦
28	647-4-t019	🟦	🟦	🟦	🟦	🟦	🟦
29	926-12-t190	🟦	🟦	🟦	🟦	🟦	🟦
30	971-10-t230	🟦	🟦	🟦	🟦	🟦	🟦
31	1212-24-t056	🟦	🟦	🟦	🟦	🟦	🟦
32	1307-24-t408	🟦	🟦	🟦	🟦	🟦	🟦
33	2004-0-t160	🟦	🟦	🟦	🟦	🟦	🟦
34	2030-24-t002	🟦	🟦	🟦	🟦	🟦	🟦
35	2060-6-t065	🟦	🟦	🟦	🟦	🟦	🟦
36	2060-24-t084	🟦	🟦	🟦	🟦	🟦	🟦
37	2064-4-t6814	🟦	🟦	🟦	🟦	🟦	🟦
38	454-24-t127	🟦	🟦	🟦	🟦	🟦	🟦
39	972-6-t127	🟦	🟦	🟦	🟦	🟦	🟦
40	926-8-t008	🟦	🟦	🟦	🟦	🟦	🟦
41	926-14-t127	🟦	🟦	🟦	🟦	🟦	🟦
42	2004-4-t321	🟦	🟦	🟦	🟦	🟦	🟦
43	2004-6-t321	🟦	🟦	🟦	🟦	🟦	🟦
44	162-24-t209	🟦	🟦	🟦	🟦	🟦	🟦
45	1231-68-t120	🟦	🟦	🟦	🟦	🟦	🟦
46	1366-12-t084	🟦	🟦	🟦	🟦	🟦	🟦

Table S2. Measuring intraspecific inhibition zone size and the effect of stress induction. The y- and x-axes are the same as in Fig S1, apart from test isolates now being screened against a positive indicator for *S. aureus* intraspecific inhibition (Newman *ΔdltA*) and three selected *S. aureus*

strains the nasal cavity (132-22 -t228; 499-24 - t021; 637-12 – t870), under standard, oxidative stress, and iron-limited conditions. See ‘Supplementary Methods – laboratory screens’ for *S. aureus* nasal strain selection criteria. Inhibition results against ‘*S. aureus* nasal’, ‘*S. aureus* nasal + oxidative stress’, and ‘*S. aureus* nasal + iron limitation’ were consistent across the three nasal strain lawns for all isolates tested, and therefore data was collapsed into a single column for each screen type. Inhibition zone definitions are the same as in Fig S1, and are stated in the ‘zone size key’ (top right).

Inhibition key:

🟦	no inhibition
🟧	inhibition

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
No.	Isolate ID	2064-2-t171	637-12-t7050	926-2-t196	1307-24-t408	2064-0-t1239	2064-0-t6825	2064-0-t6826	2064-6-t6814	2064-20-t008	2064-20-t7031	359-46-t089	1367-20-t6390	637-8-t870	424-10-t7049	454-24-t127	671-24-t032	688-14-t053	1045-26-t1510	971-16-t230	1366-2-t267
***S. aureus* indicator inhibitors**																					
1	2064-2-t171	🟦	🟧	🟧	🟦	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
2	637-12-t7050	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
***C. fimi* inhibitors**																					
3	162-48-t209	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
4	359-6-t089	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
5	450-4-t382	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
6	638-24-t171	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
7	647-8-t6855	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
8	926-24-t008	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
9	967-18-t171	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
10	971-8-t008	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
11	997-2-t015	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
12	1231-46-t120	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
13	1231-46-t4309	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
14	1231-68-t15780	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
15	1231-80-t499	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
16	1366-2-t089	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
17	1366-14-t084	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
18	2004-4-t1685	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
19	2009-32-t230	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
20	2064-0-t6825	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
21	2064-20-t7031	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
22	2064-20-t008	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Non-inhibitors																					
23	022-24-t012	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
24	132-6-t228	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
25	420-76-t1414	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
26	499-24-t021	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
27	637-16-t870	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
28	647-4-t019	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
29	926-12-t190	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
30	971-10-t230	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
31	1212-24-t056	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
32	1307-24-t408	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
33	2004-0-t160	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
34	2030-24-t002	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
35	2060-6-t065	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
36	2060-24-t084	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
37	2064-4-t6814	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
38	454-24-t127	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
39	972-6-t127	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
40	926-8-t008	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
41	926-14-t127	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
42	2004-4-t321	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
43	2004-6-t321	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
44	162-24-t209	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
45	1231-68-t120	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
46	1366-12-t084	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦

Table S3. Testing the extent of intraspecific inhibition between naturally-occurring isolates. The y-axis is the same as in Fig. S1 and S2, apart from strains t171 (participant 2064) and strain t7050 (participant 637) (i.e., the two strains displaying inhibitory activity against the *S. aureus*

positive indicator for intraspecific inhibition) being moved to the top of the grid for clearer visualisation. The x-axis represents 20 *S. aureus* nasal isolates that were inoculated into lawns to be tested against. The x-axis is ordered according to the selection criteria for lawn isolates (see Supplementary Methods – laboratory screens): i) isolates 1-2 represent the two inhibitory strains; ii) isolates 3-4 represent the two strains most closely related to strain t171 (Fig. S1); iii) isolates 5-10 represent the other strains present within participant 2064 with strain t171; iv) isolates 11-12 represent the two strains most closely related to strain t7050 (Fig. S1); v) isolate 13 represents the other strain present in participant 637 with t7050; v) isolates 14-18 represent randomly selected strains to cover all remaining five *spa*-Clonal Complexes (*spa*-CCs); vi) isolates 19-20 represent two additional randomly selected strains to increase sample size. Within each category, isolates are ordered by participant number. Cell colour represent whether inhibitory activity was observed (orange = inhibition; blue = no inhibition). See 'Supplementary Methods – laboratory screens' for more details about strain selection criteria.

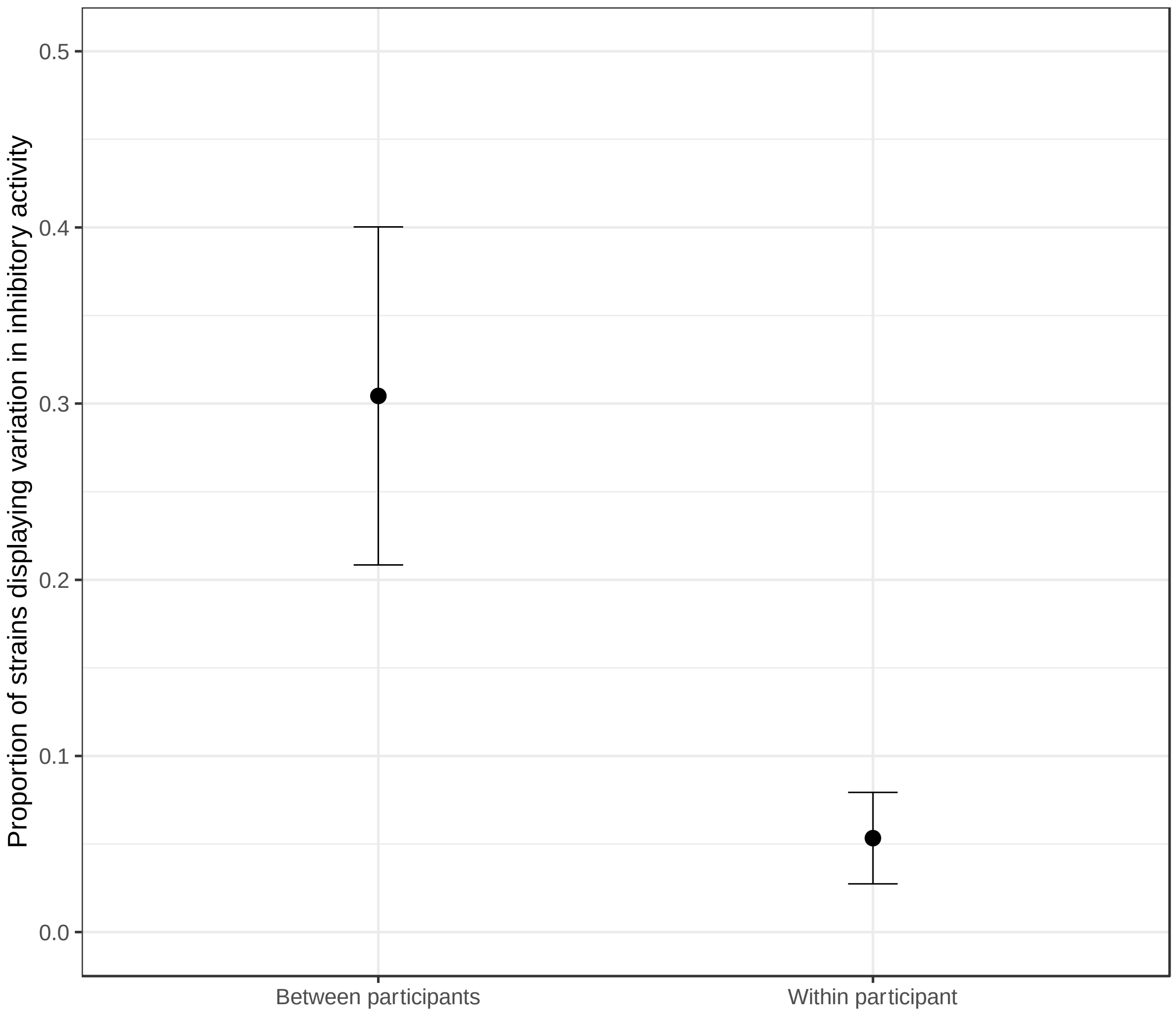


Fig S2. Inhibitory activity is generally stable within-hosts over time, but is not always consistent between-hosts. The y-axis represents the proportion of strains that display variation in inhibitory activity either between participants, or within participants over time. Error bars represent the standard error of the proportion.

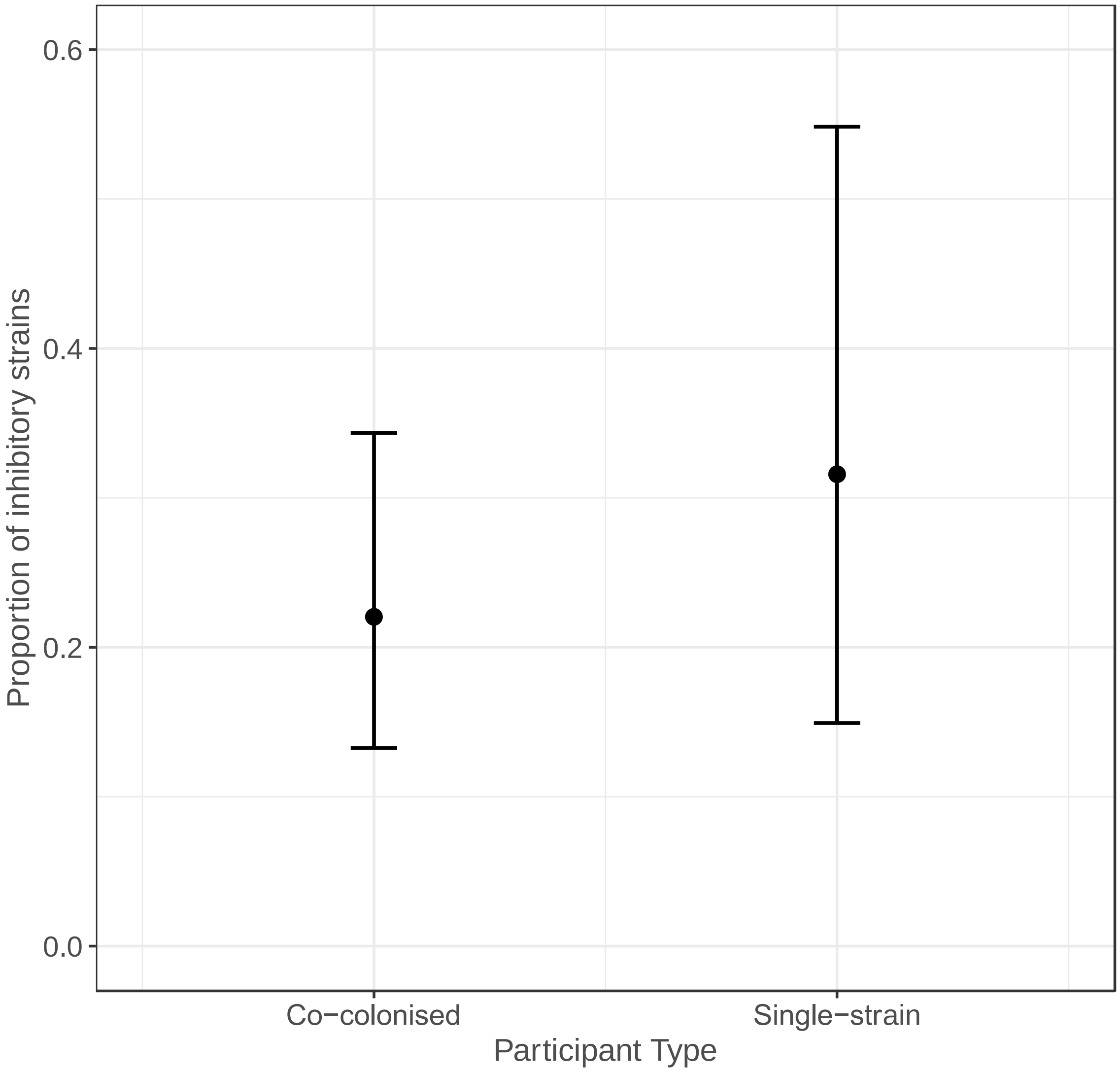


Fig. S3. *S. aureus* strains that colonise participants alone are not more likely to be inhibitory than strains that co-colonise participants with other strains. The y-axis represents the proportion of inhibitory strains in co-colonised and single-strain participant categories. Error bars represent 95% confidence intervals

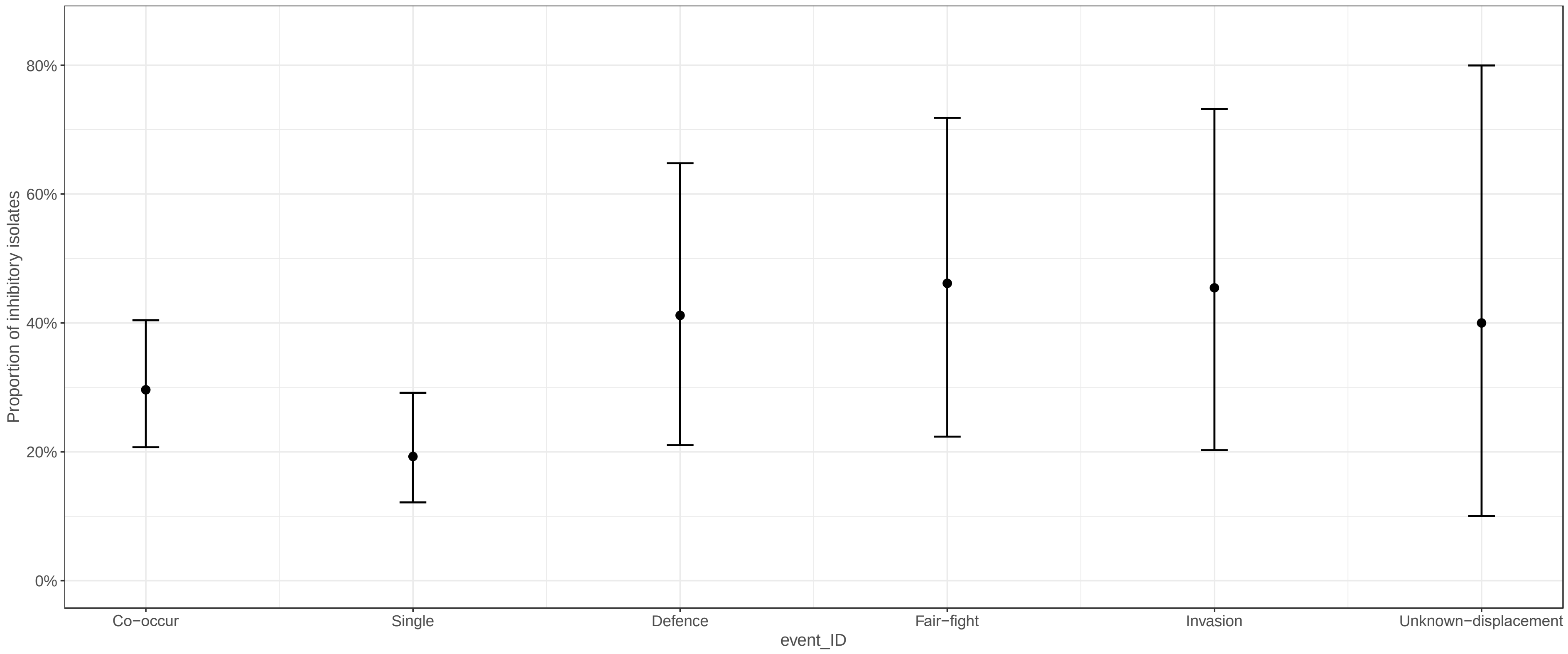


Fig S4. Inhibitory activity is not significantly associated with any individual event type. The y-axis represents the proportion of inhibitory isolates across all displacement/non-displacement event types (see 'Supplementary Methods - data analysis - section 2.2'). Non-displacement event types = 'co-occur' and 'single'; displacement event types = 'defence', 'fair-fight', 'invasion', 'unknown-displacement'. Only data points from the first 24-months of sampling in co-colonised participants were included in this plot. Error bars represent 95% confidence interval.

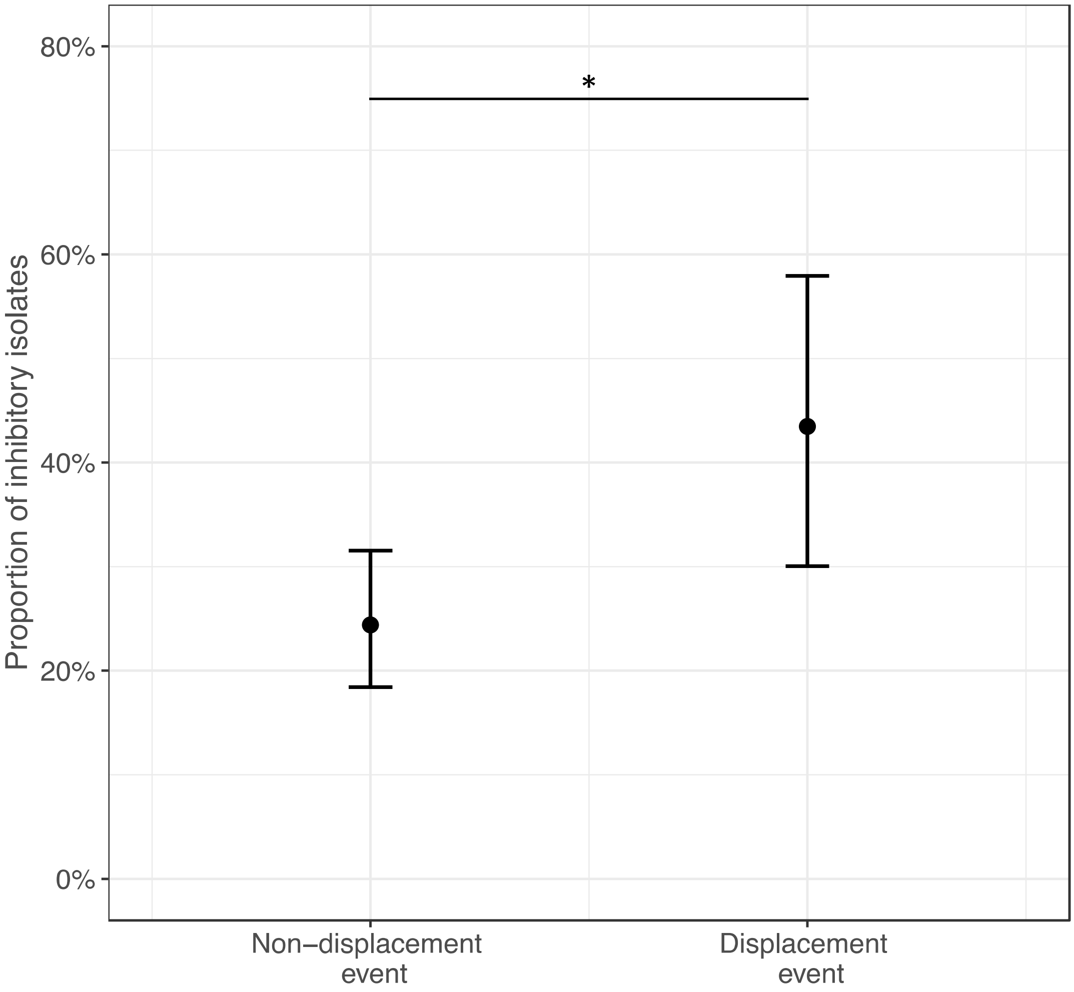


Fig. S5. Displacement events are more likely to contain inhibitory isolates than non-displacement events. The y-axis represents the proportion of inhibitory isolates. On the x-axis, displacement event types in Fig. S4 have been pooled into two categories: non-displacement and displacement events (see ‘Supplementary Methods – data analysis – section 2.2’). Only data points from the first 24-months of sampling in co-colonised participants were included in this plot. Error bars represent 95% confidence intervals. Statistical significance is denoted by an asterisk: * = $p < 0.05$.

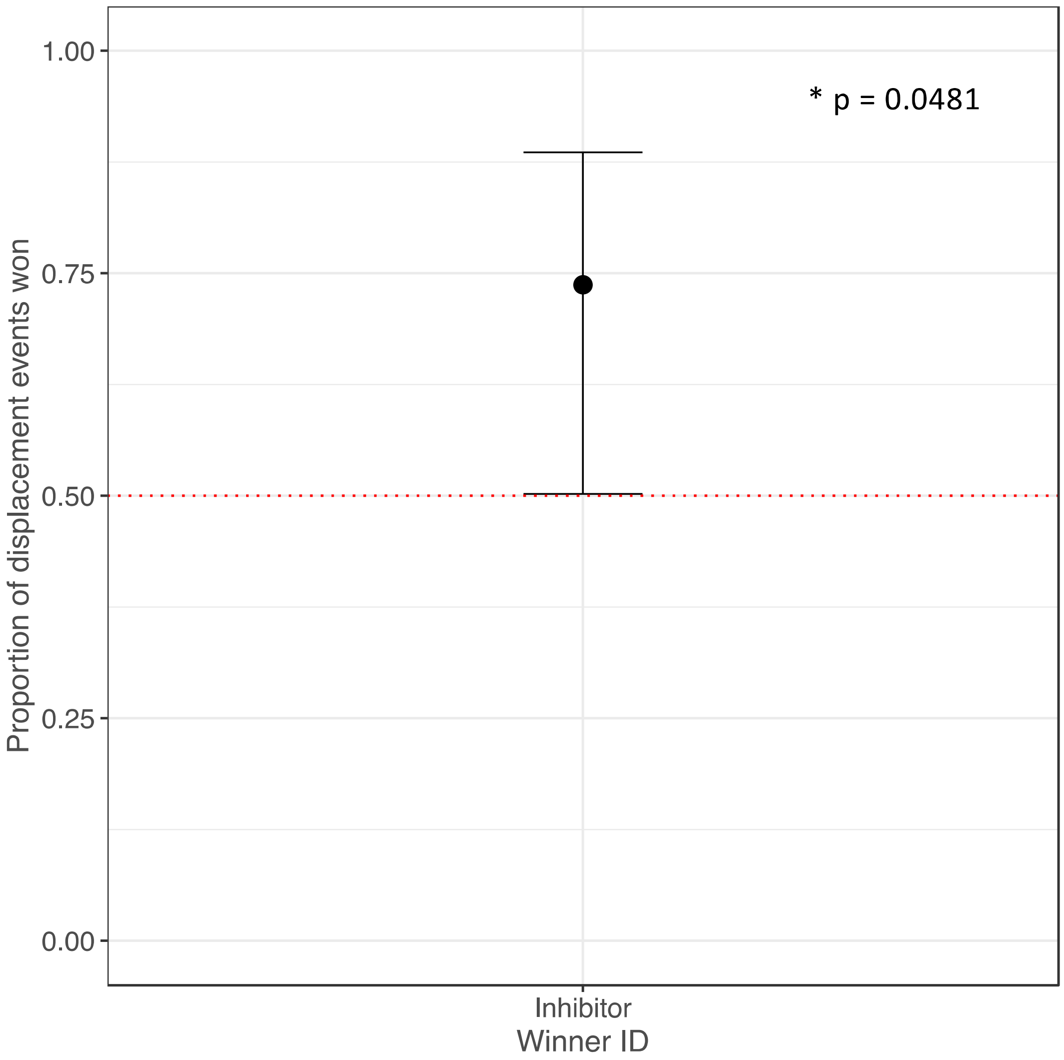


Fig. S6. Inhibitory isolates are more likely to win displacement events compared to non-inhibitory isolates. The y-axis represents the proportion of displacement events won by inhibitors (see 'Supplementary Methods - data analysis - section 2.2'). The dashed red line represents a 0.5 probability of inhibitors winning displacement events (i.e., 50:50 probability with non-inhibitors). The error bar represents the 95% confidence interval. Statistical significance is denoted by an asterisk: * = p < 0.05.

Model ID	Model description	Dataset description	Sample size	Model estimate	Standard error	P-value
1	Mean proportion of participant time points colonised ~ inhibitory status	Time frame = up to 24-months. 'Strain loss' criteria = included. Participant type(s) = co-colonised. Data points = unique strains. Inhibitor screen type = *C. fimi*.	57	0.14768	0.12866	0.256 (NS)
2	Mean proportion of participant time points colonised ~ inhibitory status	Time frame = up to 88-months	63	0.09479	0.10720	0.380 (NS)
3	Mean proportion of participant time points colonised ~ inhibitory status	'Strain loss' criteria = not included	57	0.13638	0.11872	0.256 (NS)
4	Mean proportion of participant time points colonised ~ inhibitory status	Participant type(s) = co-colonised + single-strain participants	61	0.12055	0.12972	0.357 (NS)
5	Proportion of participant time points colonised ~ inhibitory status	Data points = each 'strain x participant' interaction	69	0.14501	0.14030	0.305 (NS)
6	Mean proportion of participant time points colonised ~ inhibitory status	Inhibitor screen type = intraspecific + interspecific competitors	57	0.15589	0.15326	0.314 (NS)

Table S4 – 'Long-term success' colonisation time analyses – GLM analyses. All models in Table S4 are GLMs (family = "Gaussian"). All proportion data in normally distributed models was arcsine square root transformed to improve normality. All model estimates, standard errors, and p-values refer to the effect of inhibitors compared to non-inhibitors, on the mean proportion of participant time points colonised. The first row (Model ID = 1) describes the dataset used in the main model; the following rows (Model IDs = 2-6) state how each aspect of the dataset was individually varied compared to the dataset used in the main model, to test the robustness of the main result. Statistical significance is denoted with an asterisk(s) (NS = $p > 0.05$; * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$). See 'Supplementary Methods – data analysis – section 2.1 & 3' for more analysis details.

Model ID	Model description	Dataset description	Sample size	Posterior mean	95% Credible intervals	pMCMC
7	Mean proportion of participant time points colonised ~ inhibitory status Random effect(s) = phylogeny	Time frame = up to 24-months. 'Strain loss' criteria = included. Participant type(s) = co-colonised. Data points = unique strains. Inhibitor screen type = *C. fimi*.	57	0.1515	-0.1139 to 0.4071	0.256 (NS)
8	Mean proportion of participant time points colonised ~ inhibitory status Random effect(s) = phylogeny	Time frame = up to 88-months	63	0.1101	-0.1000 to 0.3283	0.316 (NS)
9	Mean proportion of participant time points colonised ~ inhibitory status Random effect(s) = phylogeny	'Strain loss' criteria = not included	57	0.1431	-0.1081 to 0.3749	0.25 (NS)
10	Mean proportion of participant time points colonised ~ inhibitory status Random effect(s) = phylogeny	Participant type(s) = co-colonised + single-strain participants	61	0.1153	-0.1600 to 0.3534	0.37 (NS)
11	Proportion of participant time points colonised ~ inhibitory status Random effect(s) = phylogeny	Data points = each 'strain x participant' interaction	69	0.1445	-0.1405 to 0.4105	0.318 (NS)
12	Mean proportion of participant time points colonised ~ inhibitory status Random effect(s) = phylogeny	Inhibitor screen type = intraspecific + interspecific competitors	57	0.1600	-0.1336 to 0.5025	0.308 (NS)

Table S5. 'Long-term success' colonisation time analyses – MCMCglmm analyses. All models in Table S5 are MCMCglmms (family = "Gaussian") (Hadfield, 2010, 2019). In normally distributed models we used uninformative priors for fixed and random effects (V=1, nu=0.002). All proportion data in normally distributed models was arcsine square root transformed to improve normality. All posterior means, 95% credible intervals, and pMCMCs refer to the effect of inhibitors compared to non-inhibitors, on the mean proportion of participant time points colonised. A pMCMC can generally be interpreted in the same way as a standard P value (Hadfield, 2019). The first row (Model ID = 7) describes the dataset used in the main model; the following rows (Model IDs = 8-12) state how each aspect of the dataset was individually varied compared to the dataset used in the main model, to test the robustness of the main result. Statistical significance is denoted with an asterisk(s) (NS = p > 0.05; * = p < 0.05; ** = p < 0.01; *** = p < 0.001). See 'Supplementary Methods – data analysis – section 2.1 & 3' for more analysis details.

Model ID	Model description	Dataset description	Sample size	Model estimate	Standard error	P-value
13	Displacer status ~ inhibitory status	Time frame = up to 24-months. ‘Strain loss’ criteria = included. Data points = unique strains. Inhibitor screen type = *C. fimi*.	57	2.1401	0.8624	0.0131*
14	Displacer status ~ inhibitory status	Time frame = up to 88-months.	63	1.5647	0.6731	0.0201*
15	Displacer status ~ inhibitory status	‘Strain loss’ criteria = not included.	57	1.9459	0.8583	0.0234*
16	Displacer status ~ inhibitory status	Data points = each ‘strain x participant’ interaction.	69	1.9543	0.7268	0.0071**
17	Displacer status ~ inhibitory status	Inhibitor screen type = intraspecific + interspecific competitors	57	2.3922	1.1363	0.03526*

Table S6. ‘Short-term success’ strain-level displacement analyses – Binomial GLM analyses. All models in Table 6 are binomial GLMs (family = “binomial”). ‘Model estimate’ represents the log odds of inhibitor strains being a displacing strain compared to the log odds of non-inhibitor strains being a displacing strain (i.e., the log odds ratio) and ‘standard error’ represents the standard error of the log odds ratio. All models only contain data from co-colonised participants (n=20), as single-strain participants don’t contain any displacement events. The first row (Model ID = 13) describes the dataset used in the main model; the following rows (Model IDs = 14-17) state how each aspect of the dataset was individually varied compared to the dataset used in the main model, to test the robustness of the main result. Statistical significance is denoted with an asterisk(s) (NS = $p > 0.05$; * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$). See ‘Supplementary Methods – data analysis – section 2.2 & 3’ for more analysis details.

Model ID	Model description	Dataset description	Sample size	Posterior mean	95% Credible intervals	pMCMC
18	Displacer status ~ inhibitory status Random effect = phylogeny	Time frame = up to 24-months. 'Strain loss' criteria = included. Data points = unique strains. Inhibitor screen type = *C. fimi*.	57	2.84123	0.9053 to 5.1730	0.008**
19	Displacer status ~ inhibitory status Random effect = phylogeny	Time frame = up to 88-months.	63	2.0601	0.4035 to 3.6524	0.014*
20	Displacer status ~ inhibitory status Random effect = phylogeny	'Strain loss' criteria = not included.	57	2.60994	0.5044 to 4.8716	0.006**
21	Displacer status ~ inhibitory status Random effect = phylogeny	Data points = each 'strain x participant' interaction.	69	2.5757	0.6732 to 4.6457	0.004**
22	Displacer status ~ inhibitory status Random effect = phylogeny	Inhibitor screen type = intraspecific + interspecific competitors	57	3.44069	0.3475 to 6.7814	0.006**

Table S7. 'Short-term success' strain-level displacement analyses – MCMCglmm analyses. All models in Table S7 are MCMCglmms (family = "categorical") (Hadfield, 2010, 2019). Residual variance was fixed to equal one (V=1, fix=1) and uninformative priors were used for all random effects (V=1, nu=0.002) (Hadfield, 2019). 'Posterior mean' represents the log odds of inhibitor strains being a displacing strain compared to the log odds of non-inhibitor strains being a displacing strain (i.e., the log odds ratio) and '95% credible intervals' represent the 95% credible intervals of the log odds ratio. A pMCMC can generally be interpreted in the same way as a standard P value (Hadfield, 2019). All models only contain data from co-colonised participants (n=20), as single-strain participants don't contain any displacement events. The first row (Model ID = 18) describes the dataset used in the main model; the following rows (Model IDs = 19-22) state how each aspect of the dataset was individually varied compared to the dataset used in the main model, to test the robustness of the main result. Statistical significance is denoted with an asterisk(s) (NS = p > 0.05; * = p < 0.05; ** = p < 0.01; *** = p < 0.001). See 'Supplementary Methods – data analysis – section 2.2 & 3' for more analysis details.

Model ID	Model description	Dataset description	Sample size	Model estimate	Standard error	P-value
23	Inhibitory status ~ Displacement event type	Time frame = up to 24-months. ‘Strain loss’ criteria = included. Data points = each time point isolate. Inhibitor screen type = *C. fimi*.	216	Single compared to co-occur = -0.7444. Defence compared to co-occur = 0.5083. Fair-fight compared to co-occur = 0.7108. Invasion compared to co-occur = 0.6827. Unknown-displacement compared to co-occur = 0.4595. Clearance compared to co-occur = -0.5671.	Single compared to co-occur = 0.3696. Defence compared to co-occur = 0.5496. Fair-fight compared to co-occur = 0.6072. Invasion compared to co-occur = 0.6526. Unknown-displacement compared to co-occur = 0.9447. Clearance compared to co-occur = 1.1221.	Single compared to co-occur = 0.1250 (NS). Defence compared to co-occur = 0.3550 (NS). Fair-fight compared to co-occur = 0.2417 (NS). Invasion compared to co-occur = 0.2955 (NS). Unknown-displacement compared to co-occur = 0.6267 (NS). Clearance compared to co-occur = 0.5071 (NS).
24	Inhibitory status ~ Displacement event	Time frame = up to 24-months. ‘Strain loss’ criteria = included. Data points = each time point isolate. Inhibitor screen type = *C. fimi*.	216	Displacement compared to non-displacement = 0.8690. Clearance compared to non-displacement = -0.4780.	Displacement compared to non-displacement = 0.3486. Clearance compared to non-displacement = 1.1104.	Displacement compared to non-displacement = 0.0127* Clearance compared to non-displacement = 0.6668 (NS).
25	Displacement winner identity ~ 1	Time frame = up to 24-months. ‘Strain loss’ criteria = included. Data points = each displacement event associated with inhibitors. Inhibitor screen type = *C. fimi*.	19	1.030	0.521	0.0481*

Table S8. ‘Short-term success’ isolate-level displacement analyses – Binomial GLM analyses. All models in Table S8 are binomial GLMs (family = “binomial”). In models 23 and 24, ‘model estimate’ represents the log odds of different displacement event types (model 23), or all displacement events (model 24), being associated with inhibitory isolates, compared to a specific type of non-displacement event, ‘co-occur’ (model 23), or all non-displacement events (model 24) (i.e., the log odds ratio). In model 25, ‘model estimate’ represents the log odds of an inhibitory isolate winning a displacement event, compared to non-inhibitory isolate. ‘Standard error’ represents the standard error of the log odds ratio (models 23 & 24), or log odds (model 25). All models only contain data from co-colonised participants (n=20), as single-strain participants don’t contain any displacement events. Statistical significance is denoted with an asterisk(s) (NS = p > 0.05; * = p < 0.05; ** = p < 0.01; *** = p < 0.001). See ‘Supplementary Methods – data analysis – section 2.2 & 3’ for more analysis details.

Supplementary results – genomic analyses

Participant ID	Month	*Spa*-type	*C. fimi*	Interspecific	Intraspecific indicator	Intraspecific nasal	BGC type(s)	Bacteriocin prediction	Stable BGC(s)	Stop codon
22	24	t012	no	no	no	no	none	none	NA	NA
42	48	t1716	no	no	no	no	none	none	NA	NA
132	22	t228	no	no	no	no	none	none	NA	NA
162	6	t209	no	no	no	no	none	none	yes	NA
162	48	t209	yes	no	no	no	none	none	yes	NA
359	6	t089	yes	yes	no	no	Lanthipeptide class ii	Bicereucin BsjA2-like	yes	no
359	46	t089	yes	yes	no	no	Lanthipeptide class ii	Bicereucin BsjA2-like	yes	no
420	6	t379	no	no	no	no	none	none	NA	NA
424	30	t3304	no	no	no	no	none	none	NA	NA
450	4	t382	yes	yes	no	no	none	none	NA	NA
450	88	t382	yes	yes	no	no	none	none	NA	NA
451	4	t379	no	no	no	no	none	none	NA	NA
454	24	t127	no	no	no	no	Lanthipeptide class i	Bsa	NA	no
499	24	t021	no	no	no	no	none	none	NA	NA
637	8	t870	no	no	no	no	none	none	NA	NA
637	12	t7050	yes	no	yes	no	RiPP-like	Putative bacteriocin	yes	no
637	24	t7050	yes	no	yes	no	RiPP-like	Putative bacteriocin	yes	no

638	12	t171	yes	yes	no	no	none	none	NA	NA
638	18	t065	no	no	no	no	none	none	NA	NA
638	24	t171	yes	yes	no	no	none	none	NA	NA
647	4	t019	no	no	no	no	none	none	NA	NA
647	8	t6855	yes	no	no	no	none	none	NA	NA
647	12	t230	no	no	no	no	none	none	NA	NA
647	24	t6855	yes	no	no	no	none	none	NA	NA
671	24	t032	no	no	no	no	none	none	NA	NA
686	8	t3262	no	no	no	no	none	none	NA	NA
688	6	t3262	no	no	no	no	none	none	NA	NA
688	18	t002	no	no	no	no	none	none	NA	NA
903	4	t382	no	no	no	no	none	none	NA	NA
903	16	t2643	no	no	no	no	none	none	NA	NA
926	6	t008	yes	no	no	no	Lanthipeptide class i	Bsa	yes	no
926	8	t008	no	no	no	no	Lanthipeptide class i	Bsa	yes	no
926	14	t127	no	no	no	no	Lanthipeptide class i	Bsa	NA	no
926	24	t008	yes	no	no	no	Lanthipeptide class i	Bsa	yes	no
930	8	t2074	no	no	no	no	none	none	NA	NA
930	8	t499	no	no	no	no	none	none	NA	NA
930	20	t120	no	no	no	no	none	none	NA	NA

930	22	t015	no	no	no	no	none	none	NA	NA
952	24	t084	no	no	no	no	none	none	NA	NA
967	2	t171	yes	yes	no	no	none	none	NA	NA
967	18	t171	yes	yes	no	no	none	none	NA	NA
971	8	t008	yes	yes	no	no	Lanthipeptide class i	Bsa	yes	no
971	16	t230	no	no	no	no	none	none	NA	NA
972	6	t127	no	no	no	no	Lanthipeptide class i	Bsa	yes	no
997	2	t015	yes	no	no	no	none	none	NA	NA
1045	16	t084	no	no	no	no	none	none	NA	NA
1045	16	t209	no	no	no	no	none	none	NA	NA
1045	22	t2119	no	no	no	no	none	none	NA	NA
1092	16	t021	no	no	no	no	none	none	NA	NA
1209	40	t346	no	no	no	no	none	none	NA	NA
1212	0	t056	no	no	no	no	none	none	NA	NA
1231	6	t160	no	no	no	no	none	none	NA	NA
1231	26	t120	yes	no	no	no	none	none	NA	NA
1231	46	t120	yes	no	no	no	none	none	NA	NA
1231	46	t4309	yes	no	no	no	none	none	NA	NA
1231	46	t040	no	no	no	no	none	none	NA	NA
1231	50	t4309	yes	no	no	no	none	none	NA	NA
1231	66	t120	yes	no	no	no	none	none	NA	NA

1231	68	t15780	yes	no	no	no	none	none	NA	NA
1231	68	t120	no	no	no	no	none	none	NA	NA
1231	80	t499	yes	no	no	no	none	none	NA	NA
1307	24	t408	no	no	no	no	none	none	NA	NA
1366	2	t089	yes	yes	no	no	Lanthipeptide class ii	Bicereucin BsjA2-like	yes	no
1366	2	t267	no	no	no	no	none	none	NA	NA
1366	8	t089	yes	yes	no	no	Lanthipeptide class ii	Bicereucin BsjA2-like	yes	no
1366	14	t084	yes	yes	no	no	none	none	NA	NA
1366	16	t089	yes	yes	no	no	Lanthipeptide class ii	Bicereucin BsjA2-like	yes	no
1366	24	t084	no	no	no	no	none	none	NA	NA
1367	0	t7409	no	no	no	no	none	none	NA	NA
1367	22	t346	no	no	no	no	none	none	NA	NA
1378	8	t096	no	no	no	no	none	none	NA	NA
2004	4	t1685	yes	yes	no	no	LAP	Listeriolysin S-like	yes	no
2004	4	t321	no	no	no	no	Lanthipeptide class i	Bsa	NA	yes
2004	24	t1685	yes	yes	no	no	LAP	Listeriolysin S-like	yes	no
2009	1	t230	yes	no	no	no	none	none	NA	NA
2009	32	t230	yes	no	no	no	none	none	NA	NA
2030	24	t002	no	no	no	no	none	none	NA	NA
2060	4	t012	no	no	no	no	none	none	NA	NA

2060	24	t084	no	no	no	no	none	none	NA	NA
2064	0	t6825	yes	yes	no	no	none	none	NA	NA
2064	0	t1239	no	no	no	no	none	none	NA	NA
2064	2	t171	yes	yes	yes	yes	Lanthipeptide class ii	Staphylococcin C55	yes	no
2064	6	t6814	no	no	no	no	none	none	NA	NA
2064	12	t171	yes	yes	yes	yes	Lanthipeptide class ii	Staphylococcin C55	yes	no
2064	20	t7031	yes	no	no	no	none	none	NA	NA
2064	20	t008	yes	no	no	no	Lanthipeptide class i	Bsa	yes	no
2064	22	t7031	yes	no	no	no	none	none	NA	NA
2064	24	t008	yes	no	no	no	Lanthipeptide class i	Bsa	yes	no
2104	0	t571	no	no	no	no	none	none	NA	NA

Table S9. An overview of bacteriocin gene cluster (BGC) analysis. Each row represents the bacteriocin phenotype and genomic analysis output for every isolate that was whole-genome sequenced in the collection. For genome selection criteria, see ‘Methods – genome selection & sequencing’. “*C. fimi”* represents inhibition against *C. fimi*, a positive indicator for bacteriocin production; “Interspecific” represents inhibition against any of the three interspecific competitors; “Intraspecific indicator” represent against Newman *ΔdltA,* a positive indicator for intraspecific inhibition; “Intraspecific nasal” represents inhibition against any natural strain of *S. aureus* from the nasal cavity; “BGC type(s)” and “Bacteriocin prediction(s)” represents predictions from antiSMASH(6.0) (Blin *et al.,* 2021) and BAGEL4 (van Heel *et al.,* 2018) about the type and identity of each genomic hit; “Stable BGC(s)” represents whether the BGC within each strain was stable over time, with regards to the presence/absence of the entire BGC region and the presence/absence of all bacteriocin-related genes within the BGC region; “Stop codon” represents whether a stop codon, which could cause loss of function, was identified in any of the bacteriocin-related genes within each BGC. NAs in “Stable BGC(s)” occur if a strain was only sampled at one time point within a participant, and in “Stop codon” if no BGC genes were present to test. We detected no other secondary metabolite biosynthetic gene cluster types known to have antimicrobial effects. The data in this table is summarized in Fig. 4.

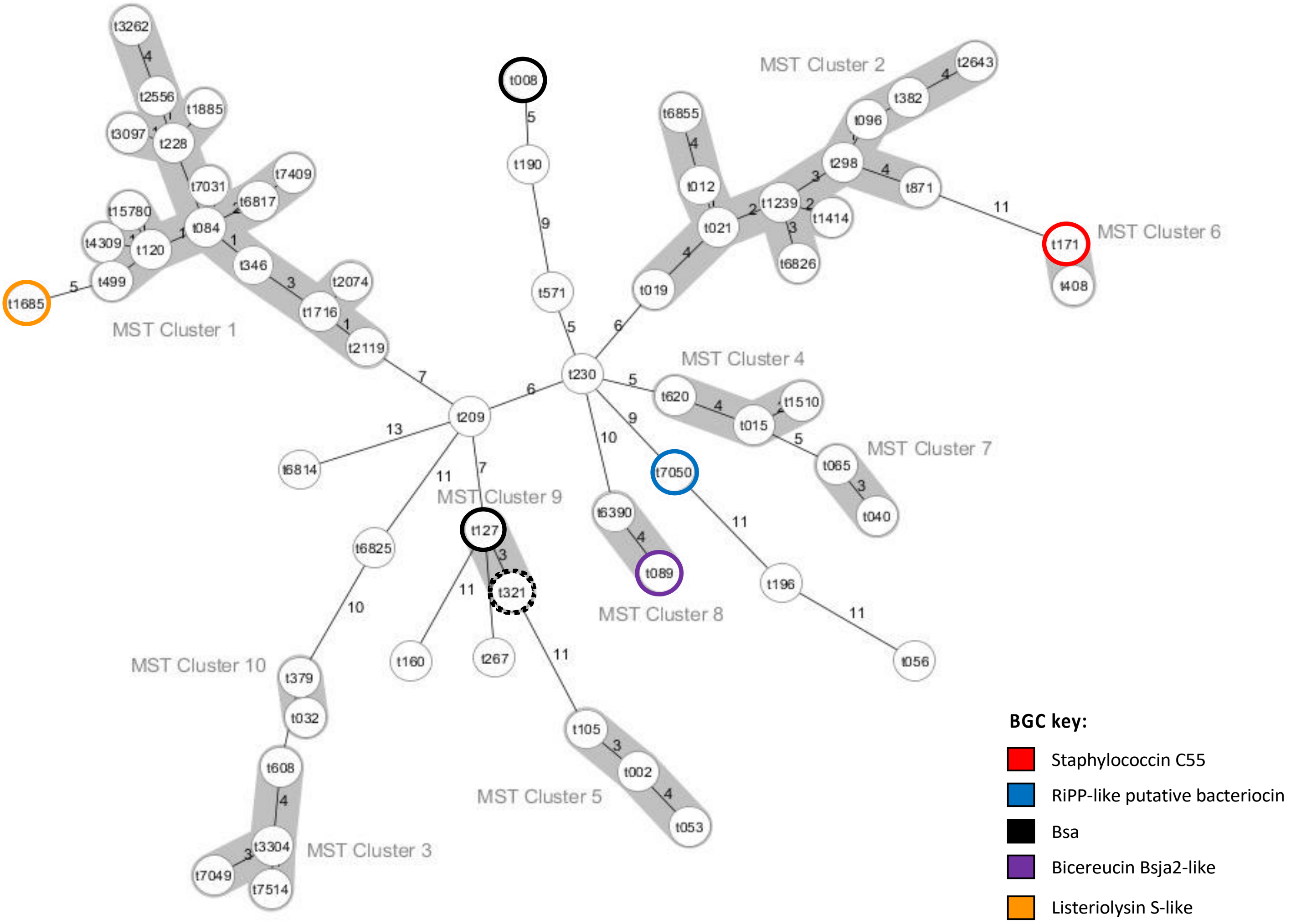


Fig S7. A *spa* gene minimum-spanning genetic distance tree representing which spa-types carry each bacteriocin gene cluster (BGC) type. The genetic distance tree is based on the *spa* gene sequence of each strain in the collection and produced using the BURP (based-upon repeat pattern) clustering algorithm (Mellmann *et al*., 2007), integrated within SeqSphere+ (8.4.0) (Ridom, GmbH). Each strain represents an individual '*spa*-type'. Two strains (t528 and t870) were removed from the tree, as they did not contain the required number of repeats (≥ 5) for BURP clustering (Mellmann *et al*., 2007). The remaining strains (n = 62) are denoted as circles and are clustered into minimum-spanning tree (MST) clusters, which represent *spa*-Clonal Complexes (*spa*-CCs), and are denoted by grey shading. Each MST cluster contains *spa*-types with a BURP distance of four or less, the default for BURP clustering (Mellmann *et al*., 2007). Branch length corresponds to the BURP distance between strains, which is also labelled as a numerical value on each branch. All *spa*-types that carry identifiable BGCs are circled with a ring. Ring colour corresponds to the prediction of the bacteriocin associated with the BGC, see 'BGC key' (bottom right). The dashed circle represents that a BGC was identified, but an intragenic stop codon was present in a bacteriocin biosynthesis gene, potentially causing loss-of-function.

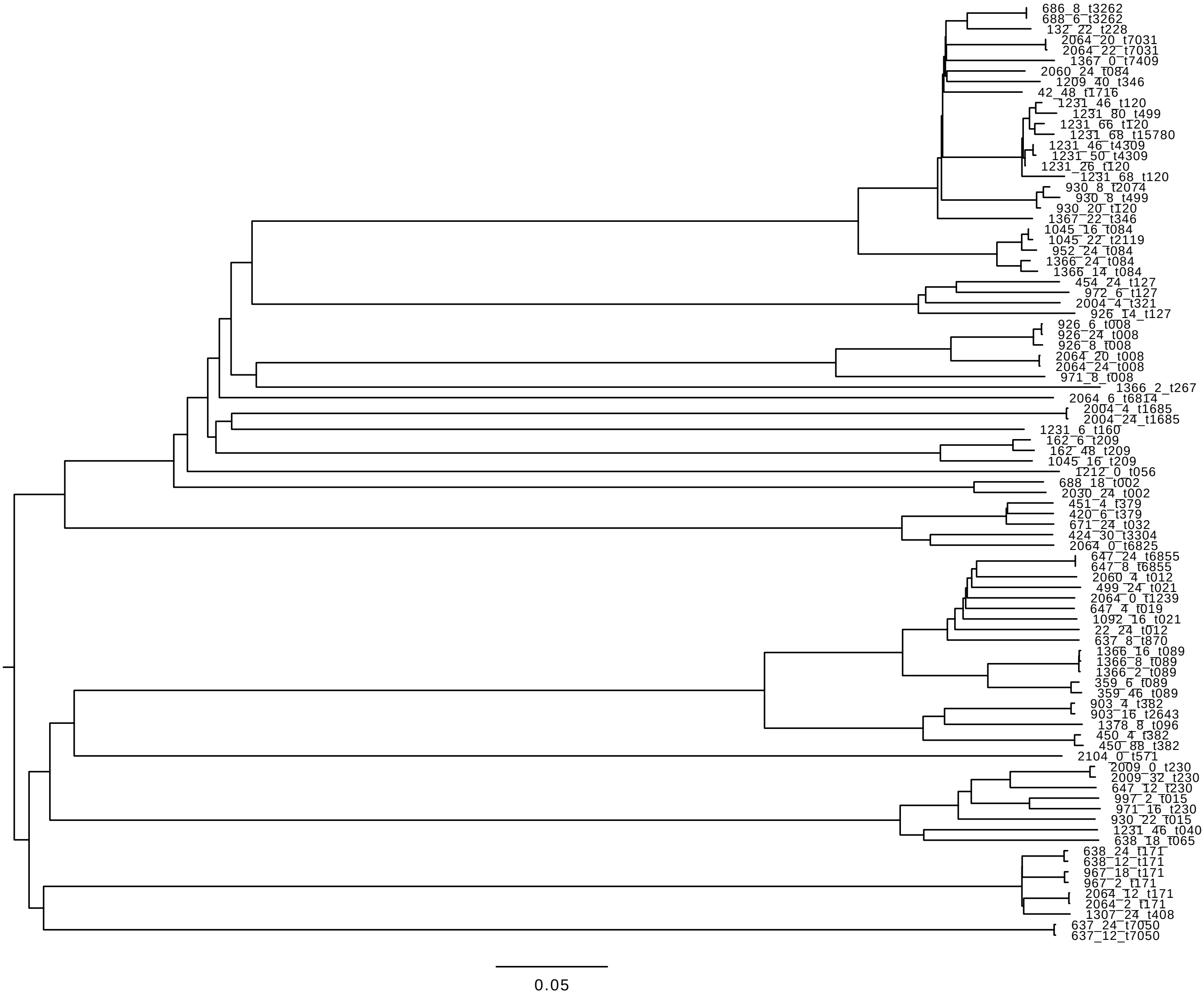


Fig S8. A core genome MLST (cgMLST) phylogenetic tree of the 89 *S. aureus* genomes used in this study. We used SeqSphere+ (v8.4) (Ridom, GmbH) to produce the tree based on the genome-wide allelic profiles of 1861 loci. The genome for each isolate is labelled with the relevant isolate ID, which consists of: participant number, sample time point, *spa*-type. The tree was midpoint rooted using FigTree (v1.4.4).

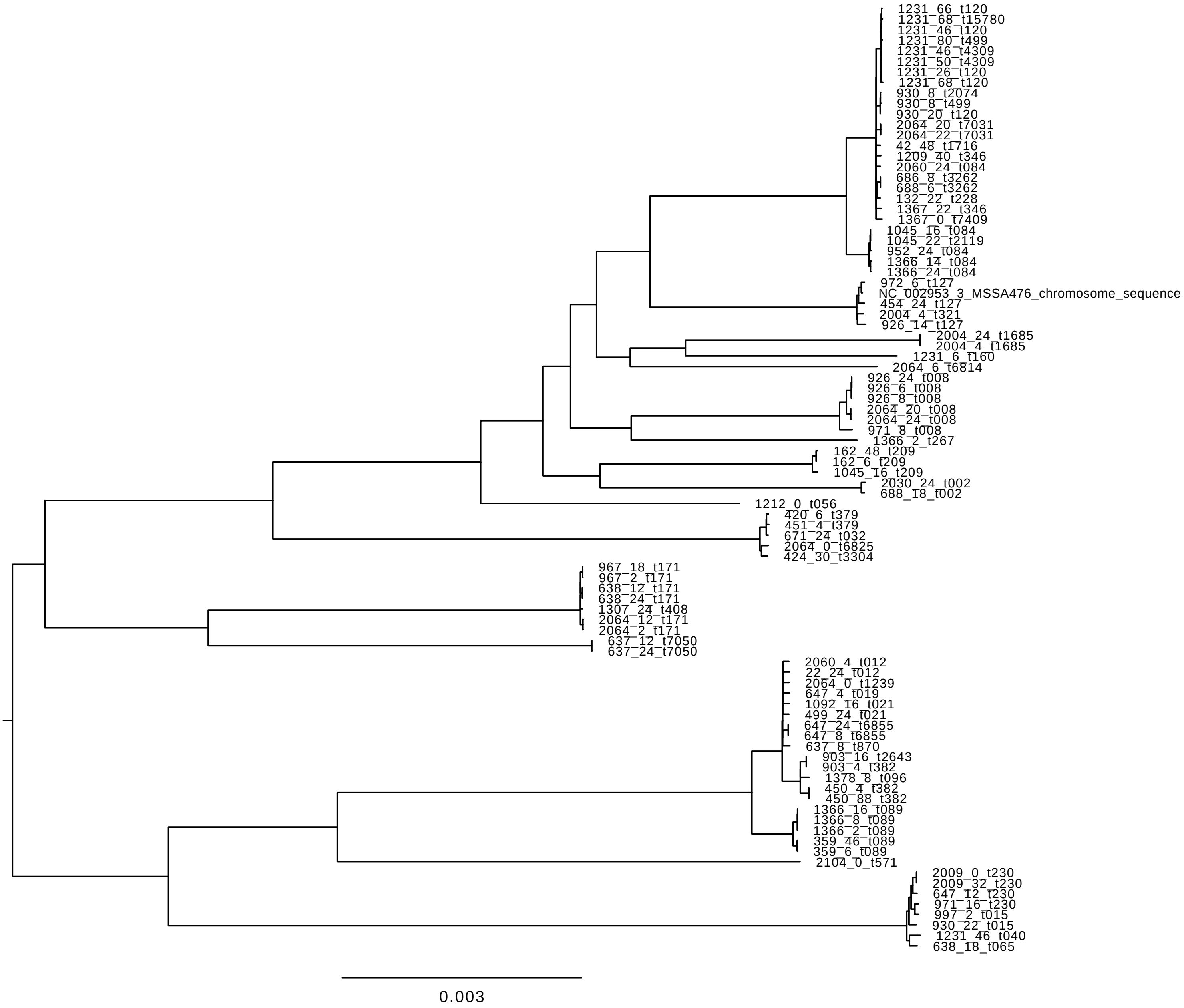


Fig S9. A maximum likelihood phylogenetic tree of the 89 *S. aureus* genomes used in this study. We produced this tree by using REALPHY (1.13) (Bertels *et al.*, 2014) to align our assembled *S. aureus* contigs to the complete reference genome MSSA 476 using Bowtie2 (v2.5.1) (Langmead & Salzberg, 2012) and estimate a maximum likelihood tree using PhyML (v3.0) (Guindon *et al.*, 2010). The genome for each isolate is labelled with the relevant isolate ID, which consists of: participant number, sample time point, *spa*-type. The tree was midpoint rooted using FigTree (v1.4.4).

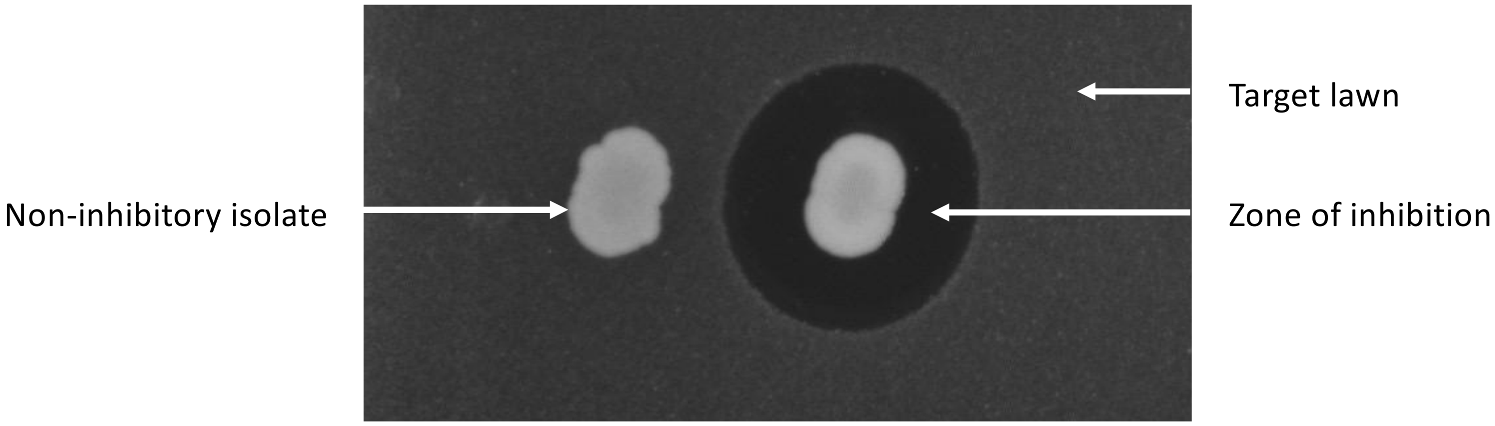


Fig S10. A representative example of the observed zones of inhibition from laboratory screens. *S. aureus* cultures to be tested for inhibitory activity were spotted on top of seeded agar lawns containing the target species or strain being tested against. Inhibitory activity was detected as a clear zone of inhibition surrounding the *S. aureus* spot. In this image, the agar lawn was seeded with positive indicator strain *C. fimi*. The *S. aureus* isolate displaying no inhibitory activity (left) is 132-10 - t228; the isolate displaying a zone of inhibition (right) is 2064-2 – t171, which was identified by genomic analysis to carry a BGC coding for 'staphylococcin C55'.

Chapter 4. Plasmids do not consistently stabilise cooperation across bacteria but may promote broad pathogen host-range

Chapter 4 consists of the following publication in *Nature Ecology & Evolution*.

nature ecology & evolution

ARTICLES

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

Check for updates

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

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

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

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

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

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

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

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

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

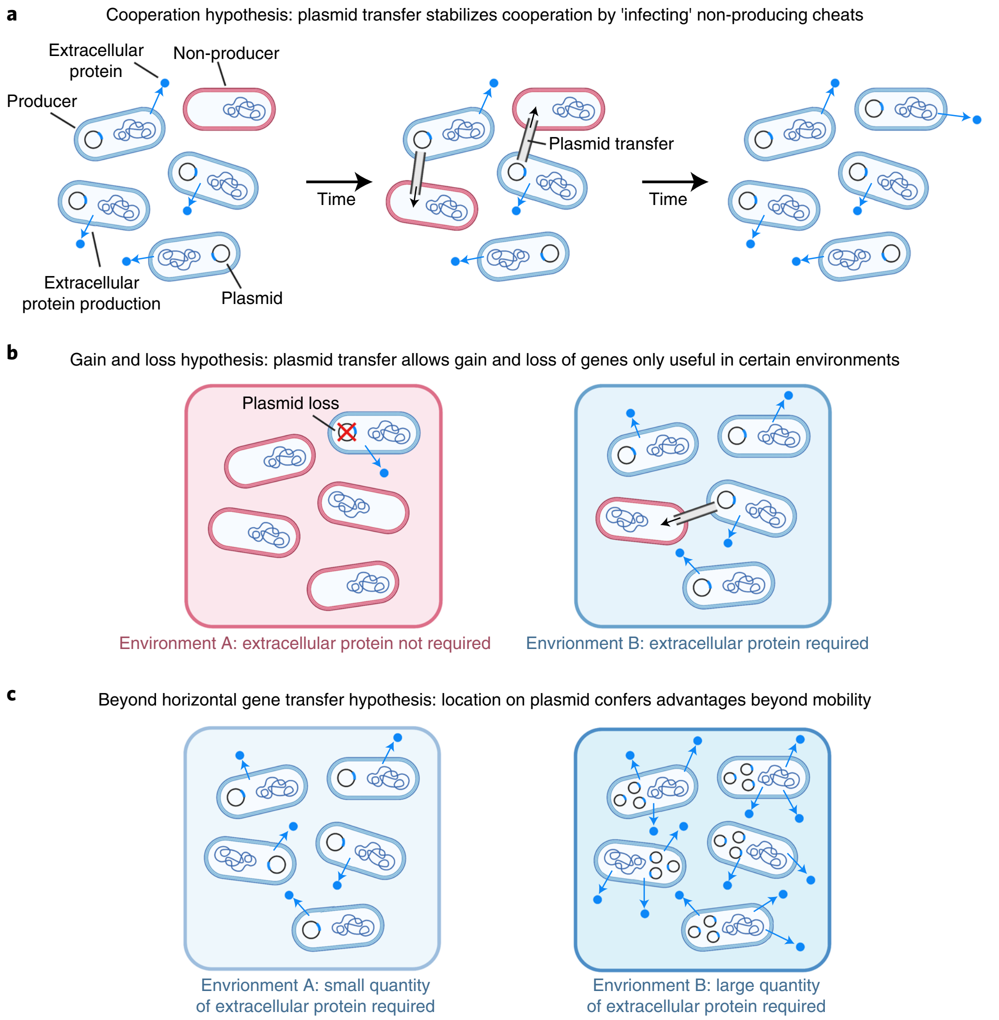

a
Cooperation hypothesis: plasmid transfer stabilizes cooperation by 'infecting' non-producing cheats
Extracellular protein
Non-producer
Producer
Extracellular protein production
Plasmid
Time
Plasmid transfer
Time
b
Gain and loss hypothesis: plasmid transfer allows gain and loss of genes only useful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

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

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

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

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

Results

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

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

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

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

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

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

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

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

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

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

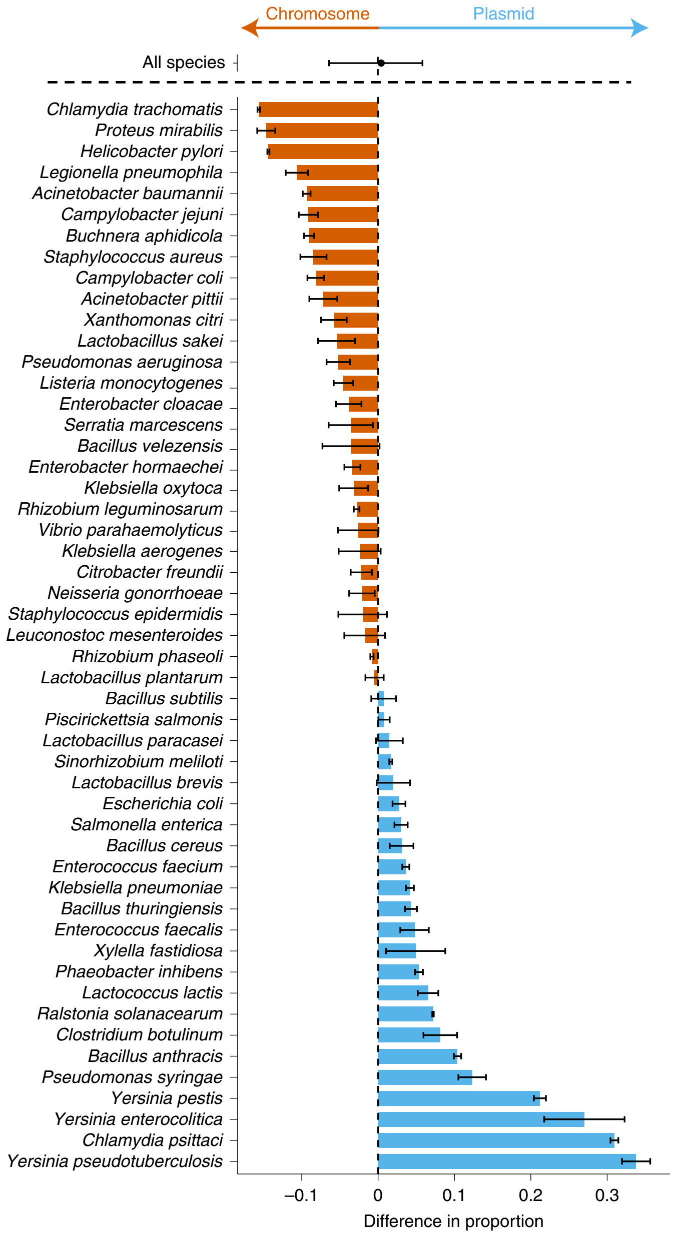


Fig. 2 | Extracellular proteins are not overrepresented on plasmids. For each species we calculated the mean difference between plasmid(s) and chromosomes in the proportion of genes coding for extracellular proteins. Species in blue have a difference greater than zero, meaning their plasmid genes code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

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

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

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

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

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

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

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

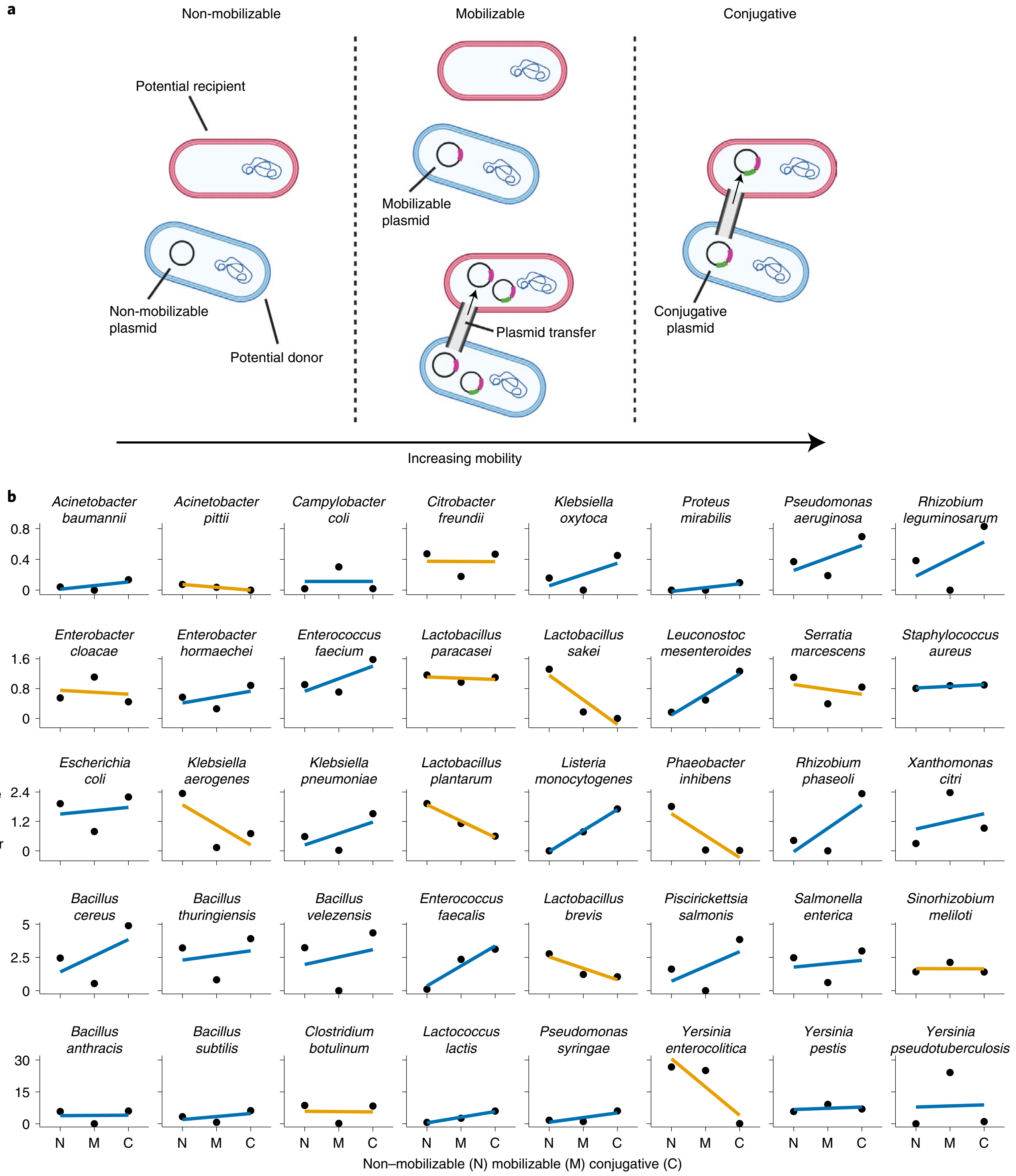


Fig. 3 | Plasmid mobility and extracellular proteins. a, We divided plasmids into three mobility types: non-mobilizable (lowest or no mobility); mobilizable (intermediate mobility); and conjugative (highest mobility). Blue cells are potential plasmid donors, while red cells are potential recipients.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

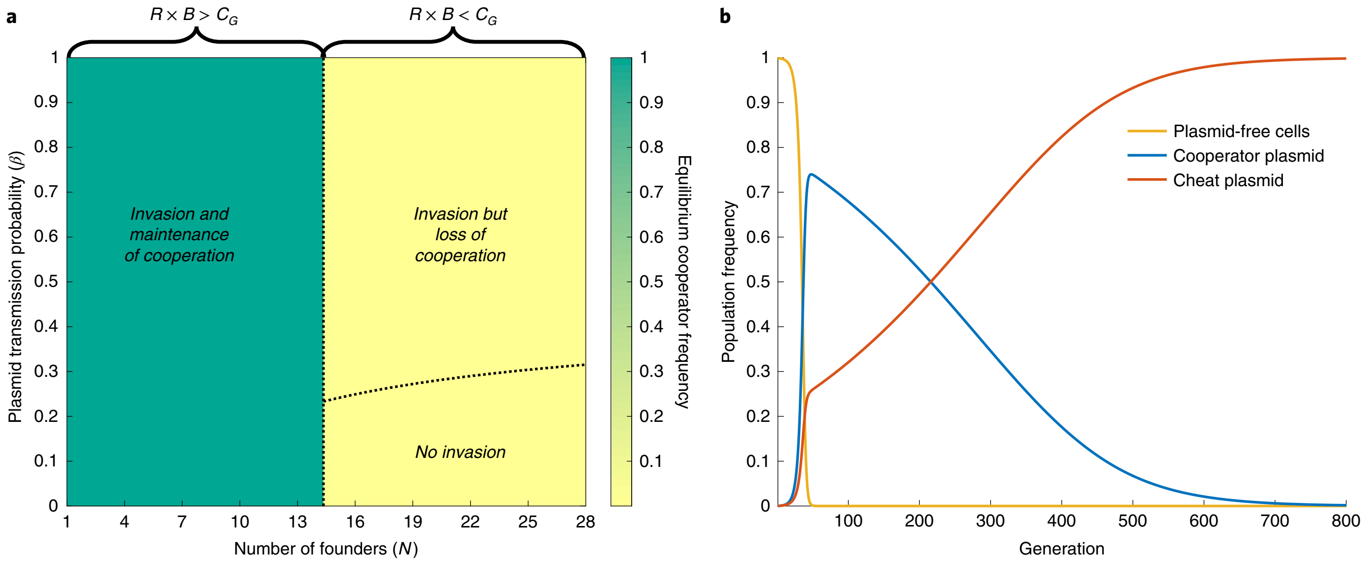


Fig. 4 | Plasmids facilitate the invasion but not the maintenance of cooperation. a,**b**, The results of our theoretical model for the case when there is no plasmid loss ($s=0$): cooperation is only maintained at equilibrium (green shaded area) when it is favoured at the chromosomal level $R\times B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

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

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

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

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

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

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

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

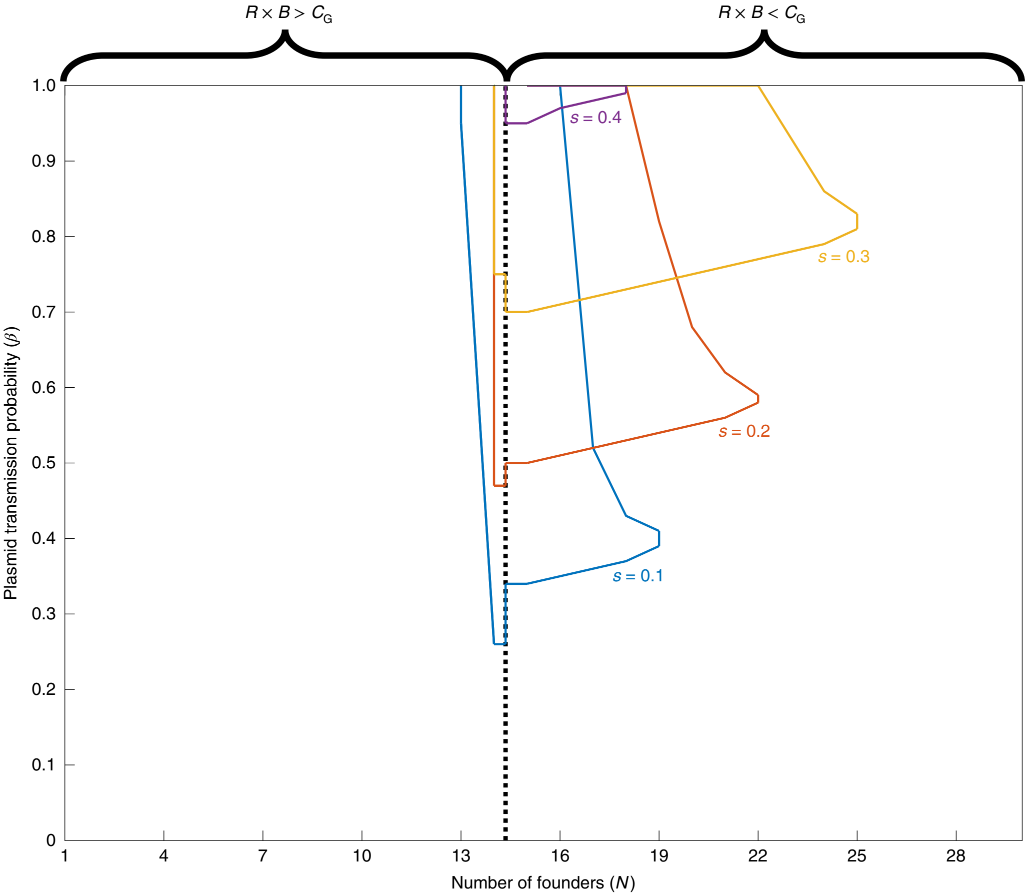


Fig. 5 | Plasmid loss can favour the maintenance of cooperation. We plot the results of our theoretical model for different levels of plasmid loss ($s=0$–1). The areas encapsulated by the coloured lines show the regions of parameter space where cooperation is polymorphic at equilibrium (that is, population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

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

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

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

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

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

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

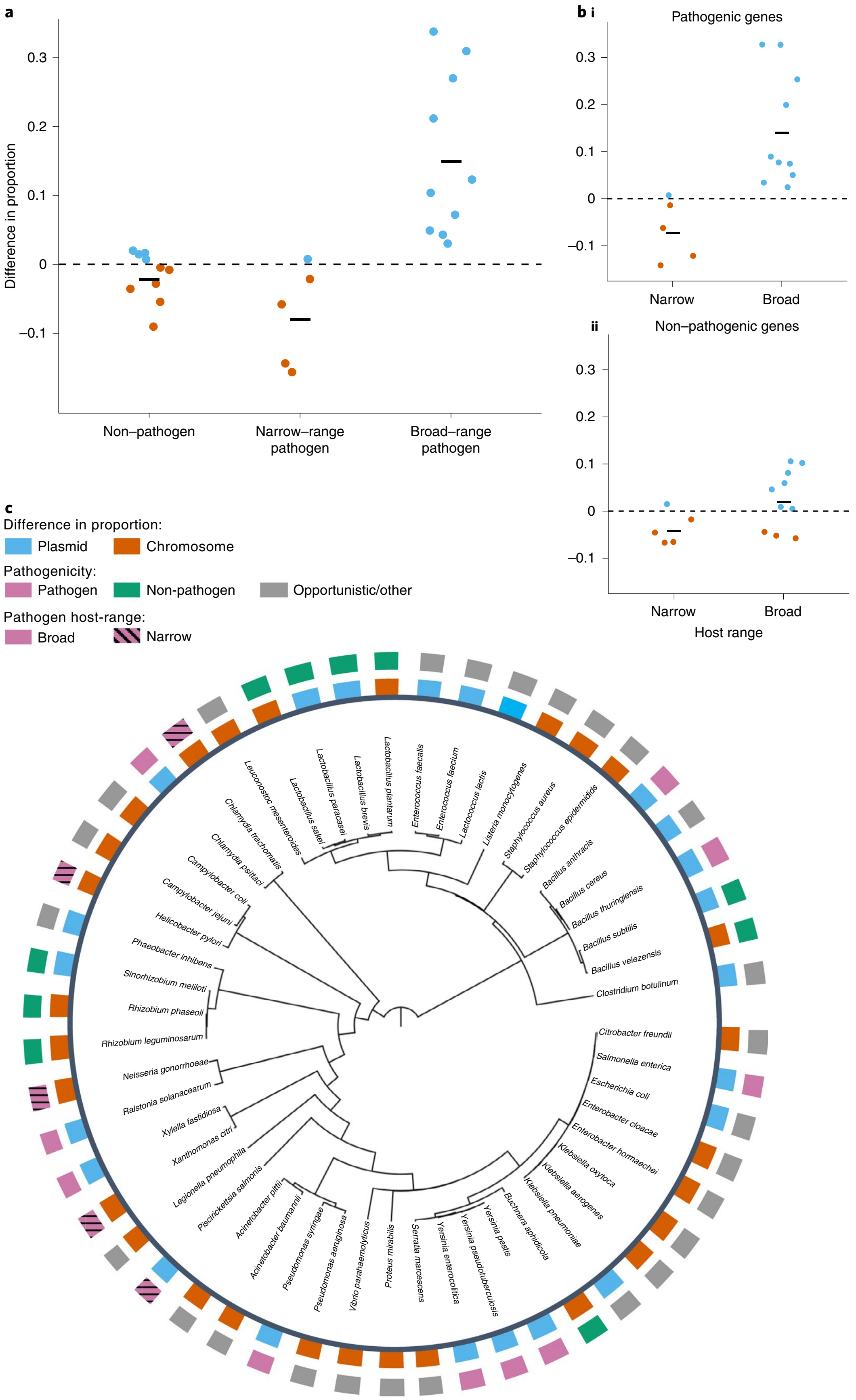
We found that plasmids of broad host-range pathogens were particularly enriched with extracellular proteins involved in facilitating pathogenicity, compared to plasmids of narrow host-range species (Fig. 6b(i)). Specifically, we found that pathogens with a broad host-range were significantly more likely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

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

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

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

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



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

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

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

Discussion

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

Methods

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Data availability

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

Code availability

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

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

References

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

21. Kruskal, W. Miracles and statistics: the casual assumption of independence. *J. Am. Stat. Assoc.* **83**, 929–940 (1988).
22. Ives, A. R. & Zhu, J. Statistics for correlated data: phylogenies, space, and time. *Ecol. Appl.* **16**, 20–32 (2006).
23. Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15 (1985).
24. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, 1991).
25. Grafen, A. The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B.* **326**, 119–157 (1989).
26. Hurlbert, S. H. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54**, 187–211 (1984).
27. Ruxton, G. & Colegrave, N. *Experimental Design for the Life Sciences* (Oxford Univ. Press, 2011).
28. Stone, G. N., Nee, S. & Felsenstein, J. Controlling for non-independence in comparative analysis of patterns across populations within species. *Philos. Trans. R. Soc. B* **366**, 1410–1424 (2011).
29. Ives, A. R., Midford, P. E. & Garland, T. Jr. Within-species variation and measurement error in phylogenetic comparative methods. *Syst. Biol.* **56**, 252–270 (2007).
30. Bakkeren, E. et al. Cooperative virulence can emerge via horizontal gene transfer but is stabilized by transmission. Preprint at *bioRxiv* https://doi.org/10.1101/2021.02.11.430745 (2021).
31. Ghoul, M., Andersen, S. B. & West, S. A. Sociomics: using omic approaches to understand social evolution. *Trends Genet.* **33**, 408–419 (2017).
32. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040 (2017).
33. Niehus, R., Mitri, S., Fletcher, A. G. & Foster, K. R. Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nat. Commun.* **6**, 8924 (2015).
34. Cordero, O. X. et al. Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. *Science* **337**, 1228–1231 (2012).
35. Rakoff-Nahoum, S., Coyne, M. J. & Comstock, L. E. An ecological network of polysaccharide utilization among human intestinal symbionts. *Curr. Biol.* **24**, 40–49 (2014).
36. Nocelli, N., Bogino, P. C., Banchio, E. & Giordano, W. Roles of extracellular polysaccharides and biofilm formation in heavy metal resistance of rhizobia. *Materials* **9**, 418 (2016).
37. Ciofu, O., Beveridge, T. J., Kadurugamuwa, J., Walther-Rasmussen, J. & Høiby, N. Chromosomal β-lactamase is packaged into membrane vesicles and secreted from *Pseudomonas aeruginosa*. *J. Antimicrob. Chemother.* **45**, 9–13 (2000).
38. Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R. C. & San Millán, Á. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat. Rev. Microbiol.* **19**, 347–359 (2021).
39. Yu, N. Y. et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**, 1608–1615 (2010).
40. Rankin, D. J., Rocha, E. P. C. & Brown, S. P. What traits are carried on mobile genetic elements, and why? *Heredity* **106**, 1–10 (2011).
41. Hadfield, J. D. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33**, 1–22 (2010).
42. Clutton-Brock, T. H. & Harvey, P. H. Primate ecology and social organization. *J. Zool.* **183**, 1–39 (1977).
43. Jennions, M. D. & Møller, A. P. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behav. Ecol.* **14**, 438–445 (2003).
44. Crawley, M. J. *Statistics: An Introduction Using R* (John Wiley & Sons, 2014).
45. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Routledge, 1988).
46. Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genom.* **4**, e000206 (2018).
47. Robertson, J., Bessonov, K., Schonfeld, J. & Nash, J. H. E. Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance. *Microb. Genom.* **6**, mgen000435 (2020).
48. Smillie, C., Garcillan-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & de la Cruz, F. Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434–452 (2010).
49. Mc Ginty, S. É. & Rankin, D. J. The evolution of conflict resolution between plasmids and their bacterial hosts. *Evolution* **66**, 1662–1670 (2012).
50. Hamilton, W. D. Genetical evolution of social behaviour I & II. *J. Theor. Biol.* **7**, 1–52 (1964).
51. Hamilton, W. D. The evolution of altruistic behavior. *Am. Nat.* **97**, 354–356 (1963).
52. Ghigo, J. M. Natural conjugative plasmids induce bacterial biofilm development. *Nature* **412**, 442–445 (2001).
53. Di Venanzio, G. et al. Multidrug-resistant plasmids repress chromosomally encoded T6SS to enable their dissemination. *Proc. Natl Acad. Sci. USA* **116**, 1378–1383 (2019).
54. Sheppard, R. J., Beddis, A. E. & Barraclough, T. G. The role of hosts, plasmids and environment in determining plasmid transfer rates: a meta-analysis. *Plasmid* **108**, 102489 (2020).
55. Rodríguez-Beltrán, J. et al. Genetic dominance governs the evolution and spread of mobile genetic elements in bacteria. *Proc. Natl Acad. Sci. USA* **117**, 15755–15762 (2020).
56. Cornelis, G. R. et al. The virulence plasmid of yersinia, an antihost genome. *Microbiol. Mol. Biol. Rev.* **62**, 1315–1352 (1998).
57. Köstlbacher, S., Collingro, A., Halter, T., Domman, D. & Horn, M. Coevolving plasmids drive gene flow and genome plasticity in host-associated intracellular bacteria. *Curr. Biol.* **31**, 346–357 (2021).
58. Gupta, A., Kapil, R., Dhakan, D. B. & Sharma, V. K. MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PLoS ONE* **9**, e93907 (2014).
59. San Millan, A., Escudero, J. A., Gifford, D. R., Mazel, D. & MacLean, R. C. Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat. Ecol. Evol.* **1**, 0010 (2016).
60. Carrier, T., Jones, K. L. & Keasling, J. D. mRNA stability and plasmid copy number effects on gene expression from an inducible promoter system. *Biotechnol. Bioeng.* **59**, 666–672 (1998).
61. Rodríguez-Beltrán, J. et al. Multicopy plasmids allow bacteria to escape from fitness trade-offs during evolutionary innovation. *Nat. Ecol. Evol.* **2**, 873–881 (2018).
62. Dietel, A.-K., Kaltenpoth, M. & Kost, C. Convergent evolution in intracellular elements: plasmids as model endosymbionts. *Trends Microbiol.* **26**, 755–768 (2018).
63. Rocha, E. P. C. & Danchin, A. Base composition bias might result from competition for metabolic resources. *Trends Genet.* **18**, 291–294 (2002).
64. Garcia-Garcera, M., Touchon, M., Brisse, S. & Rocha, E. P. C. Metagenomic assessment of the interplay between the environment and the genetic diversification of *Acinetobacter*. *Environ. Microbiol.* **19**, 5010–5024 (2017).
65. Kümmerli, R., Schiessl, K. T., Waldvogel, T., McNeill, K. & Ackermann, M. Habitat structure and the evolution of diffusible siderophores in bacteria. *Ecol. Lett.* **17**, 1536–1544 (2014).
66. Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L. & Brüssow, H. Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.* **6**, 417–424 (2003).
67. Burrus, V. & Waldor, M. K. Shaping bacterial genomes with integrative and conjugative elements. *Res. Microbiol.* **155**, 376–386 (2004).
68. O'Brien, F. G. et al. Origin-of-transfer sequences facilitate mobilisation of non-conjugative antimicrobial-resistance plasmids in *Staphylococcus aureus*. *Nucleic Acids Res.* **43**, 7971–7983 (2015).
69. Rodríguez-Rubio, L. et al. Extensive antimicrobial resistance mobilization via multicopy plasmid encapsidation mediated by temperate phages. *J. Antimicrob. Chemother.* **75**, 3173–3180 (2020).
70. Ramsay, J. P. & Firth, N. Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Curr. Opin. Microbiol.* **38**, 1–9 (2017).
71. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 3801–3806 (1999).
72. Cohen, O., Gophna, U. & Pupko, T. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* **28**, 1481–1489 (2011).
73. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* **46**, e5 (2018).
74. Gardy, J. L. & Brinkman, F. S. L. Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* **4**, 741–751 (2006).
75. Ference, C. M. et al. Recent advances in the understanding of *Xanthomonas citri* ssp. *citri* pathogenesis and citrus canker disease management. *Mol. Plant Pathol.* **19**, 1302–1318 (2018).
76. Morris, C. E., Lamichhane, J. R., Nikolić, I., Stanković, S. & Moury, B. The overlapping continuum of host range among strains in the *Pseudomonas syringae* complex. *Phytopathol. Res* **1**, 4 (2019).
77. Hadfield, J. D. *MCMCglmm Course Notes* (2019); https://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf
78. Nakagawa, S. & Schielzeth, H. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **4**, 133–142 (2013).
79. Nakagawa, S., Johnson, P. C. D. & Schielzeth, H. The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R. Soc. Interface* https://doi.org/10.1098/rsif.2017.0213 (2017).
80. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
81. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
82. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
83. Washburne, A. D. et al. Methods for phylogenetic analysis of microbiome data. *Nat. Microbiol.* **3**, 652–661 (2018).
84. Som, A. Causes, consequences and solutions of phylogenetic incongruence. *Brief. Bioinform.* **16**, 536–548 (2015).

Acknowledgements

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

Author contributions

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

Competing interests

The authors declare no competing interests.

Additional information

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

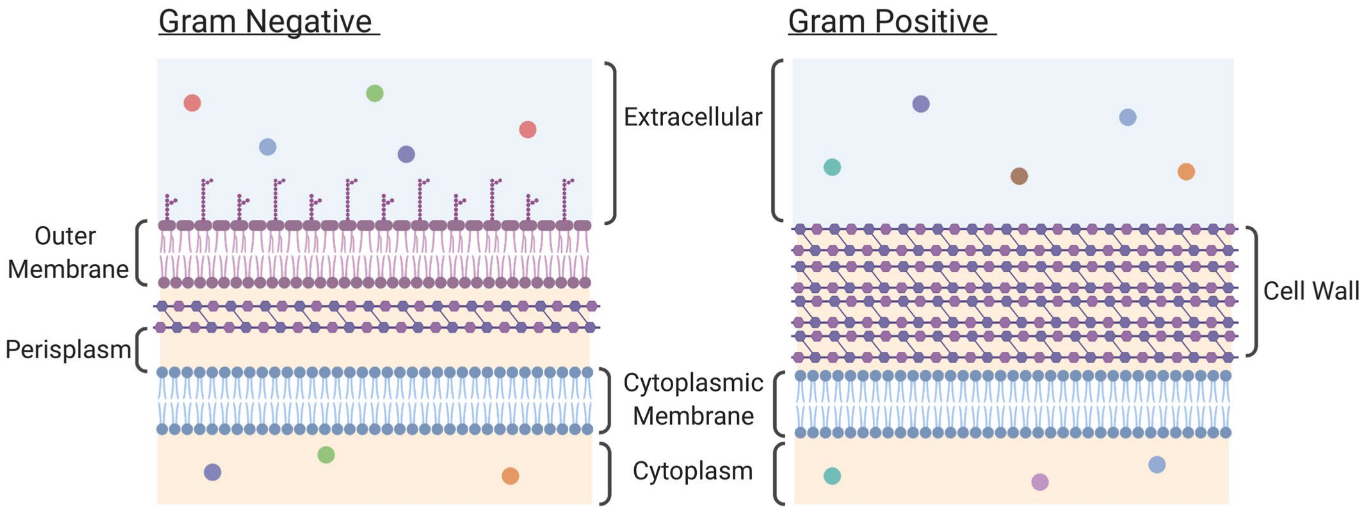
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-021-01573-2.

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

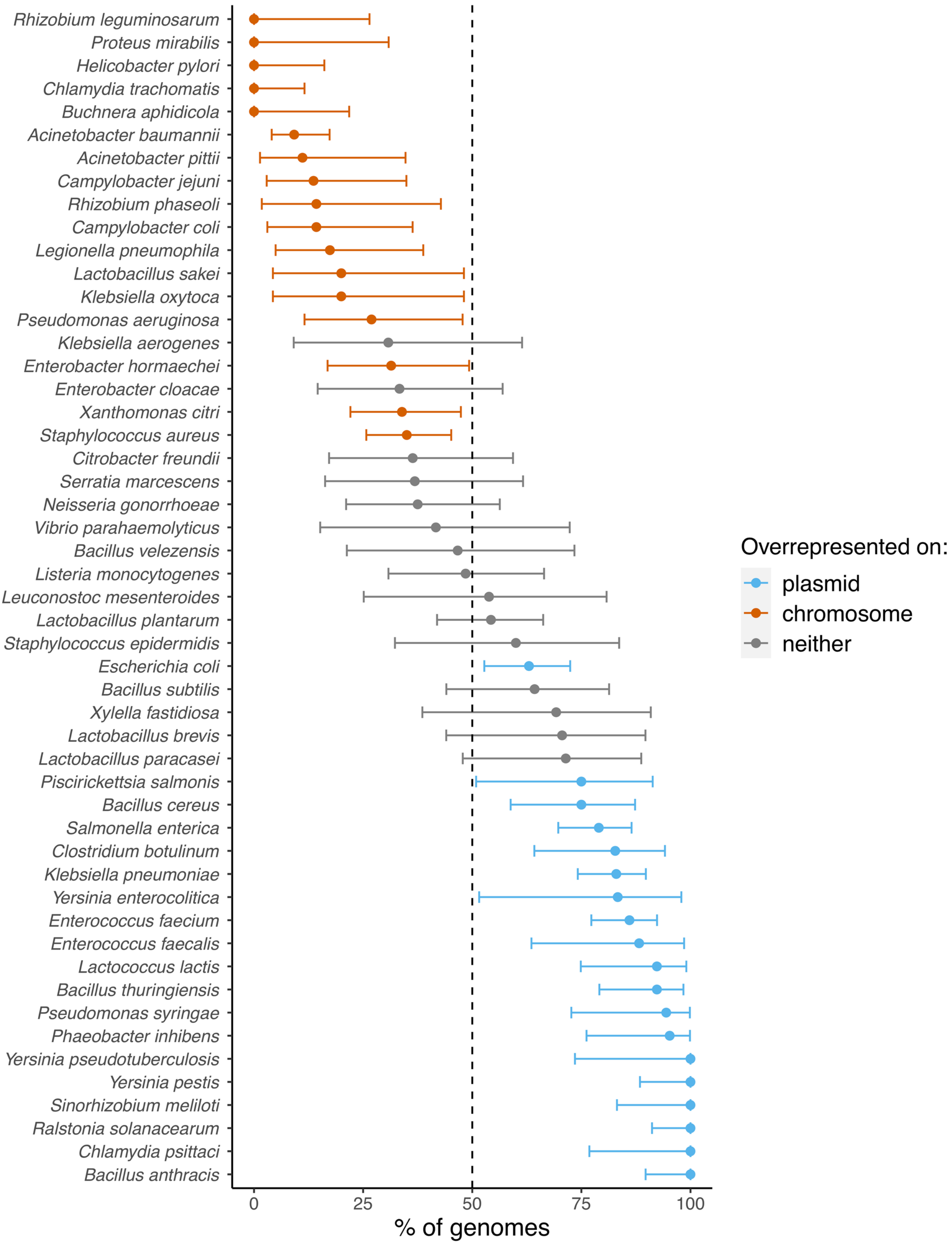
Peer review information *Nature Ecology & Evolution* thanks Isabel Gordo, Alex Washburne and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

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

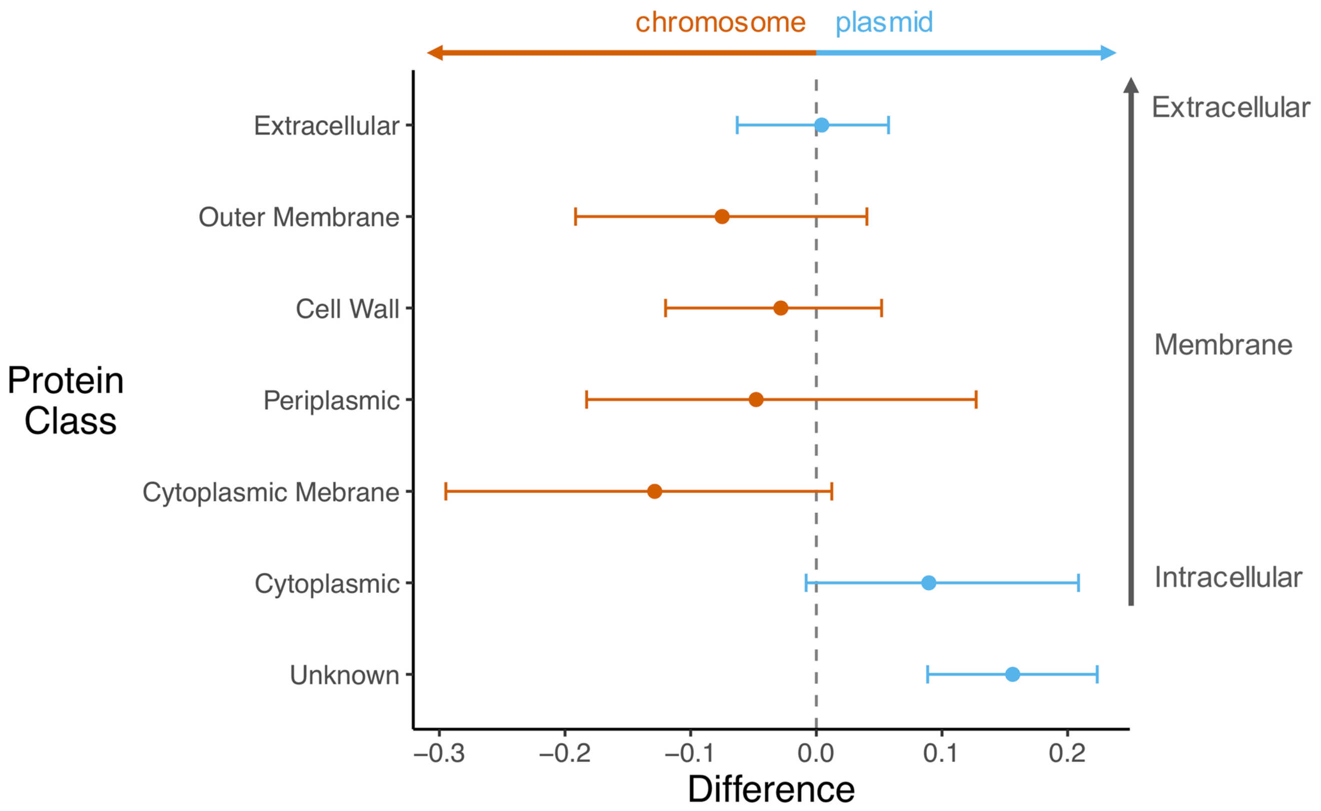


Extended Data Fig. 1 | Protein subcellular localizations. Visualization of all possible subcellular locations predicted by PSORTb. The left panel shows a cross-section of a typical Gram-negative bacterium and the right panel shows the equivalent for a Gram-positive bacterium. Both kinds of bacteria have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

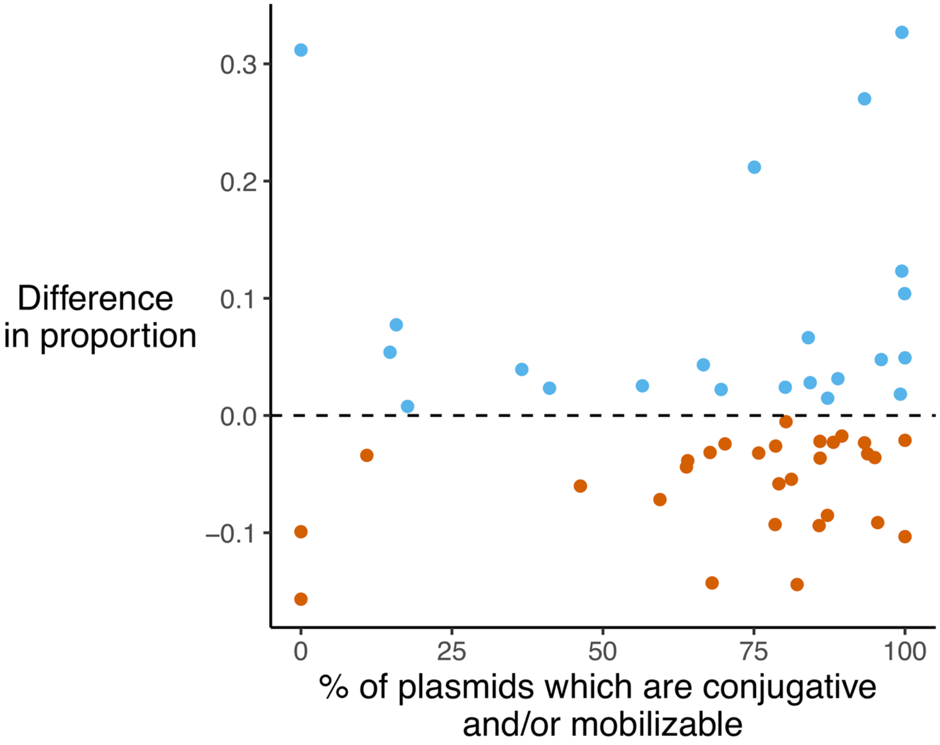


Extended Data Fig. 2 | See next page for caption.

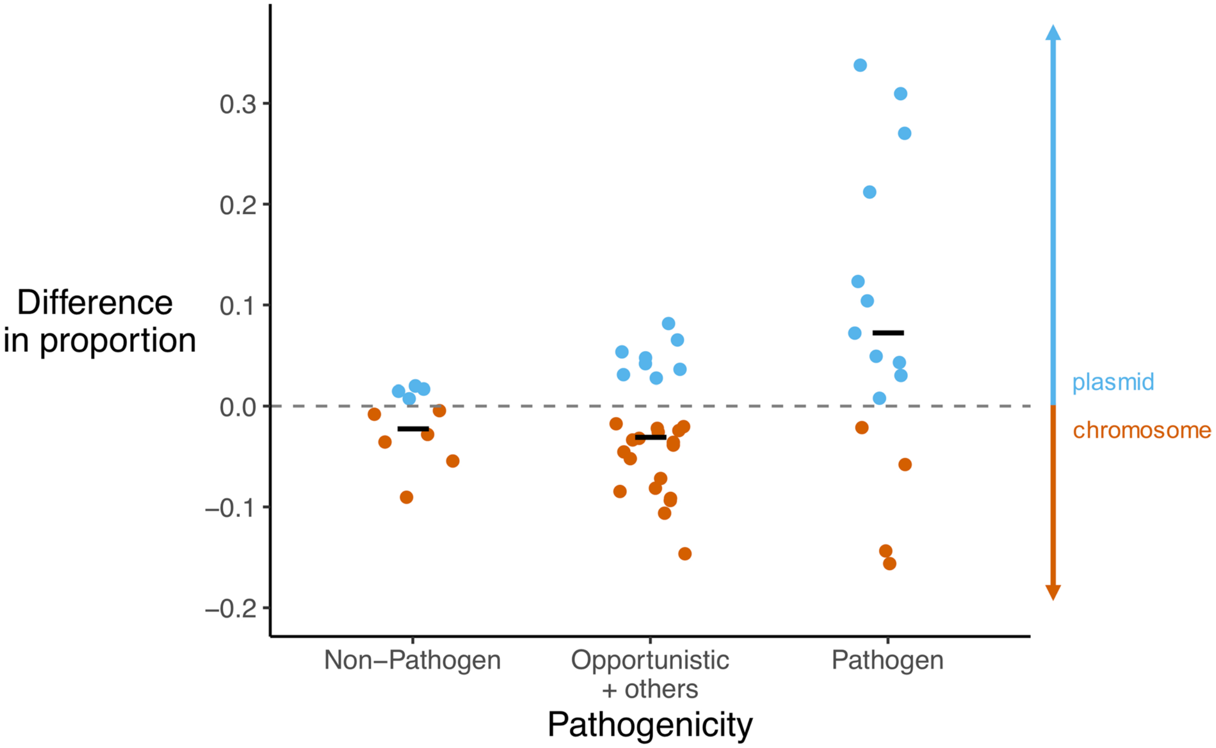
Extended Data Fig. 2 | Substantial variation within and between species in the genomic location of extracellular proteins. The x-axis is the % of genomes in each species where the proportion of plasmid proteins predicted as extracellular is greater than the proportion of chromosome proteins predicted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



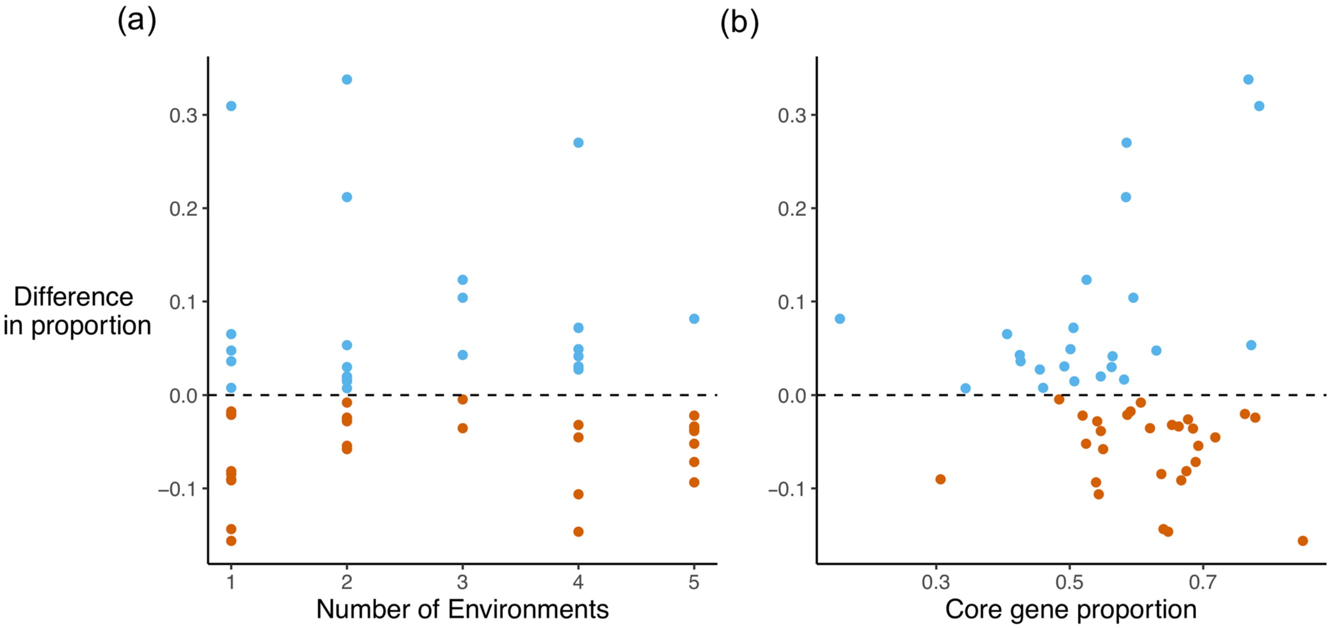
Extended Data Fig. 3 | Difference in plasmid and chromosome proportion for all protein classes predicted by PSORTb. The x-axis is the difference in plasmid and chromosome extracellular proportions, as in Fig. 2. The y-axis is all possible subcellular locations predicted by PSORTb. These protein 'classes'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



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



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



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

Chapter 5. Discussion

Each of my previous chapters contains its own individual discussion section. Here, I aim to highlight the key findings, insights, and future directions arising from my thesis.

The aim of my thesis was to expand our understanding of bacterial sociality in natural populations and across the bacterial tree of life. Testing predictions from theoretical and laboratory studies in more natural populations and in diverse non-model species is crucial if we are to understand the extent and form of bacterial social interactions *in situ*. The controlled and simplified conditions in theoretical and laboratory experiments have proven invaluable in identifying the expected causes and consequences of social traits, which can be extrapolated to broadly understand sociality (Riley & Wertz, 2002; Foster *et al.,* 2004; Griffin *et al.,* 2004; Foster & Wenseleers, 2006; Diggle *et al.,* 2007; Ross-Gillespie *et al.,* 2007, 2009, 2015; Kümmerli *et al.,* 2009a, 2009b, 2015; Dragoš *et al.,* 2018). However, natural conditions are typically more complex and variable: cells live in complex microbial communities and experience diverse abiotic and biotic selection pressures, which can vary across time and space. Each species also experiences selection pressures unique to its habitat and ecological niche, which will often considerably vary from those of model organisms. All of this complexity means that natural populations may be considerably challenging to study and provide results that deviate from theoretical or laboratory predictions for reasons that are difficult to interpret. It is nevertheless a crucial next step: studies of natural populations help us to identify and define parameters which will help make theoretical and laboratory experiments more relevant. This, in turn, will allow us to generate further testable predictions to develop our understanding of the natural world.

Taken together, my thesis demonstrates the importance of testing theory and laboratory predictions in natural populations and across diverse bacterial species. Firstly, using longitudinally sampled populations of *Staphylococcus aureus* from the human nasal cavity, I tested the hypothesis that bacteriocin production provides strains with a competitive advantage and can determine the outcome of competitive strain dynamics. While I found that inhibitory strains were associated with the propensity to displace other *S. aureus* strains from the nasal cavity, which would appear to be generally be predicted by theory and laboratory studies, the form and extent of this benefit was unexpected: interspecific inhibition was associated with intraspecific displacement, rather than direct intraspecific killing, and inhibitory activity was not observed in the majority of strains, despite being associated with a competitive advantage.

Secondly, by performing a comparative analysis of the gene content of plasmids and chromosomes in 51 diverse bacterial species, we tested the widely accepted prediction that horizontal gene transfer (HGT) *via* plasmids can stabilise cooperation across bacteria, by allowing cooperators to re-infect cheats with genes coding for social traits. We found that, contrary to this prediction, HGT does not appear to consistently stabilise cooperation across bacteria, as extracellular proteins were not more likely to be coded on: i) plasmids compared to chromosomes; ii) plasmids with a higher rate of transfer. Instead, we provide evidence the support this role of environmental variability in determining the location of genes coding for extracellular proteins, but only when focusing on broad host-range pathogens. Again, this demonstrates the value of testing hypotheses made by theoretical and laboratory experiments, this time by using genomic data instead of a culture-based approach.

In the rest of this discussion, I explore the insights from my thesis in more detail, and suggest future directions to investigate their implications for the field of sociomicrobiology.

Future directions

Characterising the role of bacteriocins in natural populations

One of the great challenges of studying the consequences of bacterial social traits, such as bacteriocins, in natural populations is isolating their effects in complex environments. In Chapter 3, I provide an important first step, by testing whether bacteriocin activity correlates with measures of success, including ability to persist, or displace competitors, in the human nasal cavity. While this approach is strengthened by focusing on strains from many genetic backgrounds and controlling for phylogenetic non-independence, additional genomic and laboratory approaches could be taken to further isolate the role of bacteriocins.

One genomics-based approach would be to perform a genome-wide association study (GWAS) (Uffelmann *et al.*, 2021), which I have begun exploring as a future project with my supervisor, Prof. Danny Wilson. GWASs are used to determine significant statistical associations between genetic variants and a trait of interest (Uffelmann *et al.*, 2021). I plan to implement the GWAS on the *S. aureus* whole-genome sequences used in Chapter 3, to test whether genetic variants are associated with: i) inhibitory phenotypes; ii) displacement ability. Although the sample sizes are modest by GWAS standards, testing inhibitory phenotypes might allow me to further identify the genetic variants associated with inhibitory activity; testing displacement ability might allow me to further isolate the role of bacteriocin-associated genetic variants in allowing strains to displace competitors, or identify additional traits that could be contributing to displacement. Additionally, this study has the potential to highlight other social traits that could influence differential strain colonisation success in the human nasal cavity.

The role of bacteriocins could also be further isolated using laboratory approaches. Future studies could perform competition experiments, whereby inhibitory strains with and without knock-out mutations in bacteriocin-related genes are competed against intraspecific and interspecific competitors to determine the fitness benefits and costs associated with bacteriocin production. To follow up on my finding that interspecific inhibition was associated with intraspecific displacement, it would be particularly interesting to perform competition experiments using microbial communities consisting of three or more species, to determine whether *S. aureus* and other species use multiple mechanisms of competition in synergy to compete on multiple fronts and successfully invade, defend, or co-exist in communities.

Despite it being well known that bacterial species generally carry many competitive traits (Ghoul & Mitri, 2016; Stubbendieck & Straight, 2016; Granato *et al.,* 2019), we know remarkably little about how they can be used in synergy. These studies should also aim to replicate natural conditions expected to affect bacteriocin activity wherever possible, such as habitat structure, nutrient availability, cell abundance and density (Nadell *et al.,* 2016; Granato *et al.,* 2019). Recent years have seen the development of a range of model systems that increasingly allow us to replicate natural conditions. For example, for *S. aureus* nasal colonisation alone, studies have developed synthetic nasal media that resembles the nutrient composition of nasal fluid (Krismer *et al.,* 2014) and synthetic noses that attempt to recreate the physical structure of the nasal cavity (de Borja Callejas *et al.,* 2014). Analogous systems have also been developed for the study of bacteria in many other habitats (Palmer *et al.,* 2007; Cidem *et al.,* 2020; Kumar *et al.,* 2021; O'Toole *et al.,* 2021; Kumar & Foster, 2022).

Selection for bacteriocin production and activity spectra

I find that, despite being associated with competitive benefit, bacteriocin activity was not detected in the majority of *S. aureus* strains (~27%) isolated from the nasal cavity. It therefore appears that bacteriocins are useful in some host environmental conditions but not in others, which relates to a generally important question in the field of bacteriocins: what are the causes of bacteriocin production? And why do only some strains in some hosts produce bacteriocins? Future natural studies should aim to test the environmental conditions identified by theoretical and laboratory studies to select for bacteriocin production. For example, high cell abundance and density are expected to select for bacteriocin production (Adams *et al.,* 1979; Chao & Levin, 1981). Although cell abundance is known to be relatively low in the nasal cavity compared to other parts of the human microbiome (Krismer *et al.,* 2017), several studies, including Chapter 3 of my thesis, have identified that the production of diffusible antimicrobial toxins such as bacteriocins does occur in species residing in the nasal cavity, that it can occur at high frequencies (Janek *et al.,* 2016), and that it can have important ecological roles, including reducing the ability of *S. aureus* to colonise the nasal cavity (Iwase *et al.,* 2010; Yan *et al.,* 2013; Zipperer *et al.,* 2016). A particularly interesting future direction would be to measure the abundance of *S. aureus* strains in each sample of my collection, for example by using amplicon sequencing, and test the prediction that strains present at higher cell abundances are more likely to produce bacteriocins.

In Chapter 3, I also find that, despite bacteriocins originally being thought of as narrow-spectrum toxins and many theoretical and laboratory studies focusing on this activity, *S. aureus* produce bacteriocins that have a relatively broad-spectrum of activity. While previous studies have identified that *S. aureus* bacteriocins can cause intraspecific inhibition (Giambiagi-Marval *et al.,* 1990; de Oliveira *et al.,* 1998; Coelho *et al.,* 2007; Koch *et al.,* 2014; Kawada-Matsuo *et al.,* 2016), it appears these bacteriocins are not prevalent in the nasal cavity, and instead interspecific inhibition is more common. This raises a question that, despite its clinical relevance, has received very little attention from an evolutionary perspective: what causes species to produce bacteriocins with different activity spectra? Future studies should aim to test natural populations for ecological predictors of broad-spectrum bacteriocin activity. For example, a recent study by Palmer and Foster (2022) has provided insight into this question by combining theoretical and comparative approaches to find that highly abundant strains are more likely to produce broad-spectrum toxins, as they can afford to make bacteriocins that broadly kills many competitors, unlike less abundant species that must focus their resources on targeting their main competitor (Palmer & Foster, 2022). Future studies could further test whether more abundant species are more likely to produce broad-spectrum bacteriocins by collecting abundance data of species and strains present within natural communities, in addition to the presence/absence data. Other ecological conditions have also been suggested to select for broad-spectrum bacteriocin activity. For example, under conditions where multiple species intensively competing for a common limiting resource, such as space or a specific nutrient in low concentration, despite being phylogenetically diverse (Heilbronner *et al.*, 2021). Future studies to use comparative approaches to test whether habitats containing particularly limiting common resources are more likely to contain more bacteriocin producers.

It is also possible that mechanistic constraints associated with certain taxonomic groups play a role in determining bacteriocin activity spectra. It has generally been observed that bacteriocins from gram-positive bacteria have a relatively broad-spectrum of activity compared to gram-negative species (Heilbronner *et al.,* 2021), however there are no clear ecological differences between the two groups to explain this. We can currently only speculate, but it has been suggested that some species are unable to target certain closely related competitors due to them being too phylogenetically similar to the focal competitor (Riley & Chavan, 2007; Hawlena *et al.,* 2010a, 2010b). Strains from many species, including many well-studied gram-negative species, can get around this by carrying a cognate immunity gene that protects them from self-harm caused by bacteriocin production (Riley & Gordon, 1999). However, gram-positive bacteria are often observed to achieve bacteriocin immunity through less discrete mechanisms that are often unidentifiable, such as through dual-function membrane transport proteins, other intrinsic cellular properties, or unknown genetic factors (Jack *et al.,* 1995; de Freire Bastos *et al.,* 2020). It is possible that this makes evolving a narrow-spectrum bacteriocin to target closely related individuals, which doesn't also cause self-harm, more challenging. Another strategy for avoiding self-harm is to target cellular receptors only carried by competing strains (Ghequire & De Mot, 2014; de Freire Bastos *et al.,* 2020). However, in gram-negative bacteria, many of these receptors are present on the outer membrane, which is gram-positives do not possess. It is possible that this fundamental biological difference in the cell envelope could also make it more difficult to evolve narrow-spectrum activity. It is, however, worth noting that intraspecific inhibitory activity been observed to be highly prevalent in some gram-positive species, such as *Lactococcus spp.* and other members of the lactic acid bacteria (LAB) (Klaenhammer, 1988; Cintas *et al.,* 2001; Mokoena, 2017), and *Streptomyces spp.* (Westhoff *et al.,* 2021). It would be interesting for future studies to consider how the mechanistic constraints associated with certain taxonomic groups could affect their ability to finely-tune bacteriocin activity spectra.

Exploitative competition in S. aureus

In addition to bacteriocins, the isolate collection of *S. aureus* from the human nasal cavity that I curated could be used to examine the roles of other social traits in determining the evolutionary and ecological success of *S. aureus*. In particular, it would be interesting to test the role of exploitative competition and its relative importance compared to interference competition such as bacteriocins. Despite adapting to nutrient limitation being identified as an important factor for *S. aureus* nasal colonisation success (Krismer *et al.*, 2014), we are yet to understand the extent exploitative competition mechanisms mediate intraspecific competition in *S. aureus*, and whether influences competitive strain dynamics. Future studies could test this by using Biolog nutrient plates (Biolog) tailored to contain nasally-relevant nutrients to quantify the level of resource use efficiency by nasal *S. aureus* strains. I would predict that strains with higher resource use efficiency would be more successful colonisers of the nasal cavity. Nutrient consumption data could also be used to determine the degree of metabolic overlap between strains, a known determinant of the degree of between-strain competition (Bruce *et al.*, 2017). A low degree of metabolic overlap could also explain why bacteriocin-mediated intraspecific inhibition is rare in *S. aureus.*

Studies could also test this by focusing on cooperative public goods for nutrient scavenging, such as siderophores (Kramer *et al.*, 2020). *S. aureus* has been identified to produce two main siderophores, staphyloferrins A (Beasley *et al.*, 2009) and B (Cheung *et al.*, 2009), and I would predict that, given iron is known to be an important limiting nutrient in the nasal cavity, strains that most effectively use staphyloferrins to uptake iron, for example by producing higher quantities or by using them more efficiency, are more likely to persistently colonise the nasal cavity. If staphyloferrin production was not associated with strain dominance, it would be interesting to further test whether cheat dynamics influence whether cheat dynamics contribute to a loss of staphyloferrin activity, as previously observed in studies of the siderophore pyoverdine in natural populations of *Pseudomonas aeruginosa* from the cystic fibrosis lung (Andersen *et al.,* 2015, 2018), and also in soil and freshwater environments (Bruce *et al.,* 2017; Butaitė *et al.,* 2017).

Bacteriocins: a medical perspective

Bacteriocins are receiving an increasing amount of attention from a medical perspective, within the developing field of 'Darwinian medicine', which aims to employ natural selection theory to design more effective and sustainable disease treatments (Williams & Nesse, 1991). Darwinian medicine can even be extended to 'Hamiltonian medicine', when treatments pertain to social traits (Brown *et al.,* 2009; Crespi *et al.,* 2014). Studies have identified bacteriocin activity to be associated with outbreaks of infection (Holt *et al.,* 2013; Quereda *et al.,* 2016). Understanding how pathogens utilise bacteriocins to gain and evolutionary and ecological benefit to outcompete competitors and cause disease is therefore of high clinical relevance.

There is also the potential to harness bacteriocins for use in our fight against pathogens, particularly those with the ability to evolve antibiotic resistance. Bacteriocins offer multiple potential advantages over the use of antibiotics, relating to having narrow and broad activity spectra, high stability, and multiple modes of implementation (Dobson *et al.,* 2012; Cotter *et al.,* 2013). Bacteriocins can either be implemented as therapeutic agents by isolating them for direct application, or by using probiotics *in situ*, whereby bacteriocin-producing strains are introduced to the microbiome to specifically target a known pathogen(s).

S. aureus is one example of such a pathogen, and the recent spread in resistance against mupirocin, the antibiotic used to decolonise patients from *S. aureus* before invasive surgery to reduce infection risk, is of particular clinical concern (Turner *et al.,* 2019). The effectiveness bacteriocin-based probiotics could also be improved by using social traits in synergy, for example by generating bacteriocin-producing strains that target the resident antibiotic resistant strain, that also act as selfish cheats for one or more public goods, known as 'trojan horses', to increase their ability to successfully invade the microbial community (Brown *et al.,* 2009). Further studies are required to understand how probiotics can be implemented effectively, sustainably, and safely. As a next step in understanding how probiotics could be effectively implemented to influence evolutionary strain dynamics, studies could perform long-term selection experiments, where probiotic strains with different strategies are added and removed from synthetic microbial communities, under environmentally-relevant conditions.

HGT & cooperation

In Chapter 4, I describe a study which uses a comparative approach to perform the most comprehensive genomic test of the 'cooperation hypothesis' to date, and found that contrary to predictions from theoretical models (Smith, 2001; Mc Ginty *et al.,* 2011, 2013), and laboratory experiments (Dimitriu *et al.,* 2014), HGT does not appear to consistently stabilise cooperation across bacteria.

It is important to note that while the aim of this study was to determine whether the predicted role of HGT in stabilising cooperation held true in bacterial genomes, a substantial portion of these genomes represent those of laboratory strains that may not resemble the genomes of the species in natural environments. To improve the power of our statistical analysis, we included as many species as possible in our analysis with equal to or more than ten genomes, irrespective of their environmental origin. However, as more genomes become available, especially in species where few genomes have currently been sequenced, future studies could repeat this analysis focusing on genomes isolates from natural environments.

Two key assumptions could also be modified in future studies to further test the role of HGT in stabilising cooperation, and to further understand the general relationship between HGT and bacterial sociality. Firstly, like previous studies we focused our study on plasmids, and did not test whether other types of mobile genetic elements (MGEs), such as bacteriophages and integrative conjugative elements (ICEs), could stabilise cooperation. While we found no effect of mobility on the location of genes coding for extracellular proteins, and models supporting the cooperation hypothesis typically assuming extremely high rates of HGT (Smith, 2001), it is possible that the biological features of other MGEs could allow for more effective cheat re-infection, and therefore a further study to confirm this result across other MGEs would be interesting and valuable.

Secondly, we analysed genes coding for extracellular proteins as a proxy for cooperative genes. This approach followed that used by previous studies to test in role of HGT in stabilising cooperation (Nogueira *et al.*, 2009) and was associated with a number of advantages: it provided a standardised definition for cooperative genes across bacteria due to the conserved nature of extracellular secretion signal peptides, it allowed us to analyse a large number of genes due to its rapid prediction speed, and it allowed us to analyse species whose cooperative traits have not been experimentally determined. However, there are some limitations with this approach. Firstly, while many extracellular proteins will provide cooperative benefits, some will not. This can occur if an extracellular protein has no effect on the fitness of a recipient, for example by forming part of the flagella to enable motility in the producer cell. It can also occur even if the extracellular protein does increase the fitness of the recipient, but was not selected for that fitness benefit. In this case, the extracellular protein provides a 'by-product', but not a cooperative, fitness benefit to the recipient (West *et al.*, 2006; West *et al.*, 2007c). Secondly, PSORTb will also miss many cooperative traits, as not all cooperative traits take the form of extracellular proteins, but rather other types of molecules. Many cooperative traits are produced by multiple genes that each directly code for intracellular proteins that interact to form the final molecule for secretion. These molecules represent some of the best studied cooperative traits, such as iron-scavenging siderophores (Kramer *et al.*, 2020). Future studies could overcome this limitation by using other genomic tools in addition to PSORTb, as demonstrated in a recent study of cooperative genes (Simonet & McNally, 2021). Many additional genomic tools are currently available to predict the presence, identity, location, and function of genes using genomic data, such as antiSMASH (Blin *et al.*, 2021), PANNZER (Koskinen *et al.*, 2015), and KofamScan (Aramaki *et al.*, 2020), and the development of new tools will only improve our ability to accurately identify genes coding for social traits in nature.

Sociomics

More broadly, in addition to the study of natural isolates in the laboratory, this study provides an example of an alternative approach, combining genome bioinformatics and comparative analysis, to test broad theoretical and laboratory evolutionary and ecological predictions about how bacteria behave *in situ.* Such approaches have been made possible by the recent 'omics' revolution, which has created exciting prospects for the field of sociomicrobiology, giving rise to the field of 'sociomics' (Ghoul *et al.*, 2017). However, our use of sociomics to study bacterial sociality to date has only scratched the surface. While Chapter 4, and other studies (Kümmerli *et al.,* 2014; Simonet & McNally, 2021; Palmer & Foster, 2022), demonstrate the value of using phylogenetic comparative methods to analyse genomic data to further understand bacterial sociality, other types of omics approaches can also provide insight into bacterial sociality in natural populations. For example, by using molecular population genetics approach, a recent study by Belcher *et al.* (2022) identified signatures of kin selection in multiple cooperative traits in natural populations of *Pseudomonas aeruginosa*. Studies have also used genomics to track social dynamics over time and infer evolutionary dynamics (Jiricny *et al.*, 2014; Ghoul *et al.*, 2015; Andersen *et al.,* 2015, 2018), transcriptomics and proteomics to show how microbial interactions can influence ecological community dynamics, such as by causing rapid micro-scale successions on model marine particles (Datta *et al.*, 2016), and metagenomics and 16S rRNA environmental sequencing to begin to understand the evolution and ecology of microbial species that are unculturable in lab conditions (Ventura *et al.*, 2009; Liu *et al.,* 2015; Coutinho *et al.*, 2018).

Concluding remark

Moving forward, using a combination of culture-based and omics approaches to test predictions in natural populations will provide insights that help refine the biological parameters of theoretical models and laboratory systems. In turn, this will allow the development of further testable hypotheses in natural populations. To date, the field of social evolution has excelled in using this feedback approach between theoretical and natural studies to gain scientific insight, and I advocate for future studies to apply the same approach in field of sociomicrobiology.

Bibliography

Adams, J., Kinney, T., Thompson, S., Rubin, L. and Helling, R.B., 1979. Frequency-dependent selection for plasmid-containing cells of *Escherichia coli*. *Genetics*, *91*(4), pp.627-637.

Aleti, G., Baker, J.L., Tang, X., Alvarez, R., Dinis, M., Tran, N.C., Melnik, A.V., Zhong, C., Ernst, M., Dorrestein, P.C. and Edlund, A., 2019. Identification of the bacterial biosynthetic gene clusters of the oral microbiome illuminates the unexplored social language of bacteria during health and disease. *MBio*, *10*(2), pp.e00321-19.

Andersen, P.S., Larsen, L.A., Fowler, V.G., Stegger, M., Skov, R.L. and Christensen, K., 2013. Risk factors for *Staphylococcus aureus* nasal colonization in Danish middle-aged and elderly twins. *European Journal of Clinical Microbiology & Infectious Diseases*, *32*(10), pp.1321-1326.

Andersen, S.B., Ghoul, M., Griffin, A.S., Petersen, B., Johansen, H.K. and Molin, S., 2017. Diversity, prevalence, and longitudinal occurrence of type II toxin-antitoxin systems of *Pseudomonas aeruginosa* infecting cystic fibrosis lungs. *Frontiers in microbiology*, *8*, p.1180.

Andersen, S.B., Ghoul, M., Marvig, R.L., Lee, Z.B., Molin, S., Johansen, H.K. and Griffin, A.S., 2018. Privatisation rescues function following loss of cooperation. *Elife*, *7*, p.e38594.

Andersen, S.B., Marvig, R.L., Molin, S., Krogh Johansen, H. and Griffin, A.S., 2015. Long-term social dynamics drive loss of function in pathogenic bacteria. *Proceedings of the National Academy of Sciences*, *112*(34), pp.10756-10761.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S. and Ogata, H., 2020. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, *36*(7), pp.2251-2252.

Abrudan, M.I., Smakman, F., Grimbergen, A.J., Westhoff, S., Miller, E.L., Van Wezel, G.P. and Rozen, D.E., 2015. Socially mediated induction and suppression of antibiosis during bacterial coexistence. *Proceedings of the National Academy of Sciences*, *112*(35), pp.11054-11059.

Bakkeren, E., Gül, E., Huisman, J.S., Steiger, Y., Rocker, A., Hardt, W.D. and Diard, M., 2021. Cooperative virulence can emerge via horizontal gene transfer but is stabilized by transmission. *bioRxiv*, pp.2021-02.

Basler, Á., Pilhofer, Á., Henderson, G.P., Jensen, G.J. and Mekalanos, J., 2012. Type VI secretion requires a dynamic contractile phage tail-like structure. *Nature*, *483*(7388), pp.182-186.

Bastos, M.C.F., Ceotto, H., Coelho, M.L.V. and Nascimento, J.S., 2009. Staphylococcal antimicrobial peptides: relevant properties and potential biotechnological applications. *Current pharmaceutical biotechnology*, *10*(1), pp.38-61.

Beasley, F.C., Vinés, E.D., Grigg, J.C., Zheng, Q., Liu, S., Lajoie, G.A., Murphy, M.E. and Heinrichs, D.E., 2009. Characterization of staphyloferrin A biosynthetic and transport mutants in *Staphylococcus aureus*. *Molecular microbiology*, *72*(4), pp.947-963.

Belcher, L.J., Dewar, A.E., Ghoul, M. and West, S.A., 2022. Kin selection for cooperation in natural bacterial populations. *Proceedings of the National Academy of Sciences*, *119*(9), p.e2119070119.

Bertels, F., Silander, O.K., Pachkov, M., Rainey, P.B. and Van Nimwegen, E., 2014. Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Molecular biology and evolution*, *31*(5), pp.1077-1088.

Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., Van Wezel, G.P., Medema, M.H. and Weber, T., 2021. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research*, *49*(W1), pp.W29-W35.

Bourke, A.F., 2011. *Principles of social evolution*. Oxford University Press.

Branche Jr, W.C., Young, V.M., Robinet, H.G. and Massey, E.D., 1963. Effect of colicine production on *Escherichia coli* in the normal human intestine. *Proceedings of the Society for Experimental Biology and Medicine*, *114*(1), pp.198-201.

Brown, S.P. and Johnstone, R.A., 2001. Cooperation in the dark: signalling and collective action in quorum-sensing bacteria. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *268*(1470), pp.961-965.

Brown, A.F., Leech, J.M., Rogers, T.R. and McLoughlin, R.M., 2014. *Staphylococcus aureus* colonization: modulation of host immune response and impact on human vaccine design. *Frontiers in immunology*, *4*, p.507.

Brown, S.P., West, S.A., Diggle, S.P. and Griffin, A.S., 2009. Social evolution in micro-organisms and a Trojan horse approach to medical intervention strategies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1533), pp.3157-3168.

Bruce, J.B., West, S.A. and Griffin, A.S., 2017. Bacteriocins and the assembly of natural *Pseudomonas* fluorescens populations. *Journal of evolutionary biology*, *30*(2), pp.352-360.

Burrus, V. and Waldor, M.K., 2004. Shaping bacterial genomes with integrative and conjugative elements. *Research in microbiology*, *155*(5), pp.376-386.

Butaitė, E., Baumgartner, M., Wyder, S. and Kümmerli, R., 2017. Siderophore cheating and cheating resistance shape competition for iron in soil and freshwater *Pseudomonas* communities. *Nature communications*, *8*(1), pp.1-12.

Butaitė, E., Kramer, J., Wyder, S. and Kümmerli, R., 2018. Environmental determinants of pyoverdine production, exploitation and competition in natural *Pseudomonas* communities. *Environmental microbiology*, *20*(10), pp.3629-3642.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.L. and Brüssow, H., 2003. Phage as agents of lateral gene transfer. *Current opinion in microbiology*, *6*(4), pp.417-424.

Cao, Z., Casabona, M.G., Kneuper, H., Chalmers, J.D. and Palmer, T., 2016. The type VII secretion system of *Staphylococcus aureus* secretes a nuclease toxin that targets competitor bacteria. *Nature microbiology*, *2*(1), pp.1-11.

Carrier, T., Jones, K.L. and Keasling, J.D., 1998. mRNA stability and plasmid copy number effects on gene expression from an inducible promoter system. *Biotechnology and bioengineering*, *59*(6), pp.666-672.

Ceotto, H., dos Santos Nascimento, J., de Paiva Brito, M.A.V. and de Freire Bastos, M.D.C., 2009. Bacteriocin production by *Staphylococcus aureus* involved in bovine mastitis in Brazil. *Research in microbiology*, *160*(8), pp.592-599.

Chao, L. and Levin, B.R., 1981. Structured habitats and the evolution of anticompetitor toxins in bacteria. *Proceedings of the National Academy of Sciences*, *78*(10), pp.6324-6328.

Cheung, J., Beasley, F.C., Liu, S., Lajoie, G.A. and Heinrichs, D.E., 2009. Molecular characterization of staphyloferrin B biosynthesis in *Staphylococcus aureus*. *Molecular microbiology*, *74*(3), pp.594-608.

Cidem, A., Bradbury, P., Traini, D. and Ong, H.X., 2020. Modifying and Integrating in vitro and ex vivo Respiratory Models for Inhalation Drug Screening. *Frontiers in bioengineering and biotechnology*, *8*, p.581995.

Cintas, L.M., Casaus, M.P., Herranz, C., Nes, I.F. and Hernández, P.E., 2001. Bacteriocins of lactic acid bacteria. *Food Science and Technology International*, *7*(4), pp.281-305.

Ciofu, O., Beveridge, T.J., Kadurugamuwa, J., Walther-Rasmussen, J. and Høiby, N., 2000. Chromosomal β-lactamase is packaged into membrane vesicles and secreted from *Pseudomonas aeruginosa*. *Journal of Antimicrobial Chemotherapy*, *45*(1), pp.9-13.

Clutton-Brock, T.H. and Harvey, P.H., 1977. Primate ecology and social organization. *Journal of zoology*, *183*(1), pp.1-39.

Coelho, M.L.V., dos Santos Nascimento, J., Fagundes, P.C., Madureira, D.J., de Oliveira, S.S., de Paiva Brito, M.A.V. and de Freire Bastos, M.D.C., 2007. Activity of staphylococcal bacteriocins against *Staphylococcus aureus* and *Streptococcus agalactiae* involved in bovine mastitis. *Research in microbiology*, *158*(7), pp.625-630.

Cohen, J., 2013. *Statistical power analysis for the behavioral sciences*. Routledge.

Cohen, O., Gophna, U. and Pupko, T., 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Molecular biology and evolution*, *28*(4), pp.1481-1489.

Conwill, A., Kuan, A.C., Damerla, R., Poret, A.J., Baker, J.S., Tripp, A.D., Alm, E.J. and Lieberman, T.D., 2022. Anatomy promotes neutral coexistence of strains in the human skin microbiome. *Cell Host & Microbe*, *30*(2), pp.171-182.

Cordero, O.X., Wildschutte, H., Kirkup, B., Proehl, S., Ngo, L., Hussain, F., Le Roux, F., Mincer, T. and Polz, M.F., 2012. Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. *Science*, *337*(6099), pp.1228-1231.

Cornelis, G.R., Boland, A., Boyd, A.P., Geuijen, C., Iriarte, M., Neyt, C., Sory, M.P. and Stainier, I., 1998. The virulence plasmid of *Yersinia*, an antihost genome. *Microbiology and Molecular Biology Reviews*, *62*(4), pp.1315-1352.

Cornforth, D.M. and Foster, K.R., 2013. Competition sensing: the social side of bacterial stress responses. *Nature Reviews Microbiology*, *11*(4), pp.285-293.

Cornforth, D.M. and Foster, K.R., 2015. Antibiotics and the art of bacterial war. *Proceedings of the National Academy of Sciences*, *112*(35), pp.10827-10828.

Cotter, P.D., Ross, R.P. and Hill, C., 2013. Bacteriocins - a viable alternative to antibiotics? *Nature Reviews Microbiology*, *11*(2), pp.95-105.

Coutinho, F.H., Gregoracci, G.B., Walter, J.M., Thompson, C.C. and Thompson, F.L., 2018. Metagenomics sheds light on the ecology of marine microbes and their viruses. *Trends in microbiology*, *26*(11), pp.955-965.

Coyne, M.J., Béchon, N., Matano, L.M., McEneany, V.L., Chatzidaki-Livanis, M. and Comstock, L.E., 2019. A family of anti-Bacteroidales peptide toxins wide-spread in the human gut microbiota. *Nature Communications*, *10*(1), p.3460.

Crawley, M.J., 2005. *Statistics: an introduction using R* (Vol. 327). Wiley.

Crespi, B., Foster, K. and Úbeda, F., 2014. First principles of Hamiltonian medicine. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1642), p.20130366.

Daly, K.M., Upton, M., Sandiford, S.K., Draper, L.A., Wescombe, P.A., Jack, R.W., O'Connor, P.M., Rossney, A., Götz, F., Hill, C. and Cotter, P.D., 2010. Production of the Bsa lantibiotic by community-acquired *Staphylococcus aureus* strains. *Journal of bacteriology*, *192*(4), pp.1131-1142.

Das, S., Lindemann, C., Young, B.C., Muller, J., Österreich, B., Ternette, N., Winkler, A.C., Paprotka, K., Reinhardt, R., Förstner, K.U. and Allen, E., 2016. Natural mutations in a *Staphylococcus aureus* virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation. *Proceedings of the National Academy of Sciences*, *113*(22), pp.E3101-E3110.

Datta, M.S., Sliwerska, E., Gore, J., Polz, M.F. and Cordero, O.X., 2016. Microbial interactions lead to rapid micro-scale successions on model marine particles. *Nature communications*, *7*(1), p.11965.

Davies, N.B., Krebs, J.R. and West, S.A., 2012. *An introduction to behavioural ecology*. John Wiley & Sons.

de Borja Callejas, F., Martinez-Anton, A., Alobid, I., Fuentes, M., Cortijo, J., Picado, C., Roca-Ferrer, J. and Mullol, J., 2014. Reconstituted human upper airway epithelium as 3-d in vitro model for nasal polyposis. *PLoS One*, *9*(6), p.e100537.

de Freire Bastos, M.D.C., Miceli de Farias, F., Carlin Fagundes, P. and Varella Coelho, M.L., 2020. Staphylococcins: An update on antimicrobial peptides produced by staphylococci and their diverse potential applications. *Applied Microbiology and Biotechnology*, *104*(24), pp.10339-10368.

de Oliveira, S.S., Abrantes, J., Cardoso, M., Sordelli, D. and Bastos, M.C.F., 1998. Staphylococcal strains involved in bovine mastitis are inhibited by *Staphylococcus aureus* antimicrobial peptides. *Letters in Applied Microbiology*, *27*(5), pp.287-291.

Di Venanzio, G., Moon, K.H., Weber, B.S., Lopez, J., Ly, P.M., Potter, R.F., Dantas, G. and Feldman, M.F., 2019. Multidrug-resistant plasmids repress chromosomally encoded T6SS to enable their dissemination. *Proceedings of the National Academy of Sciences*, *116*(4), pp.1378-1383.

Dietel, A.K., Kaltenpoth, M. and Kost, C., 2018. Convergent evolution in intracellular elements: plasmids as model endosymbionts. *Trends in microbiology*, *26*(9), pp.755-768.

Diggle, S.P., Griffin, A.S., Campbell, G.S. and West, S.A., 2007. Cooperation and conflict in quorum-sensing bacterial populations. *Nature*, *450*(7168), pp.411-414.

Dimitriu, T., Lotton, C., Bénard-Capelle, J., Misevic, D., Brown, S.P., Lindner, A.B. and Taddei, F., 2014. Genetic information transfer promotes cooperation in bacteria. *Proceedings of the National Academy of Sciences*, *111*(30), pp.11103-11108.

Ding, W., Baumdicker, F. and Neher, R.A., 2018. panX: pan-genome analysis and exploration. *Nucleic acids research*, *46*(1), pp.e5-e5.

Dinges, M.M., Orwin, P.M. and Schlievert, P.M., 2000. Exotoxins of *Staphylococcus aureus*. *Clinical microbiology reviews*, *13*(1), pp.16-34.

Dobson, A., Cotter, P.D., Ross, R.P. and Hill, C., 2012. Bacteriocin production: a probiotic trait? *Applied and environmental microbiology*, *78*(1), pp.1-6.

Donia, M.S., Cimermancic, P., Schulze, C.J., Brown, L.C.W., Martin, J., Mitreva, M., Clardy, J., Linington, R.G. and Fischbach, M.A., 2014. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell*, *158*(6), pp.1402-1414.

Dorit, R.L., Roy, S.M. and Riley, M.A., 2016. *The Bacteriocins*. Caister Academic Press.

Dorosky, R.J., Yu, J.M., Pierson III, L.S. and Pierson, E.A., 2017. *Pseudomonas chlororaphis* produces two distinct R-tailocins that contribute to bacterial competition in biofilms and on roots. *Applied and environmental microbiology*, *83*(15), pp.e00706-17.

Dragoš, A., Kiesewalter, H., Martin, M., Hsu, C.Y., Hartmann, R., Wechsler, T., Eriksen, C., Brix, S., Drescher, K., Stanley-Wall, N. and Kümmerli, R., 2018. Division of labor during biofilm matrix production. *Current Biology*, *28*(12), pp.1903-1913.

Drider, D. and Rebuffat, S. eds., 2011. *Prokaryotic antimicrobial peptides: from genes to applications*. Springer Science & Business Media.

Durack, J. and Lynch, S.V., 2019. The gut microbiome: relationships with disease and opportunities for therapy. *Journal of Experimental Medicine*, *216*(1), pp.20-40.

Everitt, R.G., Didelot, X., Batty, E.M., Miller, R.R., Knox, K., Young, B.C., Bowden, R., Auton, A., Votintseva, A., Larner-Svensson, H. and Charlesworth, J., 2014. Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*. *Nature communications*, *5*(1), pp.1-9.

Fagundes, P.C., de Sousa Santos, I.N., Francisco, M.S., Albano, R.M. and de Freire Bastos, M.D.C., 2017. Genetic and biochemical characterization of hyicin 3682, the first bacteriocin reported for *Staphylococcus hyicus*. *Microbiological research*, *198*, pp.36-46.

Fagundes, P.C., Farias, F.M., Santos, O.C.S., de Oliveira, N.E.M., da Paz, J.A.S., Ceotto-Vigoder, H., Alviano, D.S., Romanos, M.T.V. and Bastos, M.C.F., 2016. The antimicrobial peptide aureocin A53 as an alternative agent for biopreservation of dairy products. *Journal of Applied Microbiology*, *121*(2), pp.435-444.

Felsenstein, J., 1985. Phylogenies and the comparative method. *The American Naturalist*, *125*(1), pp.1-15.

Ference, C.M., Gochez, A.M., Behlau, F., Wang, N., Graham, J.H. and Jones, J.B., 2018. Recent advances in the understanding of *Xanthomonas citri ssp. citri* pathogenesis and citrus canker disease management. *Molecular plant pathology*, *19*(6), p.1302.

Fernández-Fernández, R., Lozano, C., Eguizábal, P., Ruiz-Ripa, L., Martínez-Álvarez, S., Abdullahi, I.N., Zarazaga, M. and Torres, C.T., 2022. Bacteriocin-Like Inhibitory Substances in Staphylococci of Different Origins and Species With Activity Against Relevant Pathogens. *Frontiers in Microbiology*, p.1152.

Figueiredo, A.R., Özkaya, Ö., Kümmerli, R. and Kramer, J., 2022. Siderophores drive invasion dynamics in bacterial communities through their dual role as public good versus public bad. *Ecology Letters*, *25*(1), pp.138-150.

Foster, K.R., 2010. Social behaviour in microorganisms. *Social behaviour: genes, ecology and evolution*, pp.331-56.

Foster, K.R. and Bell, T., 2012. Competition, not cooperation, dominates interactions among culturable microbial species. *Current biology*, *22*(19), pp.1845-1850.

Foster, K.R., Shaulsky, G., Strassmann, J.E., Queller, D.C. and Thompson, C.R., 2004. Pleiotropy as a mechanism to stabilize cooperation. *Nature*, *431*(7009), pp.693-696.

Foster, K.R. and Wenseleers, T., 2006. A general model for the evolution of mutualisms. *Journal of evolutionary biology*, *19*(4), pp.1283-1293.

Frank, S.A., 1994. Spatial polymorphism of bacteriocins and other allelopathic traits. *Evolutionary Ecology*, *8*, pp.369-386.

Frank, D.N., Feazel, L.M., Bessesen, M.T., Price, C.S., Janoff, E.N. and Pace, N.R., 2010. The human nasal microbiota and *Staphylococcus aureus* carriage. *PloS one*, *5*(5), p.e10598.

Fritz, S.A. and Purvis, A., 2010. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, *24*(4), pp.1042-1051.

Gamon, M.R., Cristina Moreira, E., de Oliveira, S.S., Teixeira, L.M. and do Carmo de Freire Bastos, M., 1999. Characterization of a novel bacteriocin-encoding plasmid found in clinical isolates of *Staphylococcus aureus*. *Antonie van Leeuwenhoek*, *75*, pp.233-243.

García-Bayona, L. and Comstock, L.E., 2018. Bacterial antagonism in host-associated microbial communities. *Science*, *361*(6408), p.eaat2456.

Garcia-Garcera, M. and Rocha, E.P., 2020. Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. *Nature Communications*, *11*(1), p.758.

Garcia-Garcera, M., Touchon, M., Brisse, S. and Rocha, E.P., 2017. Metagenomic assessment of the interplay between the environment and the genetic diversification of Acinetobacter. *Environmental microbiology*, *19*(12), pp.5010-5024.

Gardner, A., West, S.A. and Buckling, A., 2004. Bacteriocins, spite and virulence. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(1547), pp.1529-1535.

Gardy, J.L. and Brinkman, F.S., 2006. Methods for predicting bacterial protein subcellular localization. *Nature Reviews Microbiology*, *4*(10), pp.741-751.

Geoghegan, J.A. and Foster, T.J., 2017. Cell wall-anchored surface proteins of *Staphylococcus aureus*: many proteins, multiple functions. *Staphylococcus aureus: Microbiology, Pathology, Immunology, Therapy and Prophylaxis*, pp.95-120.

Ghequire, M.G. and De Mot, R., 2014. Ribosomally encoded antibacterial proteins and peptides from *Pseudomonas*. *FEMS microbiology reviews*, *38*(4), pp.523-568.

Ghigo, J.M., 2001. Natural conjugative plasmids induce bacterial biofilm development. *Nature*, *412*(6845), pp.442-445.

Ghoul, M., Andersen, S.B. and West, S.A., 2017. Sociomics: Using omic approaches to understand social evolution. *Trends in Genetics*, *33*(6), pp.408-419.

Ghoul, M., Griffin, A.S. and West, S.A., 2014. Toward an evolutionary definition of cheating. *Evolution*, *68*(2), pp.318-331.

Ghoul, M. and Mitri, S., 2016. The ecology and evolution of microbial competition. *Trends in microbiology*, *24*(10), pp.833-845.

Ghoul, M., West, S.A., Johansen, H.K., Molin, S., Harrison, O.B., Maiden, M.C., Jelsbak, L., Bruce, J.B. and Griffin, A.S., 2015. Bacteriocin-mediated competition in cystic fibrosis lung infections. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1814), p.20150972.

Giambiagi-Marval, M., Mafra, M.A., Penido, E.G.C. and Bastos, M.C.F., 1990. Distinct groups of plasmids correlated with bacteriocin production in *Staphylococcus aureus*. *Microbiology*, *136*(8), pp.1591-1599.

Golubchik, T., Batty, E.M., Miller, R.R., Farr, H., Young, B.C., Larner-Svensson, H., Fung, R., Godwin, H., Knox, K., Votintseva, A. and Everitt, R.G., 2013. Within-host evolution of *Staphylococcus aureus* during asymptomatic carriage. *PloS one*, *8*(5), p.e61319.

Gordon, D.M., 2016. *The Natural History of Bacteriocins* (pp. 1-10). Caister Academic Press: Norfolk, UK.

Gordon, D.M. and O'Brien, C.L., 2006. Bacteriocin diversity and the frequency of multiple bacteriocin production in *Escherichia coli*. *Microbiology*, *152*(11), pp.3239-3244.

Gordon, N.C., Pichon, B., Golubchik, T., Wilson, D.J., Paul, J., Blanc, D.S., Cole, K., Collins, J., Cortes, N., Cubbon, M. and Gould, F.K., 2017. Whole-genome sequencing reveals the contribution of long-term carriers in *Staphylococcus aureus* outbreak investigation. *Journal of clinical microbiology*, *55*(7), pp.2188-2197.

Gordon, D.M. and Riley, M.A., 1999. A theoretical and empirical investigation of the invasion dynamics of colicinogeny. *Microbiology*, *145*(3), pp.655-661.

Gordon, D.M., Riley, M.A. and Pinou, T., 1998. Temporal changes in the frequency of colicinogeny in *Escherichia coli* from house mice. *Microbiology*, *144*(8), pp.2233-2240.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. and Lopez, R., 2010. A new bioinformatics analysis tools framework at EMBL–EBI. *Nucleic acids research*, *38*(suppl_2), pp.W695-W699.

Grafen, A., 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *326*(1233), pp.119-157.

Granato, E.T., Meiller-Legrand, T.A. and Foster, K.R., 2019. The evolution and ecology of bacterial warfare. *Current biology*, *29*(11), pp.R521-R537.

Griffin, A.S., West, S.A. and Buckling, A., 2004. Cooperation and competition in pathogenic bacteria. *Nature*, *430*(7003), pp.1024-1027.

Griffith, F., 1928. The significance of pneumococcal types. *Epidemiology & Infection*, *27*(2), pp.113-159.

Gu, S., Wei, Z., Shao, Z., Friman, V.P., Cao, K., Yang, T., Kramer, J., Wang, X., Li, M., Mei, X. and Xu, Y., 2020. Competition for iron drives phytopathogen control by natural rhizosphere microbiomes. *Nature Microbiology*, *5*(8), pp.1002-1010.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, *59*(3), pp.307-321.

Gupta, A., Kapil, R., Dhakan, D.B. and Sharma, V.K., 2014. MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PloS one*, *9*(4), p.e93907.

Haaber, J., Leisner, J.J., Cohn, M.T., Catalan-Moreno, A., Nielsen, J.B., Westh, H., Penadés, J.R. and Ingmer, H., 2016. Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells. *Nature communications*, *7*(1), pp.1-8.

Hadfield, J.D., 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of statistical software, 33*, pp.1-22.

Hadfield, J.D., 2019. MCMCglmm course notes. https://cran. r-project. org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf

Hale, T.L., 1991. Genetic basis of virulence in *Shigella* species. *Microbiological reviews*, *55*(2), pp.206-224.

Hall, R.J., Whelan, F.J., McInerney, J.O., Ou, Y. and Domingo-Sananes, M.R., 2020. Horizontal gene transfer as a source of conflict and cooperation in prokaryotes. *Frontiers in Microbiology*, *11*, p.1569.

Hamilton, W.D., 1963. The evolution of altruistic behavior. *The American Naturalist*, *97*(896), pp.354-356.

Hamilton, W.D., 1964. The genetical evolution of social behaviour. II. *Journal of theoretical biology*, *7*(1), pp.17-52.

Harmsen, D., Claus, H., Witte, W., Rothganger, J., Claus, H., Turnwald, D. and Vogel, U., 2003. Typing of methicillin-resistant *Staphylococcus aureus* in a university hospital setting by using novel software for *spa* repeat determination and database management. *Journal of clinical microbiology*, *41*(12), pp.5442-5448.

Harvey, P.H. and Pagel, M.D., 1991. *The comparative method in evolutionary biology* (Vol. 239). Oxford: Oxford university press.

Hawlena, H., Bashey, F. and Lively, C.M., 2010a. The evolution of spite: Population structure and bacteriocin-mediated antagonism in two natural populations of *Xenorhabdus bacteria*. *Evolution*, *64*(11), pp.3198-3204.

Hawlena, H., Bashey, F. and Lively, C.M., 2012. Bacteriocin-mediated interactions within and between coexisting species. *Ecology and Evolution*, *2*(10), pp.2521-2526.

Hawlena, H., Bashey, F., Mendes-Soares, H. and Lively, C.M., 2010b. Spiteful interactions in a natural population of the bacterium *Xenorhabdus bovienii*. *The American Naturalist*, *175*(3), pp.374-381.

Health Protection Agency, 2007. Identification of *Staphylococcus* species, *Micrococcus* species and *Rothia* species. National Standard Method BSOP ID 7.

Heilbronner, S., Krismer, B., Brötz-Oesterhelt, H. and Peschel, A., 2021. The microbiome-shaping roles of bacteriocins. *Nature Reviews Microbiology*, *19*(11), pp.726-739.

Hill, C., 1999. Developing applications for lactococcal bacteriocins. In *Lactic Acid Bacteria: Genetics, Metabolism and Applications: Proceedings of the Sixth Symposium on lactic acid bacteria: genetics, metabolism and applications, 19–23 September 1999, Veldhoven, The Netherlands* (pp. 337-346). Springer Netherlands.

Holt, K.E., Thieu Nga, T.V., Thanh, D.P., Vinh, H., Kim, D.W., Vu Tra, M.P., Campbell, J.I., Hoang, N.V.M., Vinh, N.T., Minh, P.V. and Thuy, C.T., 2013. Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proceedings of the National Academy of Sciences*, *110*(43), pp.17522-17527.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K. and Suzuki, Y., 2016. A new view of the tree of life. *Nature microbiology*, *1*(5), pp.1-6.

Hurlbert, S.H., 1984. Pseudoreplication and the design of ecological field experiments. *Ecological monographs*, *54*(2), pp.187-211.

Ives, A.R., Midford, P.E. and Garland Jr, T., 2007. Within-species variation and measurement error in phylogenetic comparative methods. *Systematic biology*, *56*(2), pp.252-270.

Ives, A.R. and Zhu, J., 2006. Statistics for correlated data: phylogenies, space, and time. *Ecological applications*, *16*(1), pp.20-32.

Iwase, T., Uehara, Y., Shinji, H., Tajima, A., Seo, H., Takada, K., Agata, T. and Mizunoe, Y., 2010. *Staphylococcus epidermidis* Esp inhibits *Staphylococcus aureus* biofilm formation and nasal colonization. *Nature*, *465*(7296), pp.346-349.

Jain, R., Rivera, M.C. and Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences*, *96*(7), pp.3801-3806.

Janek, D., Zipperer, A., Kulik, A., Krismer, B. and Peschel, A., 2016. High frequency and diversity of antimicrobial activities produced by nasal *Staphylococcus* strains against bacterial competitors. *PLoS pathogens*, *12*(8), p.e1005812.

Jennions, M.D. and Møller, A.P., 2003. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology*, *14*(3), pp.438-445.

Jiricny, N., Molin, S., Foster, K., Diggle, S.P., Scanlan, P.D., Ghoul, M., Johansen, H.K., Santorelli, L.A., Popat, R., West, S.A. and Griffin, A.S., 2014. Loss of social behaviours in populations of *Pseudomonas aeruginosa* infecting lungs of patients with cystic fibrosis. *PloS one*, *9*(1), p.e83124.

Jolley, K.A., Bliss, C.M., Bennett, J.S., Bratcher, H.B., Brehony, C., Colles, F.M., Wimalarathna, H., Harrison, O.B., Sheppard, S.K., Cody, A.J. and Maiden, M.C., 2012. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*, *158*(Pt 4), p.1005.

Jolley, K., Bray, J. and Maiden, M.C.J.J., 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST. org website and their applications. *Wellcome open research*, *3*(124).

Jones, S., Yu, B., Bainton, N.A., Birdsall, M., Bycroft, B.W., Chhabra, S.R., Cox, A.J., Golby, P., Reeves, P.J. and Stephens, S., 1993. The lux autoinducer regulates the production of exoenzyme virulence determinants in *Erwinia carotovora* and *Pseudomonas aeruginosa*. *The EMBO journal*, *12*(6), pp.2477-2482.

Kawada-Matsuo, M., Shammi, F., Oogai, Y., Nakamura, N., Sugai, M. and Komatsuzawa, H., 2016. C55 bacteriocin produced by ETB-plasmid positive *Staphylococcus aureus* strains is a key factor for competition with *S. aureus* strains. *Microbiology and immunology*, *60*(3), pp.139-147.

Kerr, B., Riley, M.A., Feldman, M.W. and Bohannan, B.J., 2002. Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. *Nature*, *418*(6894), pp.171-174.

Kinkel, L.L., Schlatter, D.C., Xiao, K. and Baines, A.D., 2014. Sympatric inhibition and niche differentiation suggest alternative coevolutionary trajectories among Streptomycetes. *The ISME journal*, *8*(2), pp.249-256.

Kirkup, B.C. and Riley, M.A., 2004. Antibiotic-mediated antagonism leads to a bacterial game of rock–paper–scissors in vivo. *Nature*, *428*(6981), pp.412-414.

Klaenhammer, T.R., 1988. Bacteriocins of lactic acid bacteria. *Biochimie*, *70*(3), pp.337-349.

Klaenhammer, T.R., 1993. Genetics of bacteriocins produced by lactic acid bacteria. *FEMS microbiology reviews*, *12*(1-3), pp.39-85.

Koch, G., Yepes, A., Förstner, K.U., Wermser, C., Stengel, S.T., Modamio, J., Ohlsen, K., Foster, K.R. and Lopez, D., 2014. Evolution of resistance to a last-resort antibiotic in *Staphylococcus aureus* via bacterial competition. *Cell*, *158*(5), pp.1060-1071.

Kommineni, S., Bretl, D.J., Lam, V., Chakraborty, R., Hayward, M., Simpson, P., Cao, Y., Bousounis, P., Kristich, C.J. and Salzman, N.H., 2015. Bacteriocin production augments niche competition by enterococci in the mammalian gastrointestinal tract. *Nature*, *526*(7575), pp.719-722.

Koreen, L., Ramaswamy, S.V., Graviss, E.A., Naidich, S., Musser, J.M. and Kreiswirth, B.N., 2004. *spa* typing method for discriminating among *Staphylococcus aureus* isolates: implications for use of a single marker to detect genetic micro-and macrovariation. *Journal of clinical microbiology*, *42*(2), pp.792-799.

Koskinen, P., Törönen, P., Nokso-Koivisto, J. and Holm, L., 2015. PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics*, *31*(10), pp.1544-1552.

Köstlbacher, S., Collingro, A., Halter, T., Domman, D. and Horn, M., 2021. Coevolving plasmids drive gene flow and genome plasticity in host-associated intracellular bacteria. *Current Biology*, *31*(2), pp.346-357.

Kraemer, S.A., Soucy, J.P.R. and Kassen, R., 2017. Antagonistic interactions of soil pseudomonads are structured in time. *FEMS Microbiology Ecology*, *93*(5), p.fix046.

Kramer, J., Özkaya, Ö. and Kümmerli, R., 2020. Bacterial siderophores in community and host interactions. *Nature Reviews Microbiology*, *18*(3), pp.152-163.

Krishna Kumar, R. and Foster, K.R., 2022. 3D printing of microbial communities: A new platform for understanding and engineering microbiomes. *Microbial Biotechnology*.

Krishna Kumar, R., Meiller-Legrand, T.A., Alcinesio, A., Gonzalez, D., Mavridou, D.A., Meacock, O.J., Smith, W.P., Zhou, L., Kim, W., Pulcu, G.S. and Bayley, H., 2021. Droplet printing reveals the importance of micron-scale structure for bacterial ecology. *Nature communications*, *12*(1), p.857.

Krismer, B., Liebeke, M., Janek, D., Nega, M., Rautenberg, M., Hornig, G., Unger, C., Weidenmaier, C., Lalk, M. and Peschel, A., 2014. Nutrient limitation governs *Staphylococcus aureus* metabolism and niche adaptation in the human nose. *PLoS pathogens*, *10*(1), p.e1003862.

Krismer, B., Weidenmaier, C., Zipperer, A. and Peschel, A., 2017. The commensal lifestyle of *Staphylococcus aureus* and its interactions with the nasal microbiota. *Nature reviews microbiology*, *15*(11), pp.675-687.

Kruskal, W., 1988. Miracles and statistics: The casual assumption of independence. *Journal of the American Statistical Association*, *83*(404), pp.929-940.

Kümmerli, R., 2022. Iron acquisition strategies in pseudomonads: mechanisms, ecology, and evolution. *BioMetals*, pp.1-21.

Kümmerli, R., Gardner, A., West, S.A. and Griffin, A.S., 2009a. Limited dispersal, budding dispersal, and cooperation: an experimental study. *Evolution*, *63*(4), pp.939-949.

Kümmerli, R., Griffin, A.S., West, S.A., Buckling, A. and Harrison, F., 2009b. Viscous medium promotes cooperation in the pathogenic bacterium *Pseudomonas aeruginosa*. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1672), pp.3531-3538.

Kümmerli, R., Santorelli, L.A., Granato, E.T., Dumas, Z., Dobay, A., Griffin, A.S. and West, S.A., 2015. Co-evolutionary dynamics between public good producers and cheats in the bacterium *Pseudomonas aeruginosa*. *Journal of evolutionary biology*, *28*(12), pp.2264-2274.

Kümmerli, R., Schiessl, K.T., Waldvogel, T., McNeill, K. and Ackermann, M., 2014. Habitat structure and the evolution of diffusible siderophores in bacteria. *Ecology letters*, *17*(12), pp.1536-1544.

Laabei, M., Uhlemann, A.C., Lowy, F.D., Austin, E.D., Yokoyama, M., Ouadi, K., Feil, E., Thorpe, H.A., Williams, B., Perkins, M. and Peacock, S.J., 2015. Evolutionary trade-offs underlie the multi-faceted virulence of *Staphylococcus aureus*. *PLoS biology*, *13*(9), p.e1002229.

Laird, N.M., 1988. Missing data in longitudinal studies. *Statistics in medicine*, *7*(1-2), pp.305-315.

Langmead, B. and Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), pp.357-359.

Lederberg, J. and Tatum, E.L., 1946. Gene recombination in *Escherichia coli. Nature,* 158(4016), p.558.

Lee, A.S., De Lencastre, H., Garau, J., Kluytmans, J., Malhotra-Kumar, S., Peschel, A. and Harbarth, S., 2018. Methicillin-resistant *Staphylococcus aureus*. *Nature reviews Disease primers*, *4*(1), pp.1-23.

Lehtinen, S., Croucher, N.J., Blanquart, F. and Fraser, C., 2022. Epidemiological dynamics of bacteriocin competition and antibiotic resistance. *Proceedings of the Royal Society B*, *289*(1984), p.20221197.

Levin, B.R., 1988. Frequency-dependent selection in bacterial populations. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *319*(1196), pp.459-472.

Li, Y. and Rebuffat, S., 2020. The manifold roles of microbial ribosomal peptide–based natural products in physiology and ecology. *Journal of Biological Chemistry*, *295*(1), pp.34-54.

Libberton, B., Horsburgh, M.J. and Brockhurst, M.A., 2015. The effects of spatial structure, frequency dependence and resistance evolution on the dynamics of toxin-mediated microbial invasions. *Evolutionary applications*, *8*(7), pp.738-750.

Liu, C.M., Price, L.B., Hungate, B.A., Abraham, A.G., Larsen, L.A., Christensen, K., Stegger, M., Skov, R. and Andersen, P.S., 2015. *Staphylococcus aureus* and the ecology of the nasal microbiome. *Science advances*, *1*(5), p.e1400216.

MacIntyre, D.L., Miyata, S.T., Kitaoka, M. and Pukatzki, S., 2010. The *Vibrio cholerae* type VI secretion system displays antimicrobial properties. *Proceedings of the National Academy of Sciences*, *107*(45), pp.19520-19524.

MacLean, R.C. and Gudelj, I., 2006. Resource competition and social conflict in experimental populations of yeast. *Nature*, *441*(7092), pp.498-501.

Majeed, H., Gillor, O., Kerr, B. and Riley, M.A., 2011. Competitive interactions in *Escherichia coli* populations: the role of bacteriocins. *The ISME journal*, *5*(1), pp.71-81.

Maldonado-Barragán, A. and West, S.A., 2020. The cost and benefit of quorum sensing-controlled bacteriocin production in *Lactobacillus plantarum*. *Journal of evolutionary biology*, *33*(1), pp.101-111.

Mc Ginty, S.É., Lehmann, L., Brown, S.P. and Rankin, D.J., 2013. 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*(1761), p.20130400.

Mc Ginty, S.É. and Rankin, D.J., 2012. The evolution of conflict resolution between plasmids and their bacterial hosts. *Evolution*, *66*(5), pp.1662-1670.

Mc Ginty, S.E., Rankin, D.J. and Brown, S.P., 2011. Horizontal gene transfer and the evolution of bacterial cooperation. *Evolution: International Journal of Organic Evolution*, *65*(1), pp.21-32.

McInerney, J.O., McNally, A. and O'connell, M.J., 2017. Why prokaryotes have pangenomes. *Nature microbiology*, *2*(4), pp.1-5.

McNally, L., Viana, M. and Brown, S.P., 2014. Cooperative secretions facilitate host range expansion in bacteria. *Nature communications*, *5*(1), p.4594.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), pp.153-157.

Mellmann, A., Weniger, T., Berssenbrügge, C., Rothgänger, J., Sammeth, M., Stoye, J. and Harmsen, D., 2007. Based Upon Repeat Pattern (BURP): an algorithm to characterize the long-term evolution of *Staphylococcus aureus* populations based on *spa* polymorphisms. *BMC microbiology*, *7*(1), pp.1-6.

Miller, R.R., Walker, A.S., Godwin, H., Fung, R., Votintseva, A., Bowden, R., Mant, D., Peto, T.E., Crook, D.W. and Knox, K., 2014. Dynamics of acquisition and loss of carriage of *Staphylococcus aureus* strains in the community: the effect of clonal complex. *Journal of Infection*, *68*(5), pp.426-439.

Mitri, S. and Foster, K.R., 2013. The genotypic view of social interactions in microbial communities. *Annual review of genetics*, *47*, pp.247-273.

Mitri, S. and Foster, K.R., 2016. Pleiotropy and the low cost of individual traits promote cooperation. *Evolution*, *70*(2), pp.488-494.

Mokoena, M.P., 2017. Lactic acid bacteria and their bacteriocins: classification, biosynthesis and applications against uropathogens: a mini-review. *Molecules*, *22*(8), p.1255.

Morris, C.E., Lamichhane, J.R., Nikolić, I., Stanković, S. and Moury, B., 2019. The overlapping continuum of host range among strains in the *Pseudomonas syringae* complex. *Phytopathology Research*, *1*, pp.1-16.

Nadell, C.D., Drescher, K. and Foster, K.R., 2016. Spatial structure, cooperation and competition in biofilms. *Nature Reviews Microbiology*, *14*(9), pp.589-600.

Nadell, C.D., Xavier, J.B. and Foster, K.R., 2008. The sociobiology of biofilms. *FEMS microbiology reviews*, *33*(1), pp.206-224.

Nakagawa, S., Johnson, P.C. and Schielzeth, H., 2017. The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), p.20170213.

Nakagawa, S. and Schielzeth, H., 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, *4*(2), pp.133-142.

Nascimento, J.S., Ceotto, H., Nascimento, S.B., Giambiagi-deMarval, M., Santos, K.R.N. and Bastos, M.C.F., 2006. Bacteriocins as alternative agents for control of multiresistant staphylococcal strains. *Letters in applied microbiology*, *42*(3), pp.215-221.

Nascimento, J.S., Coelho, M.L.V., Ceotto, H., Potter, A., Fleming, L.R., Salehian, Z., Nes, I.F. and Bastos, M.D.C.D.F., 2012. Genes involved in immunity to and secretion of aureocin A53, an atypical class II bacteriocin produced by *Staphylococcus aureus* A53. *Journal of bacteriology*, *194*(4), pp.875-883.

Nascimento, J.S., dos Santos, K.R.N., Gentilini, E., Sordelli, D. and de Freire Bastos, M.D.C., 2002. Phenotypic and genetic characterisation of bacteriocin-producing strains of *Staphylococcus aureus* involved in bovine mastitis. *Veterinary microbiology*, *85*(2), pp.133-144.

Navaratna, M.A., Sahl, H.G. and Tagg, J.R., 1999. Identification of genes encoding two-component lantibiotic production in *Staphylococcus aureus* C55 and other phage group II *S. aureus* strains and demonstration of an association with the exfoliative toxin B gene. *Infection and immunity*, *67*(8), pp.4268-4271.

Netz, D.J.A., Pohl, R., Beck-Sickinger, A.G., Selmer, T., Pierik, A.J., de Freire Bastos, M.D.C. and Sahl, H.G., 2002. Biochemical characterisation and genetic analysis of aureocin A53, a new, atypical bacteriocin from *Staphylococcus aureus*. *Journal of molecular biology*, *319*(3), pp.745-756.

Netz, D.J.A., Sahl, H.G., Marcolino, R., dos Santos Nascimento, J., de Oliveira, S.S., Soares, M.B. and de Freire Bastos, M.D.C., 2001. Molecular characterisation of aureocin A70, a multi-peptide bacteriocin isolated from *Staphylococcus aureus*. *Journal of molecular biology*, *311*(5), pp.939-949.

Newstead, L.L., Varjonen, K., Nuttall, T. and Paterson, G.K., 2020. Staphylococcal-produced bacteriocins and antimicrobial peptides: their potential as alternative treatments for *Staphylococcus aureus* infections. *Antibiotics*, *9*(2), p.40.

Niehus, R., Mitri, S., Fletcher, A.G. and Foster, K.R., 2015. Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nature communications*, *6*(1), p.8924.

Niehus, R., Oliveira, N.M., Li, A., Fletcher, A.G. and Foster, K.R., 2021. The evolution of strategy in bacterial warfare via the regulation of bacteriocins and antibiotics. *Elife*, *10*, p.e69756.

Niewiesk, S. and Prince, G., 2002. Diversifying animal models: the use of hispid cotton rats (*Sigmodon hispidus*) in infectious diseases. *Laboratory animals*, *36*(4), pp.357-372.

Nocelli, N., Bogino, P.C., Banchio, E. and Giordano, W., 2016. Roles of extracellular polysaccharides and biofilm formation in heavy metal resistance of rhizobia. *Materials*, *9*(6), p.418.

Nogueira, T., Rankin, D.J., Touchon, M., Taddei, F., Brown, S.P. and Rocha, E.P., 2009. Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Current Biology*, *19*(20), pp.1683-1691.

Nogueira, T., Touchon, M. and Rocha, E.P., 2012. Rapid evolution of the sequences and gene repertoires of secreted proteins in bacteria. *PloS one*, *7*(11), p.e49403.

Novick, R.P. and Geisinger, E., 2008. Quorum sensing in staphylococci. *Annual review of genetics*, *42*, pp.541-564.

O'Brien, F.G., Yui Eto, K., Murphy, R.J., Fairhurst, H.M., Coombs, G.W., Grubb, W.B. and Ramsay, J.P., 2015. Origin-of-transfer sequences facilitate mobilisation of non-conjugative antimicrobial-resistance plasmids in *Staphylococcus aureus*. *Nucleic acids research*, *43*(16), pp.7971-7983.

O'Hara, F.P., Suaya, J.A., Ray, G.T., Baxter, R., Brown, M.L., Mera, R.M., Close, N.M., Thomas, E. and Amrine-Madsen, H., 2016. *spa* typing and multilocus sequence typing show comparable performance in a macroepidemiologic study of *Staphylococcus aureus* in the United States. *Microbial Drug Resistance*, *22*(1), pp.88-96.

O’Toole, G.A., Crabbé, A., Kümmerli, R., LiPuma, J.J., Bomberger, J.M., Davies, J.C., Limoli, D., Phelan, V.V., Bliska, J.B., DePas, W.H. and Dietrich, L.E., 2021. Model systems to study the chronic, polymicrobial infections in cystic fibrosis: current approaches and exploring future directions. *MBio*, *12*(5), pp.e01763-21.

Okuda, K.I., Zendo, T., Sugimoto, S., Iwase, T., Tajima, A., Yamada, S., Sonomoto, K. and Mizunoe, Y., 2013. Effects of bacteriocins on methicillin-resistant *Staphylococcus aureus* biofilm. *Antimicrobial agents and chemotherapy*, *57*(11), pp.5572-5579.

Otto, M., 2013. Community-associated MRSA: what makes them special? *International Journal of Medical Microbiology*, *303*(6-7), pp.324-330.

Otto, M., 2020. Staphylococci in the human microbiome: the role of host and interbacterial interactions. *Current opinion in microbiology*, *53*, pp.71-77.

Palmer, K.L., Aye, L.M. and Whiteley, M., 2007. Nutritional cues control *Pseudomonas aeruginosa* multicellular behavior in cystic fibrosis sputum. *Journal of bacteriology*, *189*(22), pp.8079-8087.

Palmer, J.D. and Foster, K.R., 2022. The evolution of spectrum in antibiotics and bacteriocins. *Proceedings of the National Academy of Sciences*, *119*(38), p.e2205407119.

Paradis, E. and Schliep, K., 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), pp.526-528.

Pembury Smith, M.Q. and Ruxton, G.D., 2020. Effective use of the McNemar test. *Behavioral Ecology and Sociobiology*, *74*(11), pp.1-9.

Pérez-Gutiérrez, R.A., López-Ramírez, V., Islas, A., Alcaraz, L.D., Hernández-González, I., Olivera, B.C.L., Santillán, M., Eguiarte, L.E., Souza, V., Travisano, M. and Olmedo-Alvarez, G., 2013. Antagonism influences assembly of a *Bacillus* guild in a local community and is depicted as a food-chain network. *The ISME Journal*, *7*(3), pp.487-497.

Peschel A, Otto M, Jack RW, Kalbacher H, Jung G, Gotz F. Inactivation of the dlt operon in *Staphylococcus aureus* confers sensitivity to defensins, protegrins, and other antimicrobial peptides. The Journal of biological chemistry. 1999; 274(13):8405–10. Epub 1999/03/20. PMID: 10085071.

Pfeiffer, T., Schuster, S. and Bonhoeffer, S., 2001. Cooperation and competition in the evolution of ATP-producing pathways. *Science, 292*(5516), pp.504-507.

Pollitt, E.J., West, S.A., Crusz, S.A., Burton-Chellew, M.N. and Diggle, S.P., 2014. Cooperation, quorum sensing, and evolution of virulence in *Staphylococcus aureus*. *Infection and immunity*, *82*(3), pp.1045-1051.

Quereda, J.J., Dussurget, O., Nahori, M.A., Ghozlane, A., Volant, S., Dillies, M.A., Regnault, B., Kennedy, S., Mondot, S., Villoing, B. and Cossart, P., 2016. Bacteriocin from epidemic *Listeria* strains alters the host intestinal microbiota to favor infection. *Proceedings of the National Academy of Sciences*, *113*(20), pp.5706-5711.

Rakoff-Nahoum, S., Coyne, M.J. and Comstock, L.E., 2014. An ecological network of polysaccharide utilization among human intestinal symbionts. *Current biology*, *24*(1), pp.40-49.

Ramsay, J.P. and Firth, N., 2017. Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Current opinion in microbiology, 38,* pp.1-9.

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

Revell, L.J., 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution*, (2), pp.217-223.

Rice, P., Longden, I. and Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. *Trends in genetics*, *16*(6), pp.276-277.

Riley, M.A. and Chavan, M.A., 2007. *Bacteriocins*. Springer-Verlag Berlin Heidelberg.

Riley, M.A. and Gordon, D.M., 1996. The ecology and evolution of bacteriocins. *Journal of Industrial Microbiology*, *17*, pp.151-158.

Riley, M.A. and Gordon, D.M., 1999. The ecological role of bacteriocins in bacterial competition. *Trends in microbiology*, *7*(3), pp.129-133.

Riley, M.A. and Wertz, J.E., 2002. Bacteriocins: evolution, ecology, and application. *Annual Reviews in Microbiology*, *56*(1), pp.117-137.

Robertson, J. and Nash, J.H., 2018. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial genomics*, *4*(8).

Rocha, E.P. and Danchin, A., 2002. Base composition bias might result from competition for metabolic resources. *TRENDS in Genetics*, *18*(6), pp.291-294.

Rodríguez-Beltrán, J., DelaFuente, J., Leon-Sampedro, R., MacLean, R.C. and San Millan, A., 2021. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology*, *19*(6), pp.347-359.

Rodriguez-Beltran, J., Hernandez-Beltran, J.C.R., DelaFuente, J., Escudero, J.A., Fuentes-Hernandez, A., MacLean, R.C., Peña-Miller, R. and San Millan, A., 2018. Multicopy plasmids allow bacteria to escape from fitness trade-offs during evolutionary innovation. *Nature ecology & evolution*, *2*(5), pp.873-881.

Rodríguez-Beltrán, J., Sørum, V., Toll-Riera, M., de la Vega, C., Peña-Miller, R. and San Millán, Á., 2020. Genetic dominance governs the evolution and spread of mobile genetic elements in bacteria. *Proceedings of the National Academy of Sciences*, *117*(27), pp.15755-15762.

Rodríguez-Rubio, L., Serna, C., Ares-Arroyo, M., Matamoros, B.R., Delgado-Blas, J.F., Montero, N., Bernabe-Balas, C., Wedel, E.F., Mendez, I.S., Muniesa, M. and Gonzalez-Zorn, B., 2020. Extensive antimicrobial resistance mobilization via multicopy plasmid encapsidation mediated by temperate phages. *Journal of Antimicrobial Chemotherapy*, *75*(11), pp.3173-3180.

Ross-Gillespie, A., Dumas, Z. and Kümmerli, R., 2015. Evolutionary dynamics of interlinked public goods traits: an experimental study of siderophore production in *Pseudomonas aeruginosa*. *Journal of evolutionary biology*, *28*(1), pp.29-39.

Ross-Gillespie, A., Gardner, A., Buckling, A., West, S.A. and Griffin, A.S., 2009. Density dependence and cooperation: theory and a test with bacteria. *Evolution*, *63*(9), pp.2315-2325.

Ross-Gillespie, A., Gardner, A., West, S.A. and Griffin, A.S., 2007. Frequency dependence and cooperation: theory and a test with bacteria. *The American Naturalist*, *170*(3), pp.331-342.

Russell, A.H. and Truman, A.W., 2020. Genome mining strategies for ribosomally synthesised and post-translationally modified peptides. *Computational and Structural Biotechnology Journal*, *18*, pp.1838-1851.

Ruxton, G. and Colegrave, N., 2011. *Experimental design for the life sciences*. Oxford University Press.

Sainani, K.L., 2015. Dealing with missing data. *PM&R*, *7*(9), pp.990-994.

San Millan, A., Escudero, J.A., Gifford, D.R., Mazel, D. and MacLean, R.C., 2016. Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nature ecology & evolution*, *1*(1), p.0010.

Sandoz, K.M., Mitzimberg, S.M. and Schuster, M., 2007. Social cheating in *Pseudomonas aeruginosa* quorum sensing. *Proceedings of the National Academy of Sciences*, *104*(40), pp.15876-15881.

Schade, J. and Weidenmaier, C., 2016. Cell wall glycopolymers of Firmicutes and their role as nonprotein adhesins. *FEBS letters*, *590*(21), pp.3758-3771.

Schindler, C.A. and Schuhardt, V., 1964. Lysostaphin: a new bacteriolytic agent for the *Staphylococcus*. *Proceedings of the National Academy of Sciences*, *51*(3), pp.414-421.

Sears, H.J., Brownlee, I. and Uchiyama, J.K., 1950. Persistence of individual strains of *Escherichia coli* in the intestinal tract of man. *Journal of bacteriology*, *59*(2), pp.293-301.

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

Shopsin, B., Gomez, M., Montgomery, S.O., Smith, D.H., Waddington, M., Dodge, D.E., Bost, D.A., Riehman, M., Naidich, S. and Kreiswirth, B.N., 1999. Evaluation of protein A gene polymorphic region DNA sequencing for typing of *Staphylococcus aureus* strains. *Journal of clinical microbiology*, *37*(11), pp.3556-3563.

Simonet, C. and McNally, L., 2021. Kin selection explains the evolution of cooperation in the gut microbiota. *Proceedings of the National Academy of Sciences*, *118*(6), p.e2016046118.

Smillie, C., Garcillán-Barcia, M.P., Francia, M.V., Rocha, E.P. and de la Cruz, F., 2010. Mobility of plasmids. *Microbiology and Molecular Biology Reviews*, *74*(3), pp.434-452.

Smith, J., 2001. The social evolution of bacterial pathogenesis. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *268*(1462), pp.61-69.

Som, A., 2015. Causes, consequences and solutions of phylogenetic incongruence. *Briefings in Bioinformatics*, *16*(3), pp.536-548.

Somerville, G.A., 2016. *Staphylococcus: genetics and physiology*. Caister Academic Press.

Souza, D.P., Oka, G.U., Alvarez-Martinez, C.E., Bisson-Filho, A.W., Dunger, G., Hobeika, L., Cavalcante, N.S., Alegria, M.C., Barbosa, L.R., Salinas, R.K. and Guzzo, C.R., 2015. Bacterial killing via a type IV secretion system. *Nature communications*, *6*(1), pp.1-9.

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*, *112*(45), pp.14042-14047.

Stone, G.N., Nee, S. and Felsenstein, J., 2011. Controlling for non-independence in comparative analysis of patterns across populations within species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1569), pp.1410-1424.

Strassmann, J.E., Zhu, Y. and Queller, D.C., 2000. Altruism and social cheating in the social amoeba *Dictyostelium discoideum*. *Nature*, *408*(6815), pp.965-967.

Stubbendieck, R.M. and Straight, P.D., 2016. Multifaceted interfaces of bacterial competition. *Journal of bacteriology*, *198*(16), pp.2145-2155.

Thomas, C.M. and Nielsen, K.M., 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews microbiology*, *3*(9), pp.711-721.

Tong, S.Y., Davis, J.S., Eichenberger, E., Holland, T.L. and Fowler Jr, V.G., 2015. *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clinical microbiology reviews*, *28*(3), pp.603-661.

Trivedi, U., Madsen, J.S., Everett, J., Fell, C., Russel, J., Haaber, J., Crosby, H.A., Horswill, A.R., Burmølle, M., Rumbaugh, K.P. and Sørensen, S.J., 2018. *Staphylococcus aureus* coagulases are exploitable yet stable public goods in clinically relevant conditions. *Proceedings of the National Academy of Sciences*, *115*(50), pp.E11771-E11779.

Turner, N.A., Sharma-Kuinkel, B.K., Maskarinec, S.A., Eichenberger, E.M., Shah, P.P., Carugati, M., Holland, T.L. and Fowler, V.G., 2019. Methicillin-resistant *Staphylococcus aureus*: an overview of basic and clinical research. *Nature Reviews Microbiology*, *17*(4), pp.203-218.

Twisk, J. and de Vente, W., 2002. Attrition in longitudinal studies: How to deal with missing data. *Journal of clinical epidemiology*, *55*(4), pp.329-337.

Uffelmann, E., Huang, Q.Q., Munung, N.S., De Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T. and Posthuma, D., 2021. Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), p.59.

Uhlen, M., Guss, B., Nilsson, B., Gatenbeck, S., Philipson, L. and Lindberg, M., 1984. Complete sequence of the staphylococcal gene encoding protein A. A gene evolved through multiple duplications. *Journal of Biological Chemistry*, *259*(3), pp.1695-1702.

Ulhuq, F.R., Gomes, M.C., Duggan, G.M., Guo, M., Mendonca, C., Buchanan, G., Chalmers, J.D., Cao, Z., Kneuper, H., Murdoch, S. and Thomson, S., 2020. A membrane-depolarizing toxin substrate of the *Staphylococcus aureus* type VII secretion system mediates intraspecies competition. *Proceedings of the National Academy of Sciences*, *117*(34), pp.20836-20847.

van Heel, A.J., de Jong, A., Song, C., Viel, J.H., Kok, J. and Kuipers, O.P., 2018. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic acids research*, *46*(W1), pp.W278-W281.

Ventura, M., Turroni, F., Canchaya, C., Vaughan, E.E., O'Toole, P.W. and van Sinderen, D., 2009. Microbial diversity in the human intestine and novel insights from metagenomics. *Frontiers in bioscience*, *14*(1), pp.3214-3221.

Von Eiff, C., Becker, K., Machka, K., Stammer, H. and Peters, G., 2001. Nasal carriage as a source of *Staphylococcus aureus* bacteremia. *New England Journal of Medicine*, *344*(1), pp.11-16.

Votintseva, A.A., Miller, R.R., Fung, R., Knox, K., Godwin, H., Peto, T.E.A., Crook, D.W., Bowden, R. and Walker, A.S., 2014. Multiple-strain colonization in nasal carriers of *Staphylococcus aureus*. *Journal of clinical microbiology*, *52*(4), pp.1192-1200.

Waite, R.D. and Curtis, M.A., 2009. *Pseudomonas aeruginosa* PAO1 pyocin production affects population dynamics within mixed-culture biofilms. *Journal of bacteriology*, *191*(4), pp.1349-1354.

Washburne, A.D., Morton, J.T., Sanders, J., McDonald, D., Zhu, Q., Oliverio, A.M. and Knight, R., 2018. Methods for phylogenetic analysis of microbiome data. *Nature microbiology*, *3*(6), pp.652-661.

Wertheim, H.F., Melles, D.C., Vos, M.C., van Leeuwen, W., van Belkum, A., Verbrugh, H.A. and Nouwen, J.L., 2005. The role of nasal carriage in *Staphylococcus aureus* infections. *The Lancet infectious diseases*, *5*(12), pp.751-762.

West, S.A., Diggle, S.P., Buckling, A., Gardner, A. and Griffin, A.S., 2007a. The social lives of microbes. *Annual Review of Ecology, Evolution, and Systematics*, pp.53-77.

West, S.A., Griffin, A.S. and Gardner, A., 2007b. Evolutionary explanations for cooperation. *Current biology*, *17*(16), pp.R661-R672.

West, S.A., Griffin, A.S. and Gardner, A., 2007c. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of evolutionary biology*, *20*(2), pp.415-432.

West, S.A., Griffin, A.S., Gardner, A. and Diggle, S.P., 2006. Social evolution theory for microorganisms. *Nature reviews microbiology*, *4*(8), pp.597-607.

Westhoff, S., Kloosterman, A.M., van Hoesel, S.F., van Wezel, G.P. and Rozen, D.E., 2021. Competition sensing changes antibiotic production in *Streptomyces*. *MBio*, *12*(1), pp.e02729-20.

Williams, G.C. and Nesse, R.M., 1991. The dawn of Darwinian medicine. *The Quarterly review of biology*, *66*(1), pp.1-22.

Willsey, G.G. and Wargo, M.J., 2015. Extracellular lipase and protease production from a model drinking water bacterial community is functionally robust to absence of individual members. *Plos one*, *10*(11), p.e0143617.

Wilson, R.A., Handley, B.A. and Beringer, J.E., 1998. Bacteriocin production and resistance in a field population of *Rhizobium leguminosarum biovar viciae*. *Soil Biology and Biochemistry*, *30*(3), pp.413-417.

Xavier, J.B., 2011. Social interaction in synthetic and natural microbial communities. *Molecular systems biology*, *7*(1), p.483.

Yan, M., Pamp, S.J., Fukuyama, J., Hwang, P.H., Cho, D.Y., Holmes, S. and Relman, D.A., 2013. Nasal microenvironments and interspecific interactions influence nasal microbiota complexity and *S. aureus* carriage. *Cell host & microbe*, *14*(6), pp.631-640.

Young, B.C., Golubchik, T., Batty, E.M., Fung, R., Larner-Svensson, H., Votintseva, A.A., Miller, R.R., Godwin, H., Knox, K., Everitt, R.G. and Iqbal, Z., 2012. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proceedings of the National Academy of Sciences*, *109*(12), pp.4550-4555.

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. and Brinkman, F.S., 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, *26*(13), pp.1608-1615.

Zerbino, D.R. and Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, *18*(5), pp.821-829.

Zinder, N.D. and Lederberg, J., 1952. Genetic exchange in *Salmonella*. *Journal of bacteriology*, *64*(5), pp.679-699.

Zipperer, A., Konnerth, M.C., Laux, C., Berscheid, A., Janek, D., Weidenmaier, C., Burian, M., Schilling, N.A., Slavetinsky, C., Marschal, M. and Willmann, M., 2016. Human commensals producing a novel antibiotic impair pathogen colonization. *Nature*, *535*(7613), pp.511-516.

Appendix: Chapter 4. Plasmids do not consistently stabilise cooperation across bacteria but may promote broad pathogen host-range

nature portfolio

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

Supplementary information

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

S1 – Supplementary Results for Genomic Analyses

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

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

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

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

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

We have considered the relative rates of transfer among the three mobility types, where conjugative plasmids transfer at faster rates than mobilizable, and mobilizable transfer at faster rates than non-mobilizable[2]. However, the variation in transfer rates within plasmids of the same mobility type is likely to be large, and mobilization via mechanisms other than conjugation, such as phage transfer, is possible[2–5].

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

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

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

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

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

Number of environments.

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

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

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

Core vs accessory genes.

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

Effect sizes, variance explained and significance.

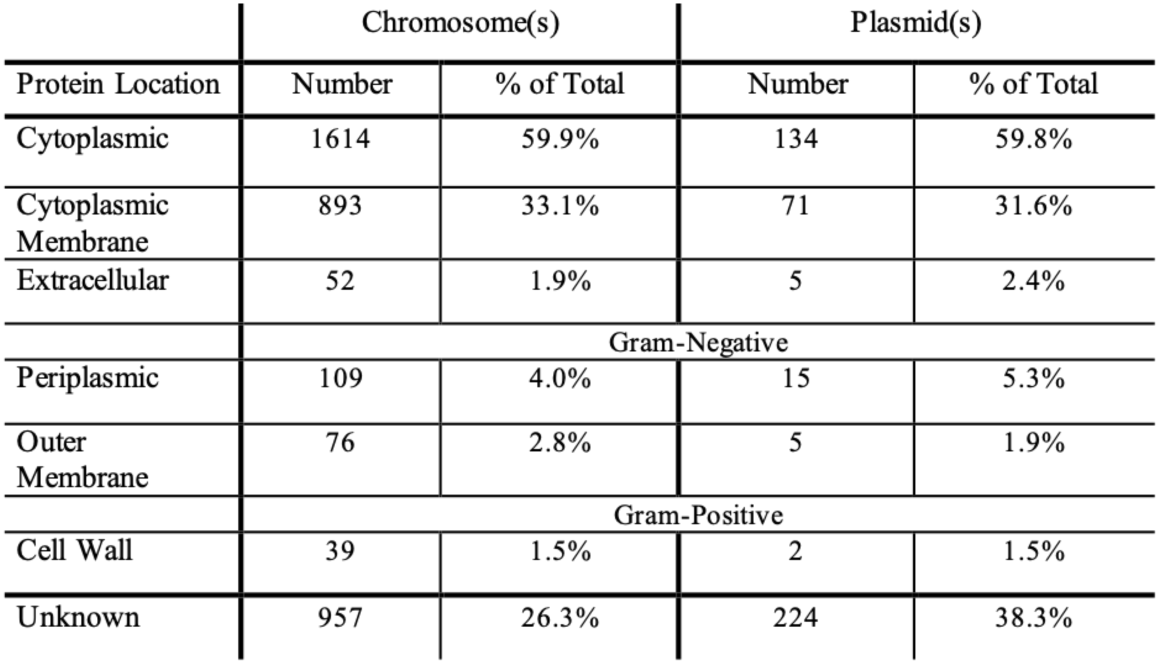
The percentage of variance explained that is considered biologically significant is subjective and can depend upon the kind of data you are examining, and the field of research. In many areas of evolution and ecology, 5-10% can be a reasonable baseline, but in some areas 1% could be argued for[1,13].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

S1

References

1. Crawley, M. J. *Statistics: An Introduction Using R*. (John Wiley & Sons, 2014).
2. Smillie, C., Garcillan-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & de la Cruz, F. Mobility of Plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434–452 (2010).
3. O'Brien, F. G. *et al.* Origin-of-transfer sequences facilitate mobilisation of non-conjugative antimicrobial-resistance plasmids in Staphylococcus aureus. *Nucleic Acids Res.* **43**, 7971–7983 (2015).
4. Ramsay, J. P. & Firth, N. Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Curr. Opin. Microbiol.* **38**, 1–9 (2017).
5. Rodríguez-Rubio, L. *et al.* Extensive antimicrobial resistance mobilization via multicopy plasmid encapsidation mediated by temperate phages. *J. Antimicrob. Chemother.* **75**, 3173–3180 (2020).
6. Sheppard, R. J., Beddis, A. E. & Barraclough, T. G. The role of hosts, plasmids and environment in determining plasmid transfer rates: A meta-analysis. *Plasmid* **108**, 102489 (2020).
7. Garcia-Garcera, M., Touchon, M., Brisse, S. & Rocha, E. P. C. Metagenomic assessment of the interplay between the environment and the genetic diversification of Acinetobacter. *Environ. Microbiol.* **19**, 5010–5024 (2017).
8. Garcia-Garcera, M. & Rocha, E. P. C. Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. *Nat. Commun.* **11**, 758 (2020).
9. Kümmerli, R., Schiessl, K. T., Waldvogel, T., McNeill, K. & Ackermann, M. Habitat structure and the evolution of diffusible siderophores in bacteria. *Ecol. Lett.* **17**, 1536–1544 (2014).
10. Domingo-Sananes, M. R. & McInerney, J. O. Mechanisms That Shape Microbial Pangenomes. *Trends Microbiol.* **0**, (2021).
11. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040 (2017).
12. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* **46**, e5 (2018).
13. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Routledge, 1988).
14. Jennions, M. D. & Møller, A. P. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behav. Ecol.* **14**, 438–445 (2003).
15. West, S. A., Shuker, D. M. & Sheldon, B. C. Sex-ratio adjustment when relatives interact: a test of constraints on adaptation. *Evol. Int. J. Org. Evol.* **59**, 1211–1228 (2005).
16. West, S. *Sex Allocation*. (Princeton University Press, 2009).

S2 – Supplementary Tables



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

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

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

S2

Table S2. MCMCglmm analyses

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

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

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

S2

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

S2

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

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

S2

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

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

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

S2

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

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

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

Table S3. Measures of Bacterial Lifestyle and Environmental Variability

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

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

S2

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

S2

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

S2

Table S3 References

1. Garcia-Garcera, M. & Rocha, E. P. C. Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. *Nat. Commun.* **11**, 758 (2020).
2. Fournier, P. E., Richet, H. & Weinstein, R. A. The Epidemiology and Control of Acinetobacter baumannii in Health Care Facilities. *Clin. Infect. Dis.* **42**, 692–699 (2006).
3. Howard, A., O'Donoghue, M., Feeney, A. & Sleator, R. D. Acinetobacter baumannii. *Virulence* **3**, 243–250 (2012).
4. Yang, H., Liang, L., Lin, S. & Jia, S. Isolation and Characterization of a Virulent Bacteriophage AB1 of Acinetobacter baumannii. *BMC Microbiol.* **10**, 131 (2010).
5. Anane A, Y., Apalata, T., Vasaikar, S., Okuthe, G. E. & Songca, S. Prevalence and molecular analysis of multidrug-resistant Acinetobacter baumannii in the extra-hospital environment in Mthatha, South Africa. *Braz. J. Infect. Dis.* **23**, 371–380 (2019).
6. Al Atrouni, A., Joly-Guillou, M.-L., Hamze, M. & Kempf, M. Reservoirs of Non-baumannii Acinetobacter Species. *Front. Microbiol.* **7**, (2016).
7. Spencer, R. C. Bacillus anthracis. *J. Clin. Pathol.* **56**, 182–187 (2003).
8. Koehler, T. M. Bacillus anthracis physiology and genetics. *Mol. Aspects Med.* **30**, 386–396 (2009).
9. Bottone, E. J. Bacillus cereus, a Volatile Human Pathogen. *Clin. Microbiol. Rev.* **23**, 382–398 (2010).
10. Messelhäußer, U. & Ehling-Schulz, M. Bacillus cereus—a Multifaceted Opportunistic Pathogen. *Curr. Clin. Microbiol. Rep.* **5**, 120–125 (2018).
11. Harwood, C. R. Bacillus subtilis and its relatives: molecular biological and industrial workhorses. *Trends Biotechnol.* **10**, 247–256 (1992).
12. Earl, A. M., Losick, R. & Kolter, R. Ecology and genomics of Bacillus subtilis. *Trends Microbiol.* **16**, 269 (2008).
13. Argôlo-Filho, R. C. & Loguercio, L. L. Bacillus thuringiensis Is an Environmental Pathogen and Host-Specificity Has Developed as an Adaptation to Human-Generated Ecological Niches. *Insects* **5**, 62–91 (2013).
14. Garbutt, J., Bonsall, M. B., Wright, D. J. & Raymond, B. Antagonistic competition moderates virulence in Bacillus thuringiensis. *Ecol. Lett.* **14**, 765–772 (2011).
15. Rabbee, M. F. *et al.* Bacillus velezensis: A Valuable Member of Bioactive Molecules within Plant Microbiomes. *Molecules* **24**, (2019).
16. Reva, O. N. *et al.* Genetic, Epigenetic and Phenotypic Diversity of Four Bacillus velezensis Strains Used for Plant Protection or as Probiotics. *Front. Microbiol.* **10**, (2019).
17. Moran, N. A. & Mira, A. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. *Genome Biol.* **2**, research0054.1 (2001).
18. Sheppard, S. K. & Maiden, M. C. J. The Evolution of Campylobacter jejuni and Campylobacter coli. *Cold Spring Harb. Perspect. Biol.* **7**, (2015).
19. Andersen, A. A. Serotyping of US Isolates of Chlamydophila Psittaci from Domestic and Wild Birds. *J. Vet. Diagn. Invest.* **17**, 479–482 (2005).
20. Harkinezhad, T., Geens, T. & Vanrompay, D. Chlamydophila psittaci infections in birds: A review with emphasis on zoonotic consequences. *Vet. Microbiol.* **135**, 68–77 (2009).

S2

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

S2

43. Ivanek, R., Gröhn, Y. T. & Wiedmann, M. Listeria monocytogenes in multiple habitats and host populations: review of available data for mathematical modeling. *Foodborne Pathog. Dis.* **3**, 319–336 (2006).
44. Hill, S. A., Masters, T. L. & Wachter, J. Gonorrhea - an evolving disease of the new millennium. *Microb. Cell* **3**, 371–389.
45. Bramucci, A. R. *et al.* The Bacterial Symbiont Phaeobacter inhibens Shapes the Life History of Its Algal Host Emiliania huxleyi. *Front. Mar. Sci.* **5**, (2018).
46. Fryer, J. L. & Hedrick, R. P. Piscirickettsia salmonis: a Gram-negative intracellular bacterial pathogen of fish. *J. Fish Dis.* **26**, 251–262 (2003).
47. Drzewiecka, D. Significance and Roles of Proteus spp. Bacteria in Natural Environments. *Microb. Ecol.* **72**, 741–758 (2016).
48. Pellett, S., Bigley, D. V. & Grimes, D. J. Distribution of Pseudomonas aeruginosa in a riverine ecosystem. *Appl. Environ. Microbiol.* **45**, 328–332 (1983).
49. Sandoz, K. M., Mitzimberg, S. M. & Schuster, M. Social cheating in Pseudomonas aeruginosa quorum sensing. *Proc. Natl. Acad. Sci.* **104**, 15876–15881 (2007).
50. Morris, C. E. *et al.* The life history of the plant pathogen Pseudomonas syringae is linked to the water cycle. *ISME J.* **2**, 321–334 (2008).
51. Rohmer, L., Kjemtrup, S., Marchesini, P. & Dangl, J. L. Nucleotide sequence, functional characterization and evolution of pFKN, a virulence plasmid in Pseudomonas syringae pathovar maculicola. *Mol. Microbiol.* **47**, 1545–1562 (2003).
52. Morris, C. E., Lamichhane, J. R., Nikolić, I., Stanković, S. & Moury, B. The overlapping continuum of host range among strains in the Pseudomonas syringae complex. *Phytopathol. Res.* **1**, 4 (2019).
53. Álvarez, B., López, M. M. & Biosca, E. G. Biocontrol of the Major Plant Pathogen Ralstonia solanacearum in Irrigation Water and Host Plants by Novel Waterborne Lytic Bacteriophages. *Front. Microbiol.* **10**, (2019).
54. Gutarra, L., Herrera, J., Fernandez, E., Kreuze, J. & Lindqvist-Kreuze, H. Diversity, Pathogenicity, and Current Occurrence of Bacterial Wilt Bacterium Ralstonia solanacearum in Peru. *Front. Plant Sci.* **8**, (2017).
55. Labes, G., Ulrich, A. & Lentzsch, P. Influence of Bovine Slurry Deposition on the Structure of Nodulating Rhizobium leguminosarum bv. viciae Soil Populations in a Natural Habitat. *Appl. Environ. Microbiol.* **62**, 1717–1722 (1996).
56. Lowendorf, H. S. & Alexander, M. Identification of Rhizobium phaseoli Strains That Are Tolerant or Sensitive to Soil Acidity. *Appl. Environ. Microbiol.* **45**, 737–742 (1983).
57. Andino, A. & Hanning, I. Salmonella enterica: Survival, Colonization, and Virulence Differences among Serovars. *Sci. World J.* **2015**, (2015).
58. Hejazi, A. & Falkiner, F. R. Serratia marcescens. *J. Med. Microbiol.* **46**, 903–912 (1997).
59. Huang, G. *et al.* Isolation of a Novel Heterotrophic Nitrification–Aerobic Denitrification Bacterium Serratia marcescens CL1502 from Deep-Sea Sediment. *Environ. Eng. Sci.* **34**, 453–459 (2017).
60. Roumiantseva, M. L. *et al.* Diversity of Sinorhizobium meliloti from the Central Asian Alfalfa Gene Center. *Appl. Environ. Microbiol.* **68**, 4694–4697 (2002).
61. Tong, S. Y. C., Davis, J. S., Eichenberger, E., Holland, T. L. & Fowler, V. G. Staphylococcus aureus infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clin. Microbiol. Rev.* **28**, 603–661 (2015).

62. Pollitt, E. J. G., West, S. A., Crusz, S. A., Burton-Chellew, M. N. & Diggle, S. P. Cooperation, Quorum Sensing, and Evolution of Virulence in Staphylococcus aureus. *Infect. Immun.* **82**, 1045–1051 (2014).
63. Otto, M. Staphylococcus epidermidis – the “accidental” pathogen. *Nat. Rev. Microbiol.* **7**, 555–567 (2009).
64. de Souza Santos, M., Salomon, D., Li, P., Krachler, A.-M. & Orth, K. 8 - Vibrio parahaemolyticus virulence determinants. in *The Comprehensive Sourcebook of Bacterial Protein Toxins (Fourth Edition)* (eds. Alouf, J., Ladant, D. & Popoff, M. R.) 230–260 (Academic Press, 2015). doi:10.1016/B978-0-12-800188-2.00008-2.
65. Vieira, G. *et al.* Terrestrial and marine Antarctic fungi extracts active against Xanthomonas citri subsp. citri. *Lett. Appl. Microbiol.* **67**, 64–71 (2018).
66. Patané, J. S. L. *et al.* Origin and diversification of Xanthomonas citri subsp. citri pathotypes revealed by inclusive phylogenomic, dating, and biogeographic analyses. *BMC Genomics* **20**, 700 (2019).
67. Ference, C. M. *et al.* Recent advances in the understanding of Xanthomonas citri ssp. citri pathogenesis and citrus canker disease management. *Mol. Plant Pathol.* **19**, 1302–1318 (2018).
68. Baldi, P. & La Porta, N. Xylella fastidiosa: Host Range and Advance in Molecular Identification Techniques. *Front. Plant Sci.* **8**, (2017).
69. Fredriksson-Ahomaa, M., Stolle, A. & Korkeala, H. Molecular epidemiology of Yersinia enterocolitica infections. *FEMS Immunol. Med. Microbiol.* **47**, 315–329 (2006).
70. Harvey, S., Greenwood, J. R., Pickett, M. J. & Mah, R. A. Recovery of Yersinia enterocolitica from streams and lakes of California. *Appl. Environ. Microbiol.* **32**, 352–354 (1976).
71. Eisen, R. J. *et al.* Persistence of Yersinia pestis in Soil Under Natural Conditions. *Emerg. Infect. Dis.* **14**, 941–943 (2008).
72. Santos-Montañez, J., Benavides-Montaño, J. A., Hinz, A. K. & Vadyvaloo, V. Yersinia pseudotuberculosis IP32953 survives and replicates in trophozoites and persists in cysts of Acanthamoeba castellanii. *FEMS Microbiol. Lett.* **362**, (2015).
73. Gemski, P., Lazere, J. R., Casey, T. & Wohlhieter, J. A. Presence of a virulence-associated plasmid in Yersinia pseudotuberculosis. *Infect. Immun.* **28**, 1044–1047 (1980).

S3 – Supplementary Figures for Genomic Analyses

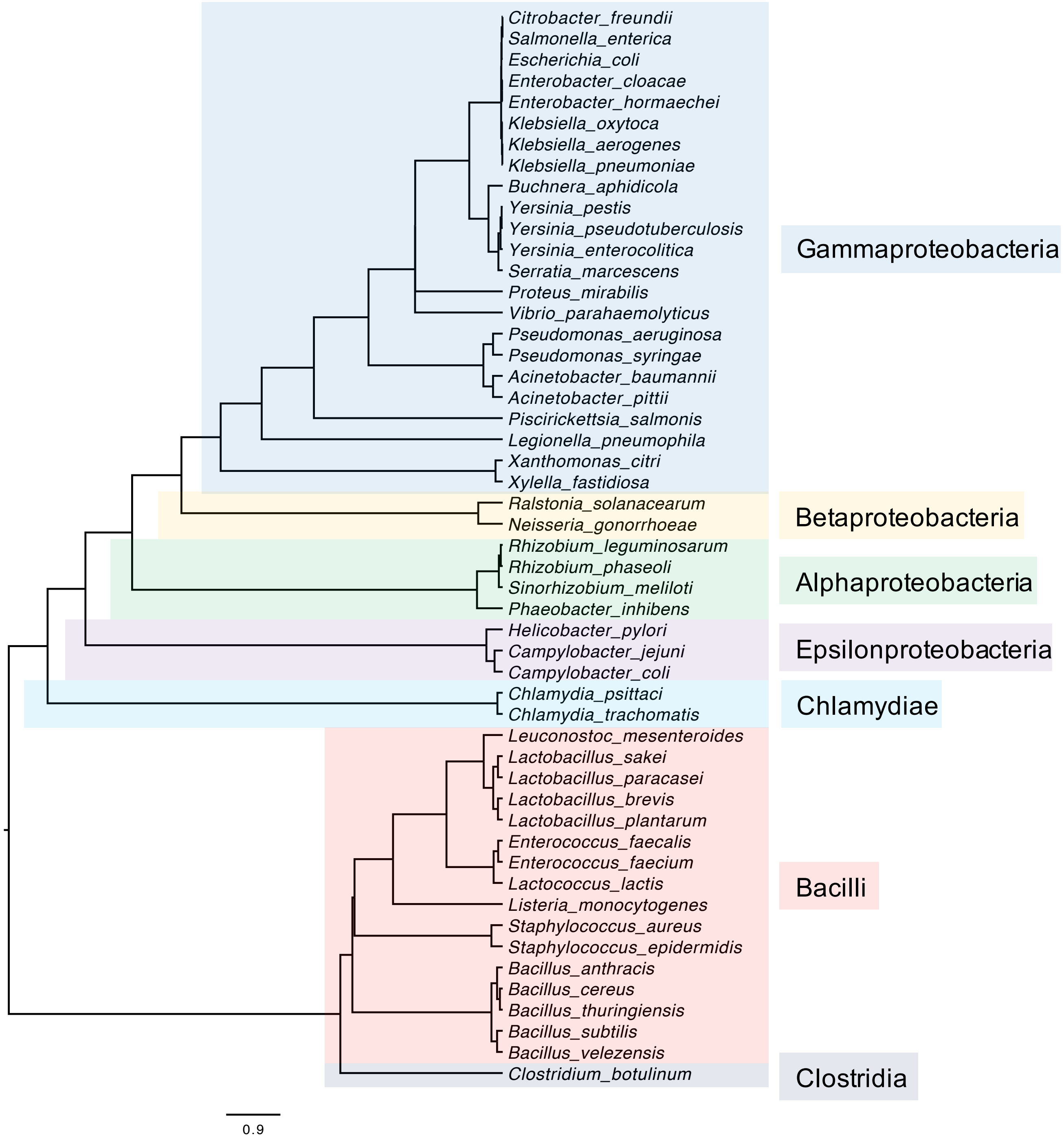


Figure S1. Phylogeny of all 51 species in our dataset.

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

S3

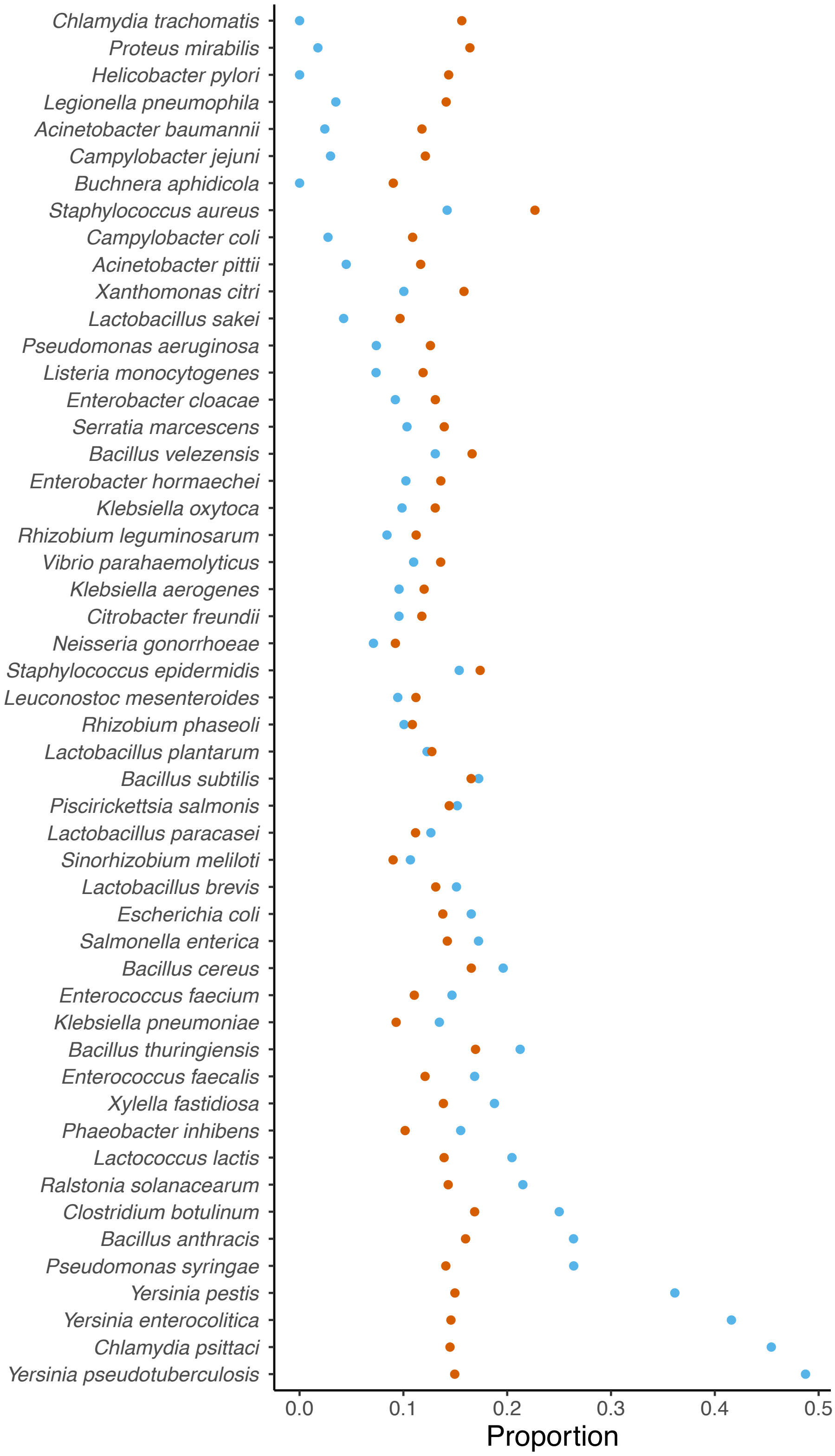


Fig S2. Proportion of proteins predicted as extracellular for plasmids and chromosomes. Each species has two proportions: the blue dot is the mean proportion of plasmid proteins predicted by PSORTb to be extracellular across all plasmids in that species, while the red dot is the mean proportion of plasmid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

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

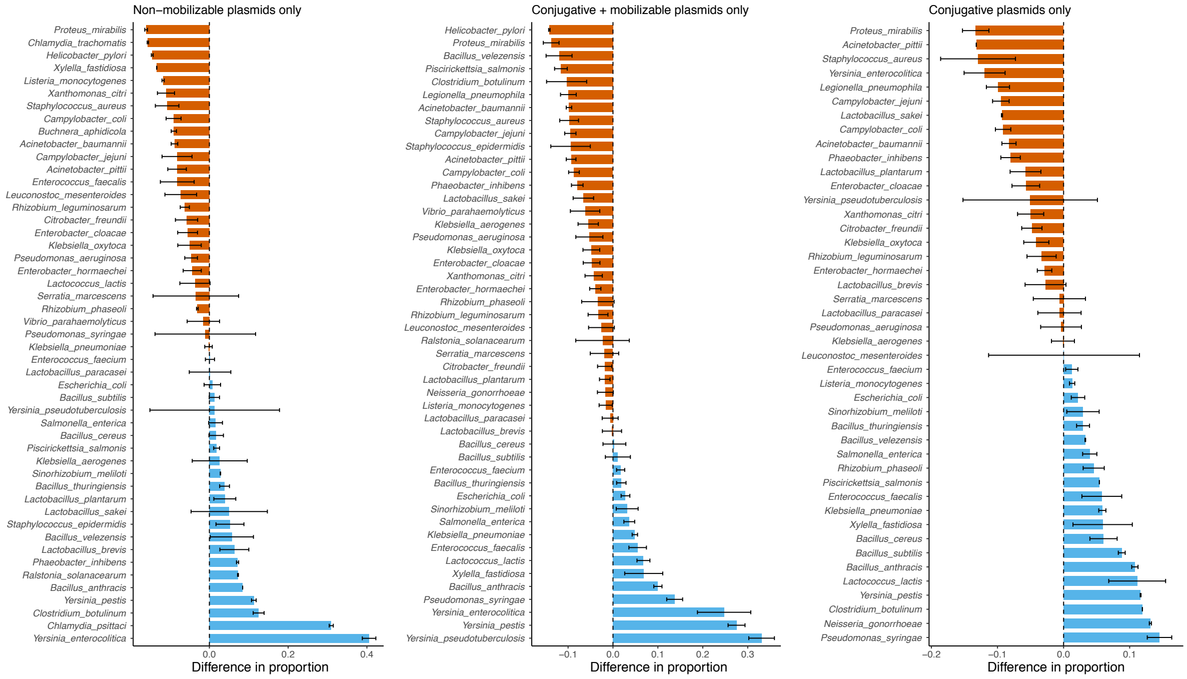


Fig S3. Extracellular proteins are not consistently overrepresented on plasmids of all three mobilities (non-mobilizable, mobilizable, conjugative)

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

S3

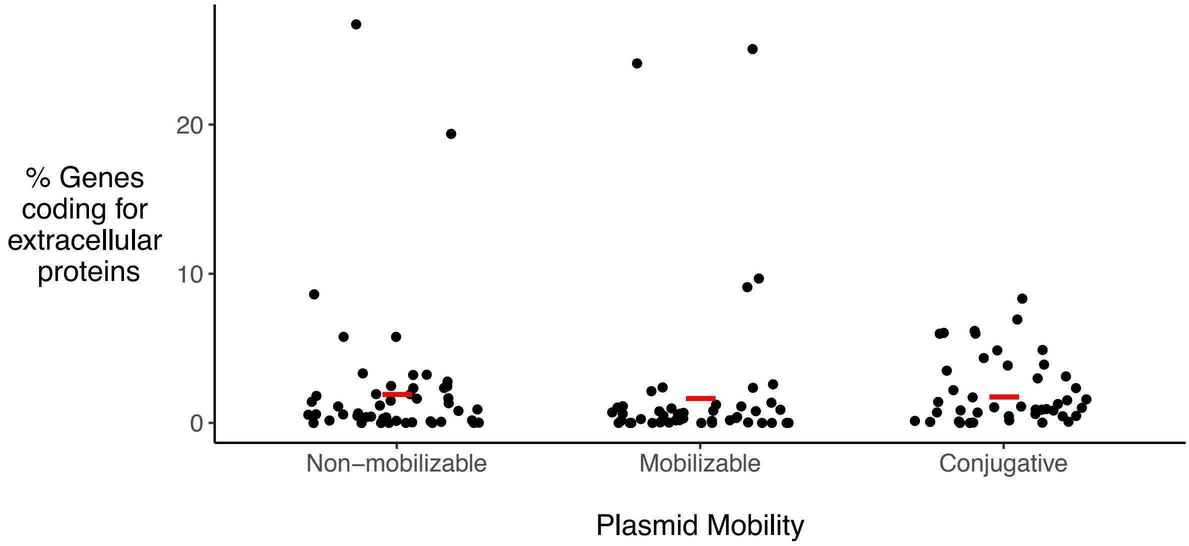


Figure S4. No difference in the mean % of genes coding for extracellular proteins across the three mobility types.

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

S3

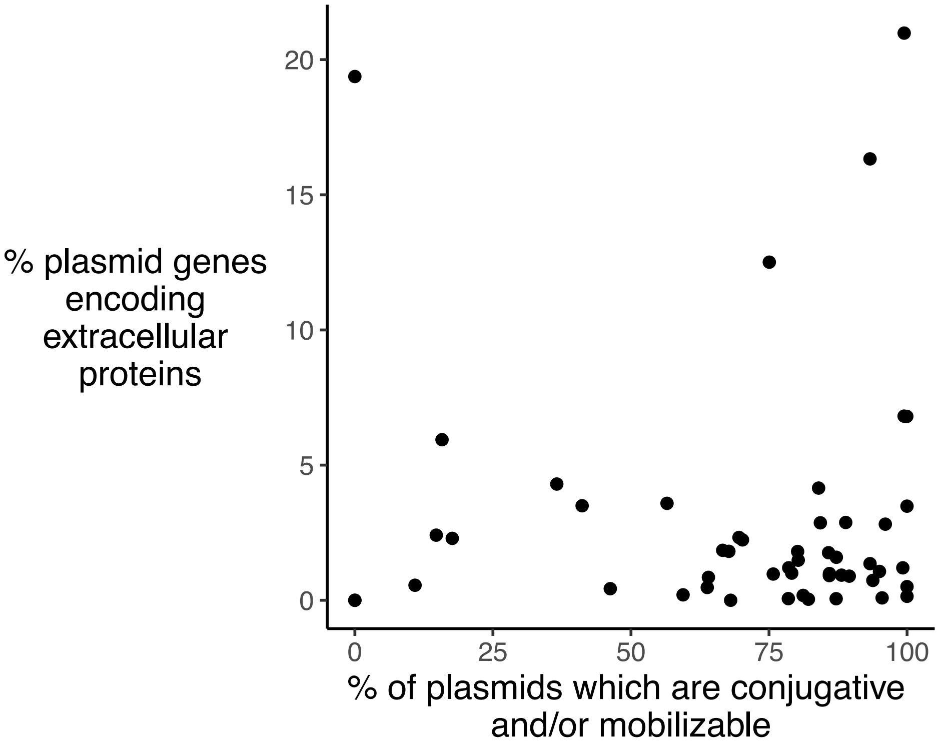


Figure S5. No effect of a species' plasmid mobility and % plasmid genes coding for extracellular proteins.

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

S3

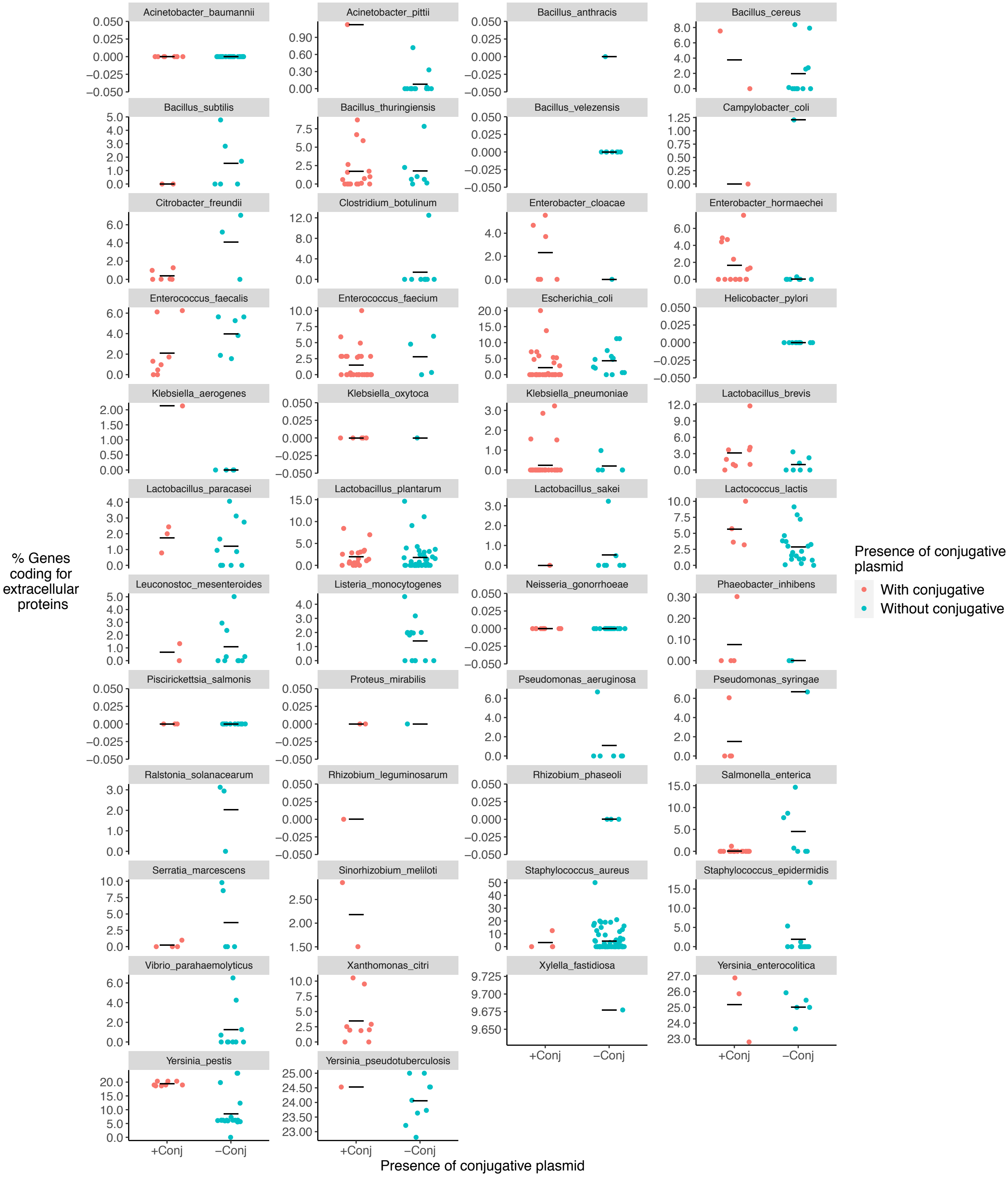


Figure S6. Co-occurrence of mobilizable plasmids with conjugative plasmids. Each panel shows data for one of the 46 species which had at least one genome with at least one mobilizable plasmid. Each dot corresponds to a genome which had at least one mobilizable

S3

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

S3

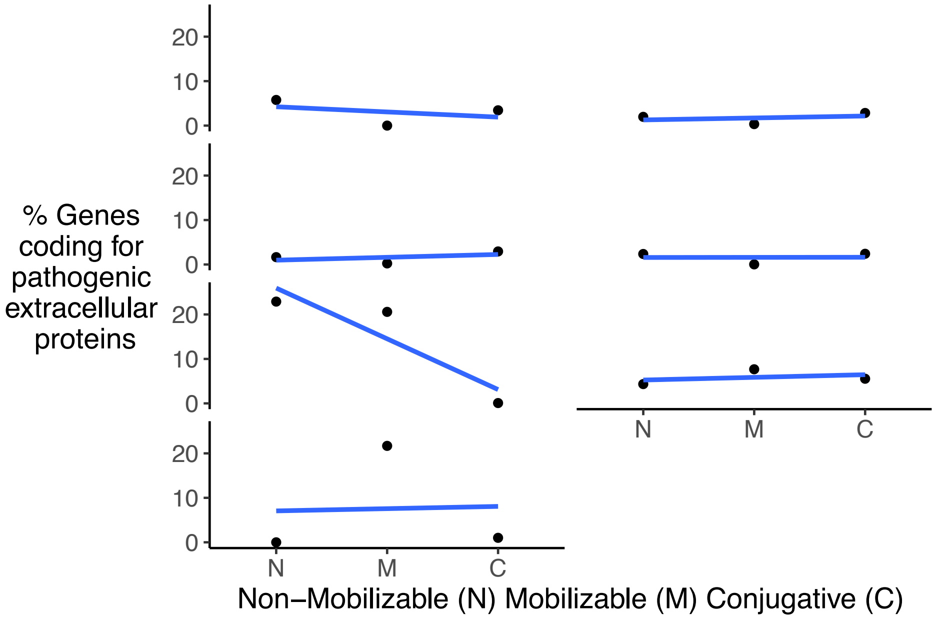


Fig S7. Pathogenic extracellular proteins are not more likely to be carried by higher mobility plasmids in broad host-range pathogen species.

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

S3

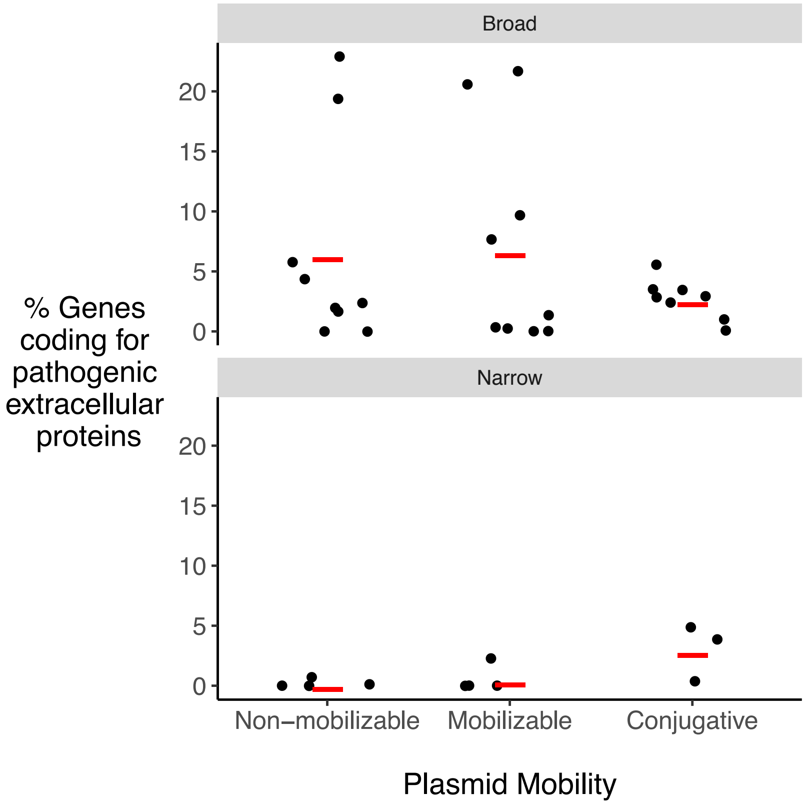


Figure S8. Pathogenic extracellular proteins are not more likely to be carried by more mobile plasmids in both broad and narrow host-range pathogen species

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

S3

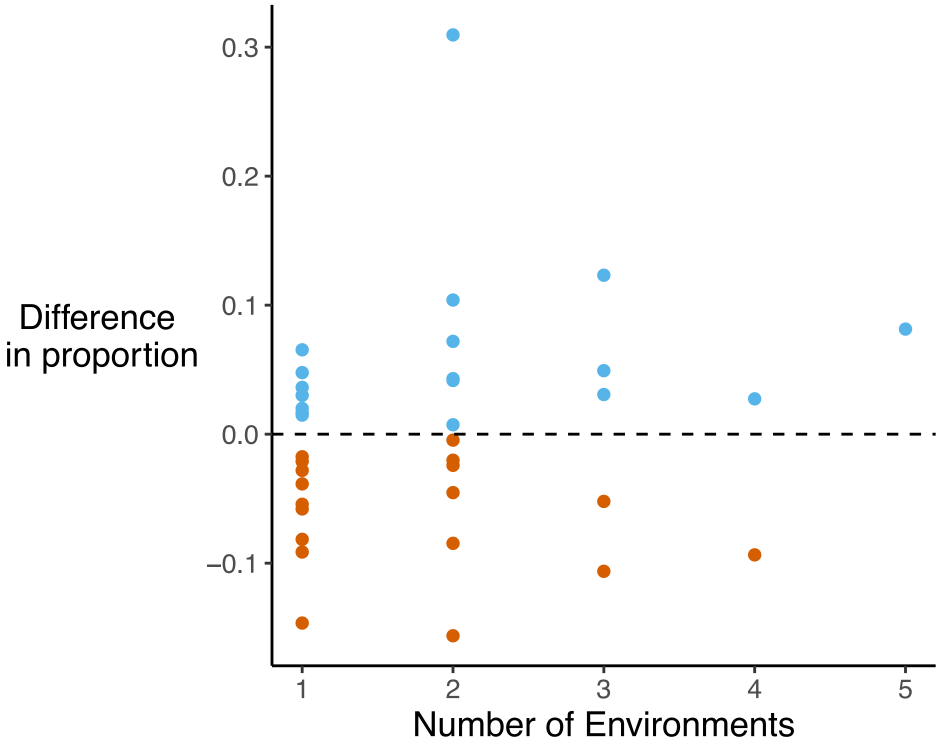


Figure S9. No significant correlation between the number of five broad environments a species is found in and how overrepresented or underrepresented extracellular proteins are on plasmids.

The x-axis shows the original published data of the number of five broad environments a species is found in, with 36 of the species in our dataset represented in the dataset. The y-axis shows the difference in the proportion of genes on plasmids and chromosomes coding for extracellular proteins. Each dot is the mean for all genomes in a species. Species in blue are those with extracellular proteins overrepresented on plasmids, while species in red are those with extracellular proteins overrepresented on chromosomes.

S3

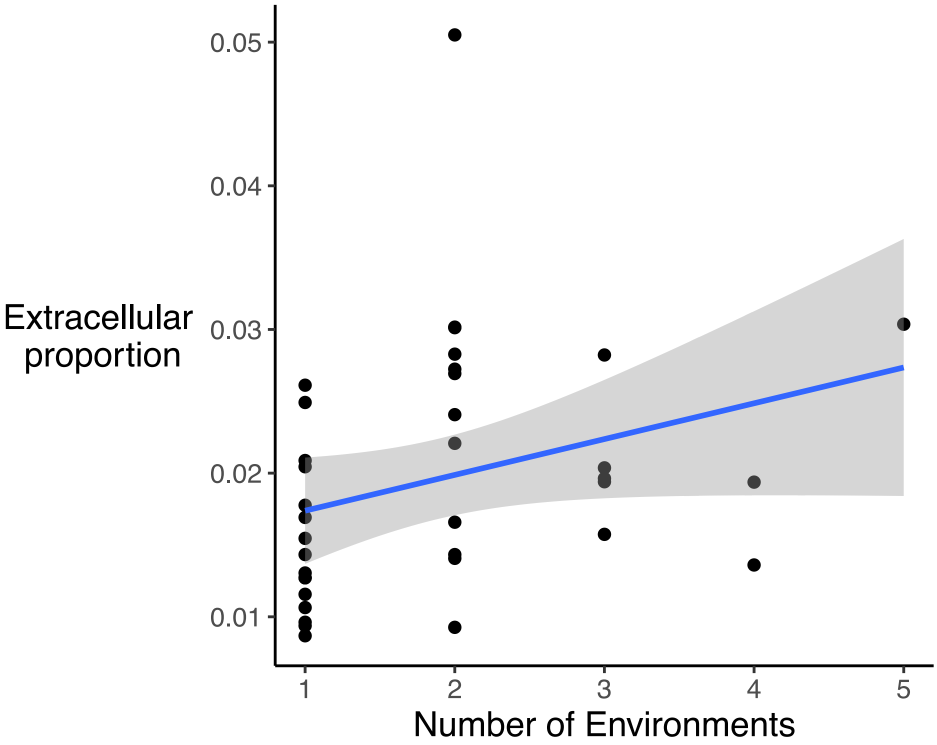


Figure S10. Positive but non-significant correlation between the number of five broad environments a species is found in and the proportion of the genome which encodes extracellular proteins.

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