Unraveling Phage-Bacteria Interactions: Phage Resistance and Defense System Dynamics in Cystic Fibrosis *P. aeruginosa* Isolates and the Statistical Methods for Assessing Network Nestedness

Ezra Levi Herman

Doctor of Philosophy

University of York
Department of Biology
October 2023

Abstract

With a long history of co-evolution, the mechanisms underlying interactions between bacteria and their viral predators, phages, remain poorly understood. A deeper understanding of these mechanisms would enable better design of phages for antimicrobial therapy, known as phage therapy. This work investigated the factors driving bacterial immunity to phages, known as phage resistance, from empirical, bioinformatic, and statistical perspectives. Empirically, the mechanisms underlying phage resistance in *Pseudomonas aeruginosa*, a major pathogen in Cystic Fibrosis (CF) lung infections, were characterised. Genetic variation in receptor biosynthesis genes associated with resistance breadth, while the defence system repertoire associated with resistance specificity. Bioinformatically, the pan-immune system of 456 clinical *P. aeruginosa* isolates from a previous longitudinal CF study was characterised. Despite most systems being localised to variable genomic regions, signatures of horizontal gene transfer were rare, and defence system repertoires remained mostly unchanged over time, suggesting a broad but stable pan-immune system. Statistically, the methods for determining nestedness in phage-bacteria infection networks (PBINs) were evaluated. An alternative null model more accurately classified nested and non-nested PBINs, indicating that previous nestedness analyses may need revisiting. Overall, this thesis suggests complementary roles for receptor modification and defence system carriage in *P. aeruginosa* phage resistance, proposes that the theory of frequent horizontal defence system transfer does not apply to clinical CF *P. aeruginosa* isolates and recommends a more robust approach for analysing PBIN nestedness.

Acknowledgements

I would first like to thank my supervisors, Ville Friman and Paul Fogg. The past four years have allowed me to explore my interests in lab experimentation, bioinformatics, statistics and teaching, and never was I told that a path was not worth pursuing. Your supervision offered me the guidance and the freedom to make the most out of my PhD, and I am incredibly thankful for your time and for all that I've learnt.

I would like to thank Jamie Wood for critically appraising my work and for guiding me at critical stages in my degree. I would also like to thank Franklin Nobrega and Jon Pitchford for orchestrating a stimulating viva and for suggesting improvements to my work.

Major highlights of my time in York involved teaching: thank you Emma Rand for taking me under your wings through our Carpentries project, and thank you Sue Russel and Anna Riach for allowing me to teach for countless hours at the Maths Skills Centre.

Thank you to the team at Bactobio for teaching me invaluable software skills, for providing career guidance at a critical point in my PhD and for the fun I had during my placement.

Thank you to the friends that I made at the Friman lab, at the Alan Turing Institute and through other walks of life for bringing joy into this period of my life, and for standing by me when times were tough.

Finally, I could not have completed this work without the love and support of my family, to whom I am endlessly grateful.

Project funding

This work was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) through the White Rose DTP in Mechanistic Biology. This work was also funded through the Enrichment Placement Scheme at The Alan Turing Institute.

Author's declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Contents

1. Chapter 1: General Introduction 7

   1 Motivation: *Pseudomonas aeruginosa* infections in Cystic Fibrosis patients 7

   1.1 *P. aeruginosa* infections in Cystic Fibrosis patients are hard to treat 7

   1.2 Phage therapy may aid in the management and treatment of *P. aeruginosa* infections in CF 8

   2 An introduction to phage-bacteria interactions 8

   2.1 Phage resistance mechanisms break specific stages of the infection cycle 8

   2.2 Defence systems are thought to spread by horizontal gene transfer 11

   2.3 Bioinformatics allows for the discovery of potential defence mechanisms 11

   3 Studying phage-bacteria co-evolution at a network level 12

   3.1 Host resistance and phage counter-resistance drive co-evolution 12

   3.2 Two major modes of co-evolution 12

   3.3 The development of phage therapy may benefit from co-evolutionary predictions 12

   3.4 Network-level signatures of co-evolutionary dynamics 13

   4 Nestedness metrics and null models 14

   4.1 Nestedness metrics 14

   4.2 Uniform null model leads to questionable inferences 15

   4.3 Choosing an alternative null model: an introduction to the Tuning-Peg landscape 15

   5 Thesis overview 18

   6 References 19

2. Chapter 2: Estimating the association between phage resistance and its determinants in clinical Cystic Fibrosis *Pseudomonas aeruginosa* isolates 23

   0 Abstract 23

   1 Introduction 23

   2 Methods 26

   2.1 Isolates used in this work 26

   2.2 Infectivity assay 27

   2.2.1 SCFM preparation 27

   2.2.2 Phage stock preparation 29

   2.2.3 Overlay agar plate preparation 29

   2.2.4 Phage titre determination 30

   2.2.5 Phage master plate preparation 30

   2.2.6 Running the infectivity assay 30

   2.2.7 Determining infectivity 30

   2.2.8 Adsorption assay 31

   2.3 Bioinformatic analysis 31

   2.3.1 Data source and pipeline construction 31

   2.3.2 Assembly and annotation of a PAO1 genome 31

   2.3.3 Determining genetic variants in receptor biosynthesis genes 31

   2.3.4 Defence system prediction 31

   2.3.5 Associations in pairwise distances 31

   2.3.6 Associations between phage resistance and the number of genetic variants or defence system families 32

2.3.7 Determining important LPS variants using elastic net regularisation 32
2.3.8 Associations between phage resistance and genetic variant or defence system presence 32
3 Results 33
3.1 Determining phage infectivity 33
3.2 No significant association between pairwise distances in receptor biosynthesis gene variants and resistance range 35
3.3 Isolates with more variants in LPS biosynthesis genes were more resistant to LPS phages 37
3.4 Genetic variation in the *rmlA* gene may have contributed to LPS phage resistance 39
3.5 A missense mutation in the *fimT* gene may have reduced phiKZ infectivity 41
3.6 Similarity in defence subsystem repertoire was positively associated with similarity in relative OD values 43
3.7 Defence system repertoire size and phage resistance breadth were not significantly associated 44
3.8 An adsorption assay suggested that the druantia defence system was not driving resistance to PNM and phiKZ phages 45
4 Discussion 49
5 References 52
6 Supplementary materials 57

3. Chapter 3: Clinical Cystic Fibrosis *P. aeruginosa* isolates show a broad but stable defence system repertoire 62

0 Abstract 62
1 Introduction 62
2 Methods 64
2.1 Genomic data source 64
2.2 Detecting putative defence systems 64
2.3 Collector's curve analysis 64
2.4 Associating the number of defence systems with genomic features 64
2.5 Detecting prophage sequences 65
2.6 Visualising prophage features 65
2.7 Localising defence systems to regions of genomic plasticity 65
2.8 Identifying putative defence system hotspots 65
2.9 Characterising defence system variation within spots 65
2.10 Estimating changes in defence system repertoire size over time within clone types and patients 66
3 Results 67
3.1 Clinical isolates carried many different defence systems 67
3.2 Genome fragmentation may have negatively biased defence system detection 67
3.3 Clinical isolates with larger genomes carried more systems 67
3.4 Defence systems were infrequently found in prophage sequences 67
3.5 Defence systems in regions of genomic plasticity 71
3.6 Defence system hotspots 71
3.7 Spots suggested infrequent sharing of large defence system modules between non-clonal isolates 71
3.8 Closely-related isolates mostly carried the same systems within spots, except for a few signals of potential recent insertion 73
3.9 Systems were rarely gained or lost within patients over time 73

4 Discussion 76
5 References 78

4. Chapter 4: Robust statistical methods suggest that a minority of empirical phage-bacteria infection networks are nested 81

0 Abstract 81
1 Introduction 81
2 Methods 84
2.1 Overview 84
2.2 Simulation of theoretical infectivity matrices 86
2.3 Experimental data processing 86
2.4 Construction of null matrices 86
2.5 Calculation of nestedness effect sizes 87
2.6 Association between SES, size and fill across simulation types 87
2.7 Association between the proportion of qFF null matrices that were unique and matrix size 87
2.8 Visualisation of PBINs and null nestedness distributions 87
3 Results 88
3.1 Across all null model and nestedness metric combinations, uniform matrices were significantly non-nested 88
3.2 The NODF and qFF combination reliably identified nested matrices as nested 89
3.3 Only with EE was there an erroneous significant nestedness effect in matrices with heterogeneous marginal totals 90
3.4 Only the NODF metric generally considered compartmentalized matrices anti-nested 92
3.5 Only with qFF was the SES largely invariant to network size and fill 93
3.6 Only with the EE null model was there a non-zero nestedness effect in empirical PBINs . 95
3.7 With the smallest simulated matrices, qFF could struggle to generate unique null matrices after NODF or NT reordering 96
4 Discussion 98
5 References 100

5. Chapter 5: General discussion 103

1 References 105

Chapter 1: General introduction

1 Motivation: *Pseudomonas aeruginosa* infections in Cystic Fibrosis patients

1.1 *P. aeruginosa* infections in Cystic Fibrosis patients are hard to treat

Cystic Fibrosis (CF) is a life-shortening genetic disorder, prevalent in the Caucasian population. The disorder is caused by mutations in the Cystic Fibrosis Transmembrane conductance Regulator (CFTR) gene, resulting in dysfunction in the maintenance of osmotic balance in epithelial surfaces (1). Approximately 1 in 3,500 newborns in Europe, where CF newborn screening is prevalent, are born with the disorder (2). Across the world, more than 100,000 individuals are estimated to live with the disorder (3). While more than 700 disease-causing mutations have been identified (4), the most common mutation is the deletion of a phenylalanine at position 508 (ΔF508). Out of 32,100 patients in the CF Foundation Patient Registry, 85.5% carried at least one copy of this variant (5).

The dysfunction of CFTR results in the secretion of thickened mucus (1). CF patients are thought to suffer from frequent infections due to abnormal clearing of the thickened mucus (1). In addition, the altered pH of the airway surface liquid may impair innate immunity function (1). Infections with *Staphylococcus aureus* or *Haemophilus influenzae* are prominent in early life, while *P. aeruginosa* is associated with infections in adolescence and adulthood (5–7). This shift is thought to be partly driven by the antibiotic therapy used against the early colonisers (6), as well as the lung damage caused by the inflammatory response to early infection (1).

Many patients eventually develop chronic infection with *P. aeruginosa*, which is associated with reduced lung capacity and earlier onset of late-stage lung disease and mortality (6). Therefore, clinicians attempt to delay or prevent chronic infection by exposing patients to antibacterial therapy upon initial colonisation (8). While this may have resulted in the reduction of *P. aeruginosa* prevalence observed in recent years (9), it is unclear whether these treatments improve long-term clinical outcomes (8). Once *P. aeruginosa* establishes a chronic infection, patients are continuously treated with antibiotics in an effort to reduce bacterial load (10). However, *P. aeruginosa* is likely to evolve resistance to the antibiotics used: many of the prevalent isolates associated with worsened disease outcomes carry antibiotic resistance (11) and long-term studies have found *P. aeruginosa* to acquire antibiotic resistance within CF patients, in part driven by hypermutators (12). Additionally, phenotypic adaptations such as mucoidy, biofilm formation, persistence and small colony variant formation can limit the effectiveness of antibiotics (12). Moreover, due to the spatial structure of the lung, *P. aeruginosa* with differing antibiotic resistance profiles can co-exist, complicating antimicrobial susceptibility testing and treatment (13). Recurrent and chronic infections contribute to the lung damage associated with CF, eventually resulting in lung failure (1). Since respiratory failure is the leading cause of mortality in CF patients (5, 7), it is important to develop treatments that eliminate pulmonary infections in CF patients.

Besides antibiotics, CF patients in select Western countries are increasingly being treated with CFTR modulators (CFTRm) (3, 5, 7). These work to restore the function of CFTR by promoting correct protein folding and cell trafficking or increasing the ability of the CFTR channel to open for chloride and bicarbonate secretion (14). However, patients still experience *P. aeruginosa* infections, and more generally the effect of CFTRm on lung infections remains unclear (15). Additionally, CFTRm is unavailable in most of the world, the current costs of CFTRm make them inaccessible to many CF patients and not all CF patients carry eligible CFTR mutations (3, 14, 16). Given the impact of *P. aeruginosa* infections on the lives of CF patients, it is pertinent to develop additional treatments for the delay, and ideally prevention, of chronic *P. aeruginosa* infections.

1.2 Phage therapy may aid in the management and treatment of *P. aeruginosa* infections in CF

Phages are viruses that specifically infect bacteria, relying on bacteria for their replication (17). Due to their ability to kill bacteria as part of their replication, phages are being developed as therapeutics for a wide range of infections, also known as "phage therapy" (18, 19). A high degree of target specificity which minimises damage to the microbiome, the potential to penetrate biofilms, a rapid discovery process and the potential to re-sensitise pathogens to antibiotics make phages attractive candidates for the treatment of bacterial infections (19). Nevertheless, their application faces several hurdles, such as the limited understanding of their interaction with the human immune system, the current absence of a well-established route to regulatory approval and the low quality and quantity of phage therapy studies (18, 19). In the context of CF *P. aeruginosa* infections, phage therapy trials are in their infancy, with only three single-patient case reports identified by a recent systematic review (20). While the phage appear to reduce bacterial load in the three reports (21–23), bacterial isolates with phage resistance are obtained over the course of treatment with limited (21) or high (23) frequency. While phage resistance may hamper treatment, it is not necessarily an unfavourable outcome, as phage resistance can sometimes promote antibiotic susceptibility or reduced virulence (19, 24–26). Nevertheless, in addition to the need for trials with larger cohorts and standardised treatment protocols (18, 20), there is a need to understand the ways in which *P. aeruginosa* adapts to phages, which could greatly improve our ability to predict the effects of phage administration and thereby minimise the risk of resistance evolution (27).

2 An introduction to phage-bacteria interactions

2.1 Phage resistance mechanisms break specific stages of the infection cycle

For a phage to infect a bacterium, the phage requires the bacterium to display a suitable receptor for binding, after which it injects its nucleic acid into the host cell (Figure 1A). For some phages, this is followed by nucleic acid replication, transcription, translation, assembly of phage components into phage progeny and lysis to release the progeny (28) (Figure 1B-D). These phages are known as 'obligately lytic' (29) (Figure 1). Some lytic phages possess the ability to incorporate their nucleic acid into the bacterial genome (Figure 1E), after which their genome is replicated alongside the host (Figure 1F). These are known as 'lytic, temperate' phages, with the process of integration known as 'lysogeny' (29) (Figure 1). This is not a one-way street: phages can sometimes enter their lytic replication cycles again, releasing further virions (30) (Figure 1G). In the context of phage therapy, 'obligately lytic' phages are generally of interest (19). Note that not all phages release virions via lysis. 'Chronic' phages extrude phages without killing the host cell (31). Similarly to lytic, temperate phages, a subset of chronic phage strains can enter a temperate state (29). Since chronic phages do not kill the host as part of their replication, they are generally not of interest in phage therapy (32).

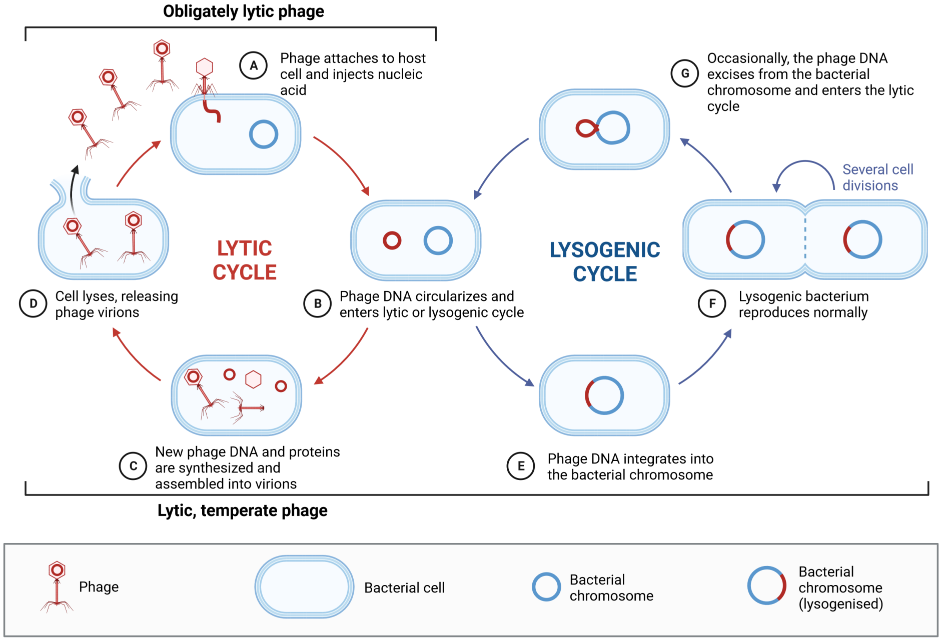
Obligately lytic phage
A
Phage attaches to host cell and injects nucleic acid
G
Occasionally, the phage DNA excises from the bacterial chromosome and enters the lytic cycle
Several cell divisions
LYTIC CYCLE
LYSOGENIC CYCLE
D
Cell lyses, releasing phage virions
B
Phage DNA circularizes and enters lytic or lysogenic cycle
F
Lysogenic bacterium reproduces normally
C
New phage DNA and proteins are synthesized and assembled into virions
E
Phage DNA integrates into the bacterial chromosome
Lytic, temperate phage
Phage
Bacterial cell
Bacterial chromosome
Bacterial chromosome (lysogenised)

Figure 1: Lytic phages can be obligately lytic or temperate. Both types bind to the host cell and inject their nucleic acid (A). Obligately lytic phages do not have a lysis-lysogeny decision (B) and proceed to direct the construction and assembly of new virions (C) which are then released by lysis (D). Instead, temperate phages can enter the lysogenic cycle (B), after which their genome is integrated into the host genome (E). The phage genome is replicated alongside the host (F), until it is released by excision (G). The temperate phage can then enter the lytic cycle (B), spreading new virions through lysis (C,D). Figure generated with BioRender.

Bacteria can break the stages of phage infection through modification or masking of phage binding sites (Figure 2A), inhibition of nucleic acid injection through superinfection exclusion (Figure 2B), degradation of phage nucleic acid (Figure 2C) or induction of lysis or growth arrest before the phage can complete phage particle assembly (28, 33). This final mechanism is known as “abortive infection” (Figure 2D) (28). The systems mediating the intracellular stages of infection have been surfacing over recent years, however many more systems are left to be discovered, as well as the underlying mechanisms (34).

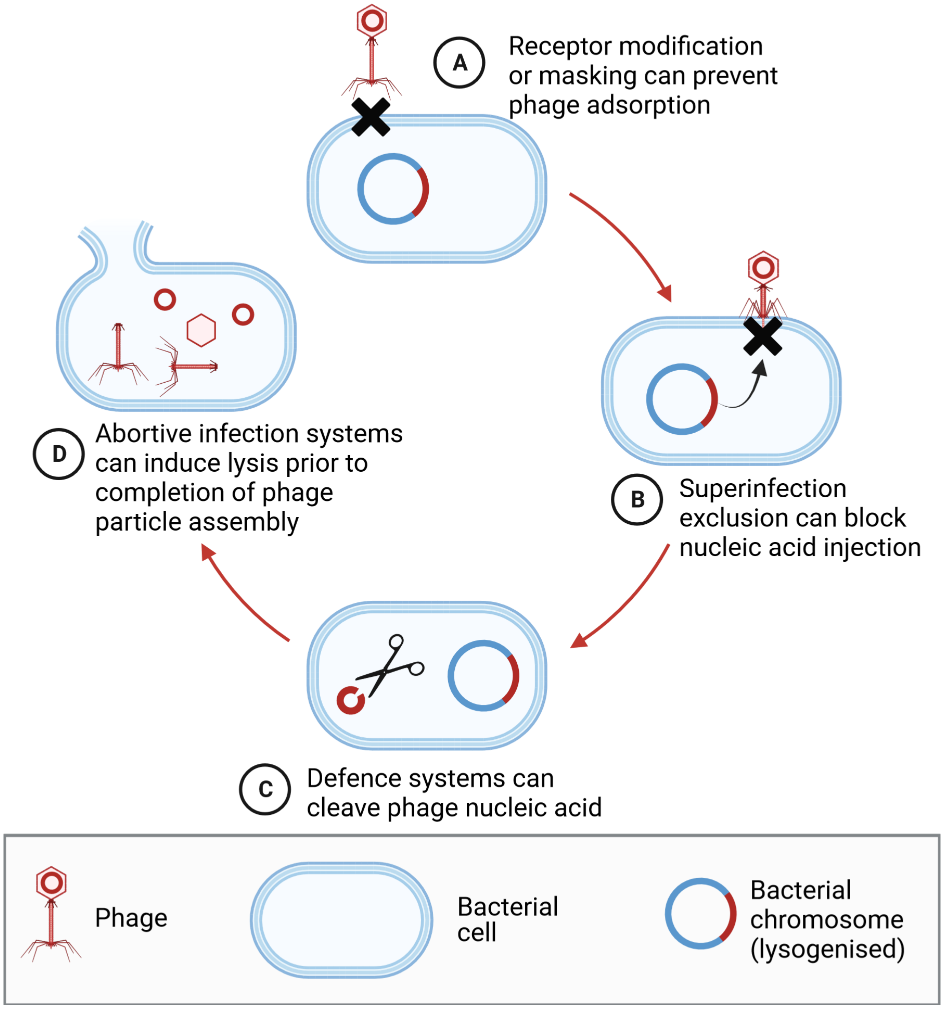
A
Receptor modification or masking can prevent phage adsorption
B
Superinfection exclusion can block nucleic acid injection
C
Defence systems can cleave phage nucleic acid
D
Abortive infection systems can induce lysis prior to completion of phage particle assembly
Phage
Bacterial cell
Bacterial chromosome (lysogenised)

Figure 2: Bacteria can resist phages at various stages of the infection cycle. A) Phage adsorption can be prevented through modification or masking of the phage receptor. B) If the host carries an appropriate temperate phage, nucleic acid injection can be blocked through superinfection exclusion. C) Defence systems can cleave foreign nucleic acids. D) Abortive infection systems can trigger growth arrest or cell death, for example thourgh membrane disruption. Figure generated with BioRender.

In the context of *P. aeruginosa*, mutations in biosynthesis genes for the lipopolysaccharide and Type IV pilus (T4P) receptors, as well as masking of T4P through glycosylation, have been reported to provide phage resistance (35, 36). Only one example of superinfection exclusion has been documented in *P. aeruginosa,* with the associated mechanism remaining unknown (37).

For defence following successful injection, *P. aeruginosa* has been found to carry many systems per isolate, with diversity between isolates (38, 39). Many of these systems function by degrading phage nucleic acids. For example, many *P. aeruginosa* isolates carry the classical restriction-modification (RM) and CRISPR-Cas systems. RM systems of types I-III use an endonuclease to cleave unmethylated motifs on incoming foreign DNA, while host DNA is methylated by a methyltransferase and is thereby protected from cleavage (40). In contrast, Type IV RM targets methylated foreign DNA, leaving unmethylated host DNA intact (40). The CRISPR-Cas system is often referred to as “adaptive”, given the three stages through which it functions: ‘adaptation’, in which signatures of foreign DNA are stored in the genome as CRISPR spacers, ‘expression’, in which CRISPR spacers are transcribed, processed into CRISPR RNAs (crRNAs) and loaded onto Cas proteins, and ‘interference’, during which Cas proteins are guided by the crRNAs towards foreign nucleic acids for cleavage (40). Many other systems that degrade nucleic acids have recently been discovered (34) and detected in *P. aeruginosa* (41), such as the Ssp and Dnd systems, however their function is often not well understood (34).

Many abortive infection systems have been detected in *P. aeruginosa* (38, 39), with some also shown to offer protection. For example, CBASS and Avs (formerly AVAST) Type V were found to protect PAO1 from a variety of phages (41). CBASS may recognise the phage major capsid protein, followed by production of cyclic oligonucleotides which then activate one of a variety of effectors that mediate abortive infection (34, 42, 43). Similarly, Avs senses viral proteins and mediates Abi through a variety of effectors (44). For example, SIR2 domains, which the Type V system carries, deplete the cell of a co-enzyme essential for metabolic reactions, NAD$^+$ (34).

Overall, with the mechanisms behind many defence systems left to be uncovered, and most defence system characterization having been performed in *E. coli* and *B. subtilis* (34), the mechanistic understanding of phage defence systems in *P. aeruginosa* is in its infancy. Over time, a thorough mechanistic understanding will allow for the study of the drivers of resistance in natural and clinical settings.

2.2 Defence systems are thought to spread by horizontal gene transfer

While mutation in receptor biosynthesis genes or global regulators can prevent phage binding (35), the defence system repertoire is thought to mainly change through horizontal gene transfer (34). With many defence systems discovered or detected in close association with mobile genetic element (MGE) genes (45–47), and many defence systems sharing similar functional modules (34), the theory goes that MGEs are driving the gain (through infection), modification (through recombination) and loss (due to fitness costs of MGE carriage) of defence systems (48). By offering protection against lytic phages, MGEs can provide a fitness advantage and promote their maintenance. For example, temporal fluctuations in resistance to an environmental phage in *V. cholerae* were associated with a variable set of defence systems carried by SXT integrative and conjugative elements (49). Beyond their own gain, MGEs can affect the acquisition of other MGEs (50). For example, conjugative elements known as pKLC102-like elements have been found to carry CRISPR–Cas systems with spacers matching known plasmids and prophages (51). Similarly, plasmids may employ CRISPR-Cas to prevent other plasmids from becoming part of their host (52).

The infectivity of MGEs has been suggested as a one of the drivers of the high number of defence systems in some genomes (48). Defence systems could also accumulate under selection for a multilayered immune system, where multiple systems could provide broader defence (48). While this paints an exciting and dynamic picture, the dynamics may be found to differ between different species and different environments. For one, not all species carry many systems and many different system families: for example, *B. subtilis* carry relatively few systems and *H. pylori* carry many systems but of relatively few families (38). Additionally, environmental factors may tip the balance in favour of mutational resistance. For example, receptor-mediated resistance may be favoured over CRISPR-Cas when phage abundance is high (53). Overall, with an increasing ability to map defence systems to genomes and MGEs will come an ability to study the balance between mutational and horizontal adaptation across genetic backgrounds, species and environments.

2.3 Bioinformatics allows for the discovery of potential defence mechanisms

In order to study phage defence mechanisms, the defence mechanisms available to bacteria need to be laid out. The discovery of many defence systems has relied on the concept of “defence islands”: chromosomal regions that contain relatively many defence systems and mobilome genes, flanked by house-keeping genes (54). Such defence islands allow for a “guilt-by-association” approach, whereby genes of unknown function that co-localise with known defence genes are tested for phage resistance capabilities (34). In *P. aeruginosa*, a recent analysis discovered two defence islands, which could vary greatly in size and defence system content across isolates (39). Acquisition of MGEs with defence systems, followed by loss of mobility through mutation, may have given rise to defence islands (48). In line with this, some systems were detected by screening genes of unknown function found in phages or phage satellites (47), or were found to localise to such MGEs after discovery (55). The availability of two bioinformatics tools, PADLOC and DefenseFinder, makes it possible to detect known systems in bacterial genomes at relative ease (38, 56). The detection of potential receptor-mediated resistance requires an alternative approach: comparison of genetic variants in receptor genes and global regulators to a sensitive host (35). While the detection of defence systems and receptor modifications have usually been studied in isolation, their combined application could provide insight into the complete defence repertoire of an isolate, allowing the drivers of phage resistance to be disentangled.

3 Studying phage-bacteria co-evolution at a network level

3.1 Host resistance and phage counter-resistance drive co-evolution

Phages adapt to the changing landscape of host protective mechanisms, described above, with their own counter-adaptations. In terms of adsorption, single point-mutations in tail fibre genes can expand the host range of *P. aeruginosa* phages (57, 58). In a more complex example in *E. coli*, phages were found to respond to loss of the LamB and OmpC receptors by switching to the OmpF receptor, followed by the host acquiring different OmpF variants, followed by the phage adapting to those new variants (59).

Beyond adsorption, phages have been found to carry counter-adaptations to defence systems. For example, phages may be under selection to lose recognition motifs used by the Type II restriction system (60). Additionally, phages can carry systems that protect them from restriction, for example through masking of restriction sites, deactivation of the host restriction endonuclease or induction of the host methylase, such that restriction sites on the phage DNA are methylated before the host manages to cleave them (61). Co-evolution has not stopped with anti-restriction systems, with hosts evolving their own counter-adaptations. For example, the PARIS system induces growth arrest upon inhibition of the EcoKI Type I RM system by the Ocr system of phage T7 (47). Similarly, inhibition of the Type I restriction endonuclease *Eco*prr1 by the phage T4 Stp system can trigger PrrC, which halts protein synthesis through tRNA$^{\mathrm{lys}}$ cleavage (62). Interestingly, phage T4 can religate the cleaved tRNAs, thereby overcoming the counter-defence mechanism of the host (62). In addition to anti-restriction systems, phages can carry anti-CRISPR proteins, which inhibit CRISPR function by inhibiting DNA binding, DNA cleavage or binding to the crRNA, as well as inducing dimer formation (63).

Phages have also evolved to counter abortive infection systems. For example *P. aeruginosa* phages were recently found to counter CBASS by sequestering its signalling molecule, and acquiring escape mutations in the major capsid following experimental knockout of the sequestering mechanism (42). Overall, phages evolve adaptations to the defence mechanisms of hosts, driving further adaptations in the hosts to maintain resistance. This continual development of resistance and counter-resistance is at the core of phage-bacteria co-evolution.

3.2 Two major modes of co-evolution

Classically, phage-bacteria co-evolution has been described as following one of two modes: arms-race dynamics (ARD) or fluctuation selection dynamics (FSD) (64). In ARD, beneficial alleles sweep to fixation in hosts and phages (64). Over time, with bacteria maintaining resistance to past phages and phages maintaining infectivity to past hosts, broader resistance and infectivity ranges are expected (65). As such, bacteria and phages are under directional selection for generalist resistance and infectivity, respectively. In contrast, in FSD, a continuous cycle is expected of changing genotype frequencies: the common bacterial genotype is susceptible to the common phage genotype, which favours a rare host genotype that resists the common phage, which subsequently favours a rare phage that infects the newly common host (64). In contrast to ARD, bacteria and phages are not expected to accumulate greater resistance and infectivity ranges over time under FSD (66).

Experimental co-evolution studies have started to reveal the conditions that favour either of these modes of co-evolution. Generally speaking, ARD is thought to be more costly due to the escalatory accumulation of adaptations. For example, with the same host and phage, ARD was observed in lab cultures and FSD was observed in soil (65, 66), where resistance accumulation in the soil environment was thought to be more costly (66). In a similar vein, different co-evolutionary dynamics between different phage strains with the same host were suggested to be driven by different costs of resistance (67). FSD was also suggested to be more likely than ARD under lower nutrient availability (68). Interestingly, co-evolution can also include (elements) of both modes. For example, FSD may follow ARD due to the costs of ever-increasing resistance ranges (69). Instead, there may be multiple rounds of ARD, resulting in specialisation along different host receptor variants (59).

3.3 The development of phage therapy may benefit from co-evolutionary predictions

Principally, the examples above highlight the difficulty in predicting co-evolutionary dynamics. However, such predictions may be important for the development of phage therapeutics. For example, ARD-promoting factors such as high nutrient availability (68) could be applied to the development of highly infective phages for phage therapy (70–72). Moreover, treatment with a phage cocktail may provide better suppression of the co-occuring pathogen genotypes than use of a single phage, if the within-patient dynamics follow FSD (72). FSD may also be preferred over ARD within the patient, to prevent the evolution of broadly-resistant pathogens. Therefore, an increased understanding of factors promoting different modes of co-evolution could aid in the development of effective phage therapeutics.

3.4 Network-level signatures of co-evolutionary dynamics

To infer whether variation in hosts, phages or experimental conditions are driving changes in co-evolutionatry dynamics, patterns need to be identified that represent ARD or FSD. Classically, time-shift experiments have been used, whereby phage resistance is measured against previous, contemporary and future hosts (64, 73). ARD is then represented by hosts increasing in resistance and phages increasing in infectivity over time. Instead, FSD is represented by fluctuations in resistance and infectivity over time.

Increasingly, networks are used to infer co-evolutionary dynamics (74). These networks are referred to as Phage-Bacteria Infection Networks (PBINs). PBINs are constructed by measuring the infectivity of a panel of phages across a panel of bacteria (Figure 3A). Such networks can be the result of co-evolution experiments, natural sampling or collections of phages and hosts that are not known to have had any contact before (75). Given that ARD selects for phages with increasing host ranges and bacteria with greater resistance ranges, a stair-like pattern is expected in an ARD network (Figure 3B). This pattern is made up of two components: firstly, there are generalist to specialist gradients (GTSGs) in resistance ranges of hosts and host ranges of phages. Secondly, host and resistance ranges form nested subsets. This stair-like pattern is called “nestedness” (74). In contrast, given that FSD selects for phages that infect a subset of the population, a “modular” pattern is expected (Figure 3C) (74). Local adaptation, whereby phages are most infective to the hosts from their local environment, is also expected to result in a modular pattern (74).

Given that nestedness was found to increase over time with ARD and to decrease over time with FSD (76), nestedness may be used as an indicator of ARD in PBINs. Experimental conditions that increase nestedness may be beneficial in the development of highly infective phages. Crucially however, many nestedness metrics exist, as well as many methods to infer statistically significant nestedness. As described in the next sections, a critical evaluation of these methods is required before drawing conclusions on the nestedness of PBINs.

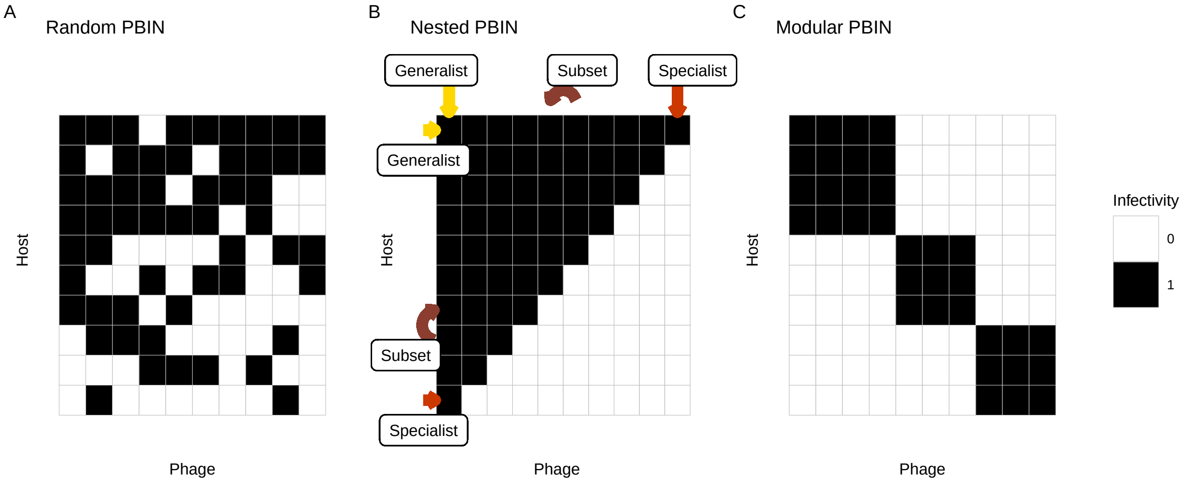

Figure 3: The common patterns examined in PBINs. A) A random network. B) A maximally nested network. The network contains both specialist and generalist hosts and phages. In addition, host and resistance ranges form nested subsets. C) A modular network. Infectivity is confined to modular subsets of the network.

4 Nestedness metrics and null models

4.1 Nestedness metrics

The concept of nestedness first arose to describe a nested occurrence pattern of species among sites (77). The concept has also been used to describe mutualistic networks, such as those formed by plants and pollinators (78). Only later was nestedness applied to the analysis of PBINs (74). Early application of the concept was made possible by The Nestedness Temperature Calculator, which allowed users to quantify the "nestedness temperature" (T) (77). T measures the extent to which infectivity falls in the top triangle of the PBIN and resistance falls in the bottom triangle (79). This is measured through distances to resistance in the top triangle, and infectivity in the bottom triangle, given an isocline that separates the two (79) (Figure 4A).

The T metric comes with limitations: firstly, T accounts for matrix fill, meaning that an extremely empty or full PBIN can be highly nested (80) (Figure 4B,C). However, it is unlikely for a co-evolution experiment following ARD to return networks showing patterns as in Figure 4B and C. Additionally, infective and resistant interactions that are further from the isocline receive more weight in the calculation (i.e. the longer arrow in Figure 4A carries more weight), which is not necessarily justified when analysing an interaction network (77).

While many other metrics were later developed, only one gained popularity in PBIN analysis: Nestedness Measure based on Overlap and Decreasing Fills (NODF) (74, 80). This measure quantifies the extent to which a PBIN shows the following characteristics: a wide range in resistance and host range sizes, overlapping fill in resistance ranges and overlapping fill in host ranges (80). Importantly, these features represent the expectations for ARD co-evolution: hosts and phages acquire broader resistance and infectivity over time, with overlaps in resistance and infectivity as hosts and phages accumulate adaptations. Nevertheless, important results obtained with T in the analysis of PBINs, such as the notion of widespread nestedness in PBINs (75), have not been re-examined using NODF. In the context of phage therapy development, it is crucial that our measure of nestedness closely resembles our expectations for ARD.

Note that T and NODF calculate nestedness given binary infection data. Phage infectivity data is commonly collected through plaque-based assays, which result in plaques or spots on petri dishes if lytic phages have completed the lytic cycle (81). Since presence or absence of a plaque is binary, T or NODF are suitable for the analysis of this data. Nevertheless, infectivity can also be determined using differences in optical density, measured in liquid growth assays (81). At first, it may seem that this data lends itself to analysis with weighted nestedness measures. For example, a weighted version of NODF exists, called WNODF (82). In addition to the characteristics mentioned above for NODF, WNODF measures the extent to which the infectivity in interactions involving generalists is greater than the infectivity in interactions involving specialists (82). In other words, this metric requires that phages acquire weak infectivity to the common host and develop stronger infectivity to that host as co-evolution progresses. This requirement does not fit with the ARD model, in which a phage evolves the ability to infect the host that is currently most abundant, while maintaining its infectivity to the host that was previously most abundant (65). Alternatively, growth data could be binarised for analysis with T or NODF. Given that plaque-based assays are most commonly performed, there are not many examples of binarisation of growth data. Nevertheless, Wright *et al.* set a threshold based on the 5th percentile of a normal distribution (35). Overall, while phage infectivity is unlikely to be binary, PBINs are usually measured in binary form, and the predictions of the ARD model are more in line with the pattern maximised by NODF than WNODF. Future work could explore binarisation methods for liquid growth assays, or network metrics that incorporate continuous measures of growth while being in line with the predictions of ARD.

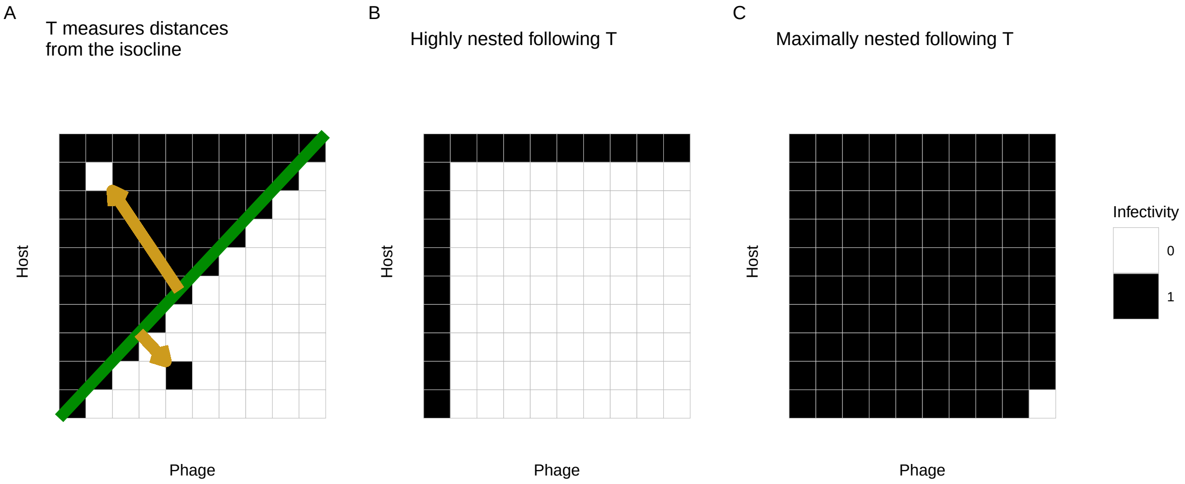

Figure 4: Exemplifying the temperature metric. A) The temperature metric quantifies deviations from "perfect" nestedness. However, perfect nestedness is dependent on matrix fill. Therefore, a highly sparse matrix can be highly nested (B) and a near-full matrix can be maximally nested (C).

4.2 Uniform null model leads to questionable inferences

Nestedness analyses often focus on whether the PBIN is significantly nested (74). In the development of phage therapeutics, significant nestedness may suggest that a broadly infective phage is being developed. Within the patient, significant nestedness may suggest that the pathogen is becoming broadly resistant. To call statistical significance, a standardised effect size (SES) can be calculated, given a particular null model (77). A null model reshuffles the number of infective interactions among the hosts and phages, given certain constraints on the number of interactions that each phage and host can have (77). The SES then measures the extent to which the PBIN is more nested than a collection of random networks generated with the null model.

The choice of null model is very important. If an inappropriate null model is chosen, the comparison of the nestedness of the PBIN and the random networks is non-informative (83). In the context of ARD, GTSGs in infectivity and resistance are expected to form, with overlapping infectivity and resistance ranges, given the escalation in infectivity and resistance. Ideally, the PBIN would be compared to null matrices that preserve any GTSG: thereby, the test would ask whether the PBIN shows greater overlap in resistance and infectivity than expected by chance. In other words, the test would isolate the pattern of interest (83), which is the stepwise growth in resistance and infectivity ranges.

A null model that has been widely used in the analysis of PBINs is EE, also referred to as SIM0, R00 or the "Bernoulli random network" (75, 84–90). This model takes the number of infective interactions in the PBIN and distributes these among the network such that every phage-host pair has the same chance of showing infectivity (77). In other words, if 20% of phage-host pairs show infectivity in the PBIN, the EE null model randomly selects 20% of the phage-host pairs to show infectivity in the null matrix. Therefore, any GTSG is not preserved in the null matrices, and the overlapping fill is no longer isolated in the test (83). More generally, this null model has been shown to have a high Type I error rate, whereby non-nested networks are likely to erroneously be classified as nested (91). Importantly, this null model has been used to infer the widely-cited claim that PBINs are generally nested (75). Given that conclusions reached exclusively with EE are not to be trusted (77), this claim requires re-evaluation with a more robust null model. In the context of phage therapy, if nestedness is to be used as a sign of ARD, it is crucial that the null model allows for reliable distinction of nested and non-nested networks.

4.3 Choosing an alternative null model: an introduction to the Tuning-Peg landscape

The opposite of EE is the FF null model: this model completely fixes the number of interactions each phage and host has (92). Therefore, a test with FF asks whether a PBIN has a stronger pattern of overlap in infectivity and resistance ranges than null networks with the same GTSG. However, FF is known to have a high Type II error, whereby nested networks are likely to be classified as non-nested (91). Therefore, experimental conditions or phages that drive an ARD pattern would be less likely to be discovered when using the FF null model.

With EE and FF sitting at opposite ends in terms of maintaining GTSGs in null networks, there is a whole suite of null models that fall in between (92). Moving away from individual null models, the Tuning-Peg (TP) landscape visualises the effect of varying constraints on the GTSGs of hosts and phages (92). The landscape shows the outcome of many nestedness tests, each using null models with a different level of discrepancy in the GTSGs. Figure 5 shows three such landscapes (right column) associated with different PBINs (left column). The bottom left cell in the landscapes shows the SES of nestedness tests with FF, while the top right cell shows the SES of tests with EE. The cells in between have varying levels of discrepancy in host range sizes (x-axis) and resistance range sizes (y-axis) compared to the PBIN that is being analysed.

TP landscapes show characteristic patterns with certain PBINs. For example, the TP landscape of a nested PBIN shows significant SES values across most of the landscape, while the landscape of a random PBIN shows non-significant SES values across most of the landscape (top and middle rows in Figure 5). An SES greater than 1.64 is considered significant, as this is the 95th percentile of a standard normal distribution (92). Strona *et al.* found that networks with a GTSG, but without overlapping fill, tended to show a gradient across the landscape (92). Such networks have “heterogeneous marginal totals” (bottom row in Figure 5).

Strona *et al.* revealed two key findings (92): firstly, the TP landscape could be a tool with which to distinguish truly nested networks from those which simply have a GTSG (i.e. nested vs heterogeneous networks). Secondly, the landscape contains null matrices that appeared to outperform EE and FF in their ability to distinguish nested and non-nested networks: the quasi-FF (qFF) null networks, contained in the bottom left (Figure 5).

In the context of PBIN nestedness analyses, qFF would allow for a comparison of a PBIN’s nestedness to networks with very similar GTSGs, thereby closely matching the requirements of the test for ARD. Additionally, visualisation of TP landscapes may allow researchers to distinguish PBINs with a true nested pattern from PBINs that only contain a GTSG. Nevertheless, the performance of qFF has not been tested on PBINs or systematically compared to EE. If qFF is found to outperform EE, it could be adopted in future studies of ARD-promoting experimental conditions and in assessments of co-evolutionary dynamics within patients undergoing phage therapy.

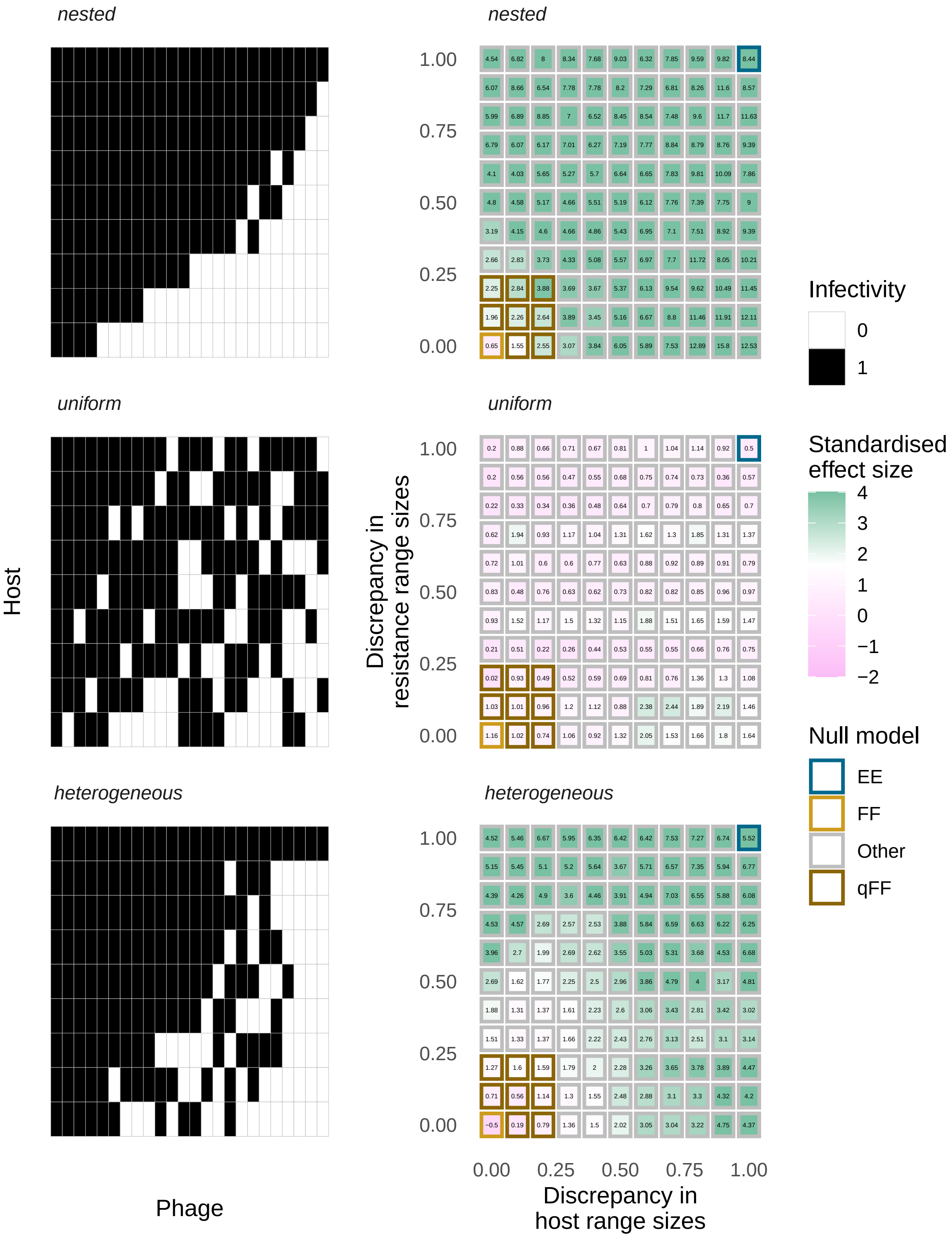

Figure 5: Exemplifying the TP landscape. The left column shows examples of nested, uniform and heterogeneous PBINs. The right column shows the associated TP landscapes. The nested PBIN is significantly nested (SES > 1.64, the 95th percentile of a standard normal distribution) across the whole landscape, except for the cell with FF null matrices. The uniform PBIN is not significantly nested across most of the landscape. The heterogeneous PBIN is only significantly nested compared to null matrices that are further away from the bottom left corner. Notice that the collection of qFF matrices consider the nested matrix nested, while they consider the heterogeneous matrix non-nested.

5 Thesis overview

The proposed work investigates the interactions between bacteria and phages through experimental, bioinformatic and statistical means, partially with an emphasis on clinical CF *P. aeruginosa.* In the first data chapter, the resistance of clinical CF *P. aeruginosa* isolates is measured against environmental phages. Phage resistance is associated with receptor biosynthesis variants and defence system repertoires. This study suggests that in the studied isolate collection, accumulation of LPS variants associates with resistance breadth towards LPS phages, while similarity in defence system repertoire associates with similarity in resistance specificity. In the second data chapter, the defence system pangenome of a large collection of clinical CF *P. aeruginosa* isolates is characterised. This work suggests that while these isolates carry many different defence systems, variation between defence system repertoires is rare between closely-related isolates and largely the norm between more distantly related isolates. Additionally, the within-patient repertoire appears to remain largely stable through time. In the final data chapter, the performance of the commonly used EE null model is compared to that of the qFF null model on simulated and empirical PBINs. This work suggests that qFF is more robust than EE, and that the majority of empirical PBINs studied are only significantly nested when using the problematic EE null model. Overall, this thesis suggests that it is important to consider defence system repertoires in light of isolate's receptor variants, that clinical *P. aeruginosa* isolates may show different defence system dynamics than seen in previous studies of other isolate collections and that conclusions reached previously in PBIN nestedness analyses using EE need to be re-evaluated.

6 References

1. Elborn JS (2016) Cystic fibrosis. *The Lancet* 388(10059):2519–2531.

2. Scotet V, Gutierrez H, Farrell PM (2020) Newborn Screening for CF across the Globe—Where Is It Worthwhile? *International Journal of Neonatal Screening* 6(1):18.

3. Guo J, Garratt A, Hill A (2022) Worldwide rates of diagnosis and effective treatment for cystic fibrosis. *Journal of Cystic Fibrosis* 21(3):456–462.

4. CFTR2 CFTR2. Available at: `http://www.cftr2.org/` [Accessed July 29, 2023].

5. Registry CFFP (2022) *2021 Annual Data Report* (Bethesda, Marylan) Available at: `https://www.cff.org/sites/default/files/2021-11/Patient-Registry-Annual-Data-Report.pdf`.

6. Thornton CS, Parkins MD (2023) Microbial Epidemiology of the Cystic Fibrosis Airways: Past, Present, and Future. *Seminars in Respiratory and Critical Care Medicine* 44(02):269–286.

7. Trust CF (2022) *CF Trust Annual Data Report 2021.*

8. Hewer SCL, Smith S, Rowbotham NJ, Yule A, Smyth AR (2023) Antibiotic strategies for eradicating Pseudomonas aeruginosa in people with cystic fibrosis. *Cochrane Database of Systematic Reviews* (6). doi:10.1002/14651858.CD004197.pub6.

9. Crull MR, et al. (2018) Changing Rates of Chronic Pseudomonas aeruginosa Infections in Cystic Fibrosis: A Population-Based Cohort Study. *Clinical Infectious Diseases* 67(7):1089–1095.

10. Smith S, Rowbotham NJ (2022) Inhaled anti-pseudomonal antibiotics for long-term therapy in cystic fibrosis. *Cochrane Database of Systematic Reviews* (11). doi:10.1002/14651858.CD001021.pub4.

11. Parkins MD, Somayaji R, Waters VJ (2018) Epidemiology, Biology, and Impact of Clonal Pseudomonas aeruginosa Infections in Cystic Fibrosis. *Clinical Microbiology Reviews* 31(4):10.1128/cmr.00019–18.

12. Camus L, Vandenesch F, Moreau K (2021) From genotype to phenotype: Adaptations of Pseudomonas aeruginosa to the cystic fibrosis environment. *Microbial Genomics* 7(3):000513.

13. Winstanley C, O'Brien S, Brockhurst MA (2016) Pseudomonas aeruginosa Evolutionary Adaptation and Diversification in Cystic Fibrosis Chronic Lung Infections. *Trends in Microbiology* 24(5):327–337.

14. Despotes KA, Donaldson SH (2022) Current state of CFTR modulators for treatment of Cystic Fibrosis. *Current Opinion in Pharmacology* 65:102239.

15. Elborn JS, Blasi F, Burgel P-R, Peckham D (2023) Role of inhaled antibiotics in the era of highly effective CFTR modulators. *European Respiratory Review* 32(167). doi:10.1183/16000617.0154-2022.

16. Guo J, Wang J, Zhang J, Fortunak J, Hill A (2022) Current prices versus minimum costs of production for CFTR modulators. *Journal of Cystic Fibrosis* 21(5):866–872.

17. Harper DR (2021) Introduction to Bacteriophages. *Bacteriophages: Biology, Technology, Therapy*, eds Harper DR, Abedon ST, Burrowes BH, McConville ML (Springer International Publishing, Cham), pp 3–16.

18. and re-implementation of bacteriophage therapy E round table on acceptance, et al. (2018) Silk Route to the Acceptance and Re-Implementation of Bacteriophage Therapy—Part II. *Antibiotics* 7(2):35.

19. Kortright KE, Chan BK, Koff JL, Turner PE (2019) Phage Therapy: A Renewed Approach to Combat Antibiotic-Resistant Bacteria. *Cell Host & Microbe* 25(2):219–232.

20. Singh J, Yeoh E, Fitzgerald DA, Selvadurai H (2023) A systematic review on the use of bacteriophage in treating Staphylococcus aureus and Pseudomonas aeruginosa infections in cystic fibrosis. *Paediatric Respiratory Reviews.* doi:10.1016/j.prrv.2023.08.001.

21. Law N, et al. (2019) Successful adjunctive use of bacteriophage therapy for treatment of multidrug-resistant Pseudomonas aeruginosa infection in a cystic fibrosis patient. *Infection* 47(4):665–668.

22. Kvachadze L, et al. (2011) Evaluation of lytic activity of staphylococcal bacteriophage Sb-1 against freshly isolated clinical pathogens. *Microbial Biotechnology* 4(5):643–650.

23. Zaldastanishvili E, et al. (2021) Phage Therapy Experience at the Eliava Phage Therapy Center: Three Cases of Bacterial Persistence. *Viruses* 13(10):1901.

24. Chan BK, et al. (2016) Phage selection restores antibiotic sensitivity in MDR Pseudomonas aeruginosa. *Scientific Reports* 6(1):26717.

25. Nordstrom HR, et al. (2022) Genomic characterization of lytic bacteriophages targeting genetically diverse Pseudomonas aeruginosa clinical isolates. *iScience* 25(6). doi:10.1016/j.isci.2022.104372.

26. Castledine M, et al. (2022) Parallel evolution of Pseudomonas aeruginosa phage resistance and virulence loss in response to phage treatment in vivo and in vitro. *eLife* 11:e73679.

27. Chan BK, Stanley G, Modak M, Koff JL, Turner PE (2021) Bacteriophage therapy for infections in CF. *Pediatric Pulmonology* 56(S1):S4–S9.

28. Dy RL, Richter C, Salmond GPC, Fineran PC (2014) Remarkable Mechanisms in Microbes to Resist Phage Infections. *Annual Review of Virology* 1(1):307–331.

29. Hobbs Z, Abedon ST (2016) Diversity of phage infection types and associated terminology: The problem with "Lytic or lysogenic". *FEMS Microbiology Letters* 363(7):fnw047.

30. Howard-Varona C, Hargreaves KR, Abedon ST, Sullivan MB (2017) Lysogeny in nature: Mechanisms, impact and ecology of temperate phages. *The ISME Journal* 11(7):1511–1520.

31. Smith WPJ, Wucher BR, Nadell CD, Foster KR (2023) Bacterial defences: Mechanisms, evolution and antimicrobial resistance. *Nature Reviews Microbiology*:1–16.

32. Harper DR (2021) Introduction to Bacteriophages. *Bacteriophages: Biology, Technology, Therapy*, eds Harper DR, Abedon ST, Burrowes BH, McConville ML (Springer International Publishing, Cham), pp 3–16.

33. Patel PH, Maxwell KL (2023) Prophages provide a rich source of antiphage defense systems. *Current Opinion in Microbiology* 73:102321.

34. Georjon H, Bernheim A (2023) The highly diverse antiphage defence systems of bacteria. *Nature Reviews Microbiology*:1–15.

35. Wright RCT, Friman V-P, Smith MCM, Brockhurst MA (2018) Cross-resistance is modular in bacteria–phage interactions. *PLOS Biology* 16(10):e2006057.

36. Harvey H, et al. (2018) Pseudomonas aeruginosa defends against phages through type IV pilus glycosylation. *Nature Microbiology* 3(1):47–52.

37. Carballo-Ontiveros MA, Cazares A, Vinuesa P, Kameyama L, Guarneros G (2020) The Concerted Action of Two B3-Like Prophage Genes Excludes Superinfecting Bacteriophages by Blocking DNA Entry into Pseudomonas aeruginosa. *Journal of Virology* 94(15):10.1128/jvi.00953–20.

38. Tesson F, et al. (2022) Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nature Communications* 13(1):2561.

39. Johnson MC, et al. (2023) Core defense hotspots within Pseudomonas aeruginosa are a consistent and rich source of anti-phage defense systems. *Nucleic Acids Research*:gkad317.

40. Dimitriu T, Szczelkun MD, Westra ER (2020) Evolutionary Ecology and Interplay of Prokaryotic Innate and Adaptive Immune Systems. *Current Biology* 30(19):R1189–R1202.

41. Costa AR, et al. (2023) Accumulation of defense systems in phage resistant strains of Pseudomonas aeruginosa. doi:10.1101/2022.08.12.503731.

42. Huiting E, et al. (2023) Bacteriophages inhibit and evade cGAS-like immune function in bacteria. *Cell* 186(4):864–876.e21.

43. Millman A, et al. (2020) Bacterial Retrons Function In Anti-Phage Defense. *Cell* 183(6):1551–1561.e12.

44. Gao LA, et al. (2022) Prokaryotic innate immunity through pattern recognition of conserved viral proteins. *Science* 377(6607):eabm4096.

45. Hochhauser D, Millman A, Sorek R (2023) The defense island repertoire of the Escherichia coli pan-genome. *PLOS Genetics* 19(4):e1010694.

46. Doron S, et al. (2018) Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* 359(6379). doi:10.1126/science.aar4120.

47. Rousset F, et al. (2022) Phages and their satellites encode hotspots of antiviral systems. *Cell Host & Microbe* 30(5):740–753.e5.

48. Rocha EPC, Bikard D (2022) Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLOS Biology* 20(1):e3001514.

49. LeGault KN, et al. (2021) Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. *Science*. doi:10.1126/science.abg2166.

50. Haudiquet M, Sousa JM de, Touchon M, Rocha EPC (2022) Selfish, promiscuous and sometimes useful: How mobile genetic elements drive horizontal gene transfer in microbial populations. *Philosophical Transactions of the Royal Society B: Biological Sciences* 377(1861):20210234.

51. León LM, Park AE, Borges AL, Zhang JY, Bondy-Denomy J (2021) Mobile element warfare via CRISPR and anti-CRISPR in Pseudomonas aeruginosa. *Nucleic Acids Research* 49(4):2114–2125.

52. Pinilla-Redondo R, et al. (2020) Type IV CRISPR–Cas systems are highly diverse and involved in competition between plasmids. *Nucleic Acids Research* 48(4):2000–2012.

53. Westra ER, et al. (2015) Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Current Biology* 25(8):1043–1049.

54. Makarova KS, Wolf YI, Koonin EV (2013) Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Research* 41(8):4360–4377.

55. Vassallo C, Doering C, Littlehale ML, Teodoro G, Laub MT (2022) Mapping the landscape of anti-phage defense mechanisms in the E. Coli pangenome. doi:10.1101/2022.05.12.491691.

56. Payne LJ, et al. (2021) Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Research* (gkab883). doi:10.1093/nar/gkab883.

57. Boon M, Holtappels D, Lood C, Noort V van, Lavigne R (2020) Host Range Expansion of Pseudomonas Virus LUZ7 Is Driven by a Conserved Tail Fiber Mutation. *PHAGE*. doi:10.1089/phage.2020.0006.

58. Le S, et al. (2013) Mapping the Tail Fiber as the Receptor Binding Protein Responsible for Differential Host Specificity of Pseudomonas aeruginosa Bacteriophages PaP1 and JG004. *PLOS ONE* 8(7):e68562.

59. Borin JM, et al. (2023) Rapid bacteria-phage coevolution drives the emergence of multi-scale networks. doi:10.1101/2023.04.13.536812.

60. Rusinov IS, Ershova AS, Karyagina AS, Spirin SA, Alexeevski AV (2018) Avoidance of recognition sites of restriction-modification systems is a widespread but not universal anti-restriction strategy of prokaryotic viruses. *BMC Genomics* 19(1):885.

61. Samson JE, Magadán AH, Sabri M, Moineau S (2013) Revenge of the phages: Defeating bacterial defences. *Nature Reviews Microbiology* 11(10):675–687.

62. Lopatina A, Tal N, Sorek R (2020) Abortive Infection: Bacterial Suicide as an Antiviral Immune Strategy. *Annual Review of Virology* 7(1):null.

63. Davidson AR, et al. (2020) Anti-CRISPRs: Protein Inhibitors of CRISPR-Cas Systems. *Annual Review of Biochemistry* 89(1):309–332.

64. Gaba S, Ebert D (2009) Time-shift experiments as a tool to study antagonistic coevolution. *Trends in Ecology & Evolution* 24(4):226–232.

65. Buckling A, Rainey PB (2002) Antagonistic coevolution between a bacterium and a bacteriophage. *Proceedings of the Royal Society of London Series B: Biological Sciences* 269(1494):931–936.

66. Gomez P, Buckling A (2011) Bacteria-Phage Antagonistic Coevolution in Soil. *SCIENCE* 332(6025):106–109.

67. Betts A, Kaltz O, Hochberg ME (2014) Contrasted coevolutionary dynamics between a bacterial pathogen and its bacteriophages. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA* 111(30):11109–11114.

68. Lopez-Pascua L, et al. (2014) Higher resources decrease fluctuating selection during host–parasite coevolution. *Ecology Letters* 17(11):1380–1388.

69. Hall AR, Scanlan PD, Morgan AD, Buckling A (2011) Host–parasite coevolutionary arms races give way to fluctuating selection. *Ecology Letters* 14(7):635–642.

70. Betts A, Vasse M, Kaltz O, Hochberg ME (2013) Back to the future: Evolving bacteriophages to increase their effectiveness against the pathogen Pseudomonas aeruginosa PAO1. *Evolutionary Applications* 6(7):1054–1063.

71. Burrowes BH, Molineux IJ, Fralick JA (2019) Directed in Vitro Evolution of Therapeutic Bacteriophages: The Appelmans Protocol. *VIRUSES-BASEL* 11(3). doi:10.3390/v11030241.

72. Brockhurst MA, Koskella B, Zhang Q-G (2021) Bacteria-Phage Antagonistic Coevolution and the Implications for Phage Therapy. *Bacteriophages: Biology, Technology, Therapy*, eds Harper DR, Abedon ST, Burrowes BH, McConville ML (Springer International Publishing, Cham), pp 231–251.

73. Brockhurst MA, Koskella B (2013) Experimental coevolution of species interactions. *Trends in Ecology & Evolution* 28(6):367–375.

74. Weitz JS, et al. (2013) Phage–bacteria infection networks. *Trends in Microbiology* 21(2):82–91.

75. Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS (2011) Statistical structure of host–phage interactions. *Proceedings of the National Academy of Sciences* 108(28):E288–E297.

76. Fortuna MA, et al. (2019) Coevolutionary dynamics shape the structure of bacteria-phage infection networks. *EVOLUTION* 73(5):1001–1011.

77. Ulrich W (2009) Ecological interaction networks: Prospects and pitfalls. *Ecological Questions* 11(1):17–25.

78. Bascompte J, Jordano P, Melián CJ, Olesen JM (2003) The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences* 100(16):9383–9387.

79. Almeida-Neto M, Jr PRG, Lewinsohn TM (2007) On nestedness analyses: Rethinking matrix temperature and anti-nestedness. *Oikos (Copenhagen, Denmark)* 116(4):716–722.

80. Almeida-Neto M, Guimarães P, Guimarães Jr PR, Loyola RD, Ulrich W (2008) A consistent metric for nestedness analysis in ecological systems: Reconciling concept and measurement. *Oikos (Copenhagen, Denmark)* 117(8):1227–1239.

81. Yerushalmy O, et al. (2023) Towards Standardization of Phage Susceptibility Testing: The Israeli Phage Therapy Center "Clinical Phage Microbiology"—A Pipeline Proposal. *Clinical Infectious Diseases* 77(Supplement_5):S337–S351.

82. Almeida-Neto M, Ulrich W (2011) A straightforward computational approach for measuring nestedness using quantitative matrices. *Environmental Modelling & Software* 26(2):173–178.

83. Moore JE, Swihart RK (2007) Toward Ecologically Explicit Null Models of Nestedness. *Oecologia* 152(4):763–778.

84. Flores CO, Valverde S, Weitz JS (2013) Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *The ISME Journal* 7(3):520–532.

85. Van Cauwenberghe J, et al. (2021) Spatial patterns in phage- Rhizobium coevolutionary interactions across regions of common bean domestication. *The ISME Journal*:1–15.

86. Shaer Tamar E, Kishony R (2022) Multistep diversification in spatiotemporal bacterial-phage coevolution. *Nature Communications* 13(1):7971.

87. Gurney J, et al. (2017) Network structure and local adaptation in co-evolving bacteria-phage interactions. *MOLECULAR ECOLOGY* 26(7, SI):1764–1777.

88. Beckett SJ, Williams HTP (2013) Coevolutionary diversification creates nested-modular structure in phage-bacteria interaction networks. *Interface Focus* 3(6):20130033.

89. Gupta A, et al. (2022) Leapfrog dynamics in phage-bacteria coevolution revealed by joint analysis of cross-infection phenotypes and whole genome sequencing. *Ecology Letters* 25(4):876–888.

90. Larsen ML, Wilhelm SW, Lennon JT (2019) Nutrient stoichiometry shapes microbial coevolution. *ECOLOGY LETTERS* 22(6):1009–1018.

91. Ulrich W, Gotelli NJ (2007) Null Model Analysis of Species Nestedness Patterns. *Ecology* 88(7):1824–1831.

92. Strona G, Ulrich W, Gotelli NJ (2018) Bi-dimensional null model analysis of presence-absence binary matrices. *Ecology* 99(1):103–115.

Chapter 2: Estimating the association between phage resistance and its determinants in clinical Cystic Fibrosis *Pseudomonas aeruginosa* isolates

Abstract

To develop resistance to an infective phage, bacterial hosts need to either modify their receptors or possess an effective defense system. This study investigated the connections between bacterial growth in the presence of phages and these two methods of phage resistance, in *P. aeruginosa* isolates obtained from the Danish Cystic Fibrosis population. Bacterial isolates with a greater number of lipopolysaccharide (LPS) variants showed increased resistance to LPS phages, suggesting that LPS variants contribute to resistance breadth. Additionally, isolates with greater similarity in defence system repertoire had greater similarity in phage resistance range, suggesting that defence systems contribute to resistance specificity. The study also identified potential factors contributing to resistance against specific phages. A particular variant in the *rmlA* gene appeared to associate with generalist resistance to LPS phages, while a variant in the *fimT* gene associated with specialist resistance against phage phiKZ. Although the presence of the druantia defense system was associated with resistance to PNM phages, an adsorption assay suggested that the bacterial isolates resisted PNM through receptor modification. Overall, this work suggests that in this collection of bacteria and phages, receptor modifications and defence systems play complementary roles in determining phage resistance.

1 Introduction

Faced with phage infection, bacterial populations can evolve resistance through two modes: surface-based resistance and defence system-based immunity (Figure 1). Surface-based resistance alters the phage receptor, impairing entry into the host. Conversely, defence system immunity prevents the entered phage from completing its infection cycle. Surface-based resistance is often observed to evolve under laboratory conditions and may be associated with resistance breadth (1–6). Nevertheless, bacteria are known to carry a diverse set of defence systems (7), and some studies have associated fluctuations in their abundance with phage resistance in natural environments (8, 9). Therefore, defence systems may be associated with resistance specificity. Additionally, the fitness costs associated with these modes of resistance are thought to depend on phage abundance, e.g. through resource availability (1), as well as the presence of host competitors in the environment (10). While it is clear that both modes of resistance can evolve, it is largely unknown what modes bacteria generally use, when a sample is taken from a natural environment. Is resistance mainly mediated by intracellular immunity, as predicted by the costs of surface-based modifications, or are hosts still able to carry surface modifications in the natural environment?

Previous work on the interactions between bacteria and phages has often considered individual phage-host pairs. However, individual bacterial and phage strains can also be considered as members of a larger network of bacteria and phages, known as phage bacteria infection networks (PBINs) (11). To construct a PBIN, the infectivity of every sampled phage is tested against every sampled host. Given host genome sequences, associations can be inferred between resistance determinants and phage resistance. For example, LeGault *et al.* and Hussain *et al.* discovered modules in PBINs from natural environments, where resistance between modules associated with presence of variable genomic regions containing defence system genes (8, 9). In contrast, despite finding associations in CRISPR spacers matching co-circulating phages, Laanto *et al.* were unable to associate CRISPR spacer repertoire with phage resistance (12). Nevertheless, adsorption also did not associate with resistance, highlighting the need to map resistance determinants beyond receptor modifications and CRISPR spacers. Additionally, given a genomically diverse set of hosts, mapping both receptor and defence system components is required to obtain a full picture of the variation in phage resistance determinants. Host diversity within a sample reduces the chances of finding a single determinant of resistance, however it opens up the opportunity to compare the relative importance of the two modes of resistance.

Beyond the importance for the evolutionary and ecological understanding of phage-bacteria interactions, understanding the natural modes of resistance is also important for phage therapy. Phage therapy relies on therapeutic phages being able to infect the pathogenic host (13). For this, an understanding of the modes of resistance carried in the natural environment would be useful. Additionally, the mode of co-evolution within the patient, following administration of the therapeutic phage, may well depend on the modes of resistance carried by the pathogen at the start of therapy. One heavily researched pathogen in the context of phage therapy is *Pseudomonas aeruginosa*, a bacterium that can chronically infect the lungs of Cystic Fibrosis (CF) patients. In a chronic infection the pathogen is found in the majority of patient samples over a minimum period of 12 months (14). Patients tend to acquire *P. aeruginosa* in childhood or young adulthood, with the prevalence in adults over 25 being estimated as greater than 70% (15, 16). Over the patient's life, the pathogen is thought to undergo many genetic and phenotypic changes (17). For example, a study of 474 *P. aeruginosa* isolates obtained from CF patients identified mutations in genes associated with increased antibiotic resistance, a shift towards a biofilm lifestyle and reduced virulence factor production (18). Phage therapy might prove useful in the treatment of CF, as phages could be used in combination with antibiotics to induce fitness trade-offs in resistance evolution (19). While the literature lacks data from completed phage therapy trials in CF patients, individual patient reports have suggested that phage therapy can improve patient health (20, 21).

Despite the importance of phage resistance in CF phage therapy, the modes of resistance employed by CF pathogenic isolates are currently unknown. The research literature is rich in phage host range tests on *P. aeruginosa* (22–25), in some cases using clinical isolates (26–29). These studies show that *P. aeruginosa* varies widely in its resistance to phages. While it is known that the presence of intact host receptors can determine *P. aeruginosa* phage susceptibility (4, 30), an important additional factor is the presence of phage defence systems, of which *P. aeruginosa* is known to carry many kinds (31–33). Previous work has suggested that 32-68% of phage resistance, in a panel of 32 *P. aeruginosa* hosts, is driven by defence systems (32). However this work did not attempt to characterise the receptor modifications that may have driven resistance in the remaining cases. Moreover, while the authors found a positive association between the number of defence systems and phage resistance breadth, the corresponding relationship for receptor modifications and phage resistance was left uncovered. Therefore, there is scope to estimate the associations between both defence system components and receptor modifications, and resistance, on the same isolate collection. Additionally, understanding the resistance determinants in clinical *P. aeruginosa* isolates may also inform the development of phage therapy.

The present report describes an infectivity assay using 27 phages, mostly isolated from hospital sewage waters (4), against 24 *P. aeruginosa* hosts. These hosts represent 23 clone types identified by Marvig *et al.* (18), with the addition of PAO1. The aim of this study was to associate phage resistance with properties of the hosts and phage. Moreover, this study aimed to consider both defence system presence and variants in receptor biosynthesis genes. Firstly, phage resistance was studied in light of the phage's predicted host receptors. Then, associations were estimated between pairwise distances in either receptor biosynthesis genetic variants or defence system repertoire, and isolate's full resistance range. In addition, the number of biosynthesis gene variants or defence systems was associated with overall phage resistance. Moreover, associations were estimated between the presence of individual genetic variants or defence systems and resistance to individual phages. Finally, for a subset of isolates that appeared to carry a functional druantia system, an adsorption assay was performed to rule out receptor-mediated defence. The clinical *P. aeruginosa* isolates were found to vary widely in their phage susceptibility. Furthermore, the 23 isolates were found to carry variants across most receptor biosynthesis genes. Additionally, the isolates carried defence systems belonging to a total of 59 different system families. Isolates with more variants in LPS biosynthesis genes were found to have higher resistance to LPS phages. Additionally, isolates with greater similarity in defence system repertoire had greater similarity, on average, in phage resistance range. Finally, this work identified *rmlA*, *fimT* and the druantia defence system as potential drivers of phage resistance. However, when testing the involvement of druantia through an adsorption assay, it was found that the mode of resistance was more likely to be receptor-mediated. Overall, this work suggests that in this collection of bacteria and phages, receptor modifications are associated with resistance breadth while defence system carriage is associated with resistance specificity.

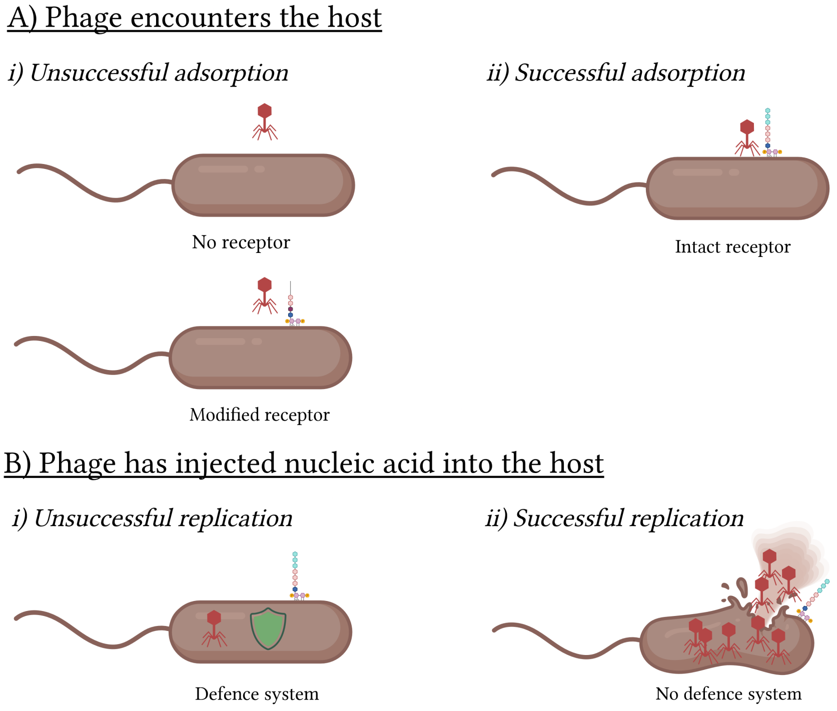
A) Phage encounters the host
i) Unsuccessful adsorption
ii) Successful adsorption
No receptor
Intact receptor
Modified receptor
B) Phage has injected nucleic acid into the host
i) Unsuccessful replication
ii) Successful replication
Defence system
No defence system

Figure 1: Broadly speaking, phages have two requirements for successful infection and propagation. Firstly, phages require the host to carry an appropriate receptor for binding. Secondly, in addition to compatible replication machinery, phages require that the host does not carry a defence system that can shut down their infection.

2 Methods

2.1 Isolates used in this work

The bacterial isolates used in this work were isolated and sequenced by Marvig *et al* (18), with the exception of PAO1. Isolates were originally grouped into separate clone types if they differed by at least 10,000 SNPs (18). The isolates used in the assay were chosen to represent separate clone types (Table 1). Thereby, the isolates would potentially carry a wide range of phage defence systems. Although Marvig *et al.* identified 53 clone types, the local frozen collection was limited to isolates representing 23 clone types. While the isolates were also chosen to be the first of their clone type isolated from a patient, some isolates originated from the same patient. PAO1 was included, as it was a known susceptible host to the phages used in this assay (4). PAO1 therefore acted as a positive control and as the replication host in the infectivity assay.

Table 1: The bacterial isolates used in this work. Isolates 1-23 were isolated by Marvig *et al* (18). While every isolate represented a separate clone type, some isolates originated from the same patient. Isolates were grouped by patient and arranged by isolation date in this table.

Isolate ID	Patient	Date	Genotype
4	CF236	2006-08-29	DK15
23	CF236	2006-08-29	DK53
16	CF382	2007-02-19	DK32
3	CF405	2005-10-11	DK13
2	CF408	2006-08-05	DK09
7	CF422	2005-05-23	DK21
8	CF422	2009-12-01	DK22
9	CF422	2010-02-03	DK23
10	CF422	2012-08-08	DK24
6	CF455	2007-12-17	DK19
5	CF455	2011-12-07	DK18
18	CF472	2005-05-18	DK47
17	CF472	2007-01-16	DK38
19	CF472	2010-08-24	DK48
20	CF472	2011-07-02	DK49
12	CF496	2004-11-24	DK27
13	CF496	2006-03-21	DK28
14	CF496	2007-05-29	DK29
15	CF496	2007-05-29	DK30
1	CF499	2006-03-15	DK06
22	CF499	2007-07-24	DK51
21	CF499	2009-08-07	DK50
11	CF499	2012-11-20	DK26
PAO1	Non-CF patient	NA	PAO1

The phages used in this assay came from different sources (Table 2). Wright *et al.* predicted the host receptors used by these phages through infectivity assays on PAO1 with various receptor mutations (4). While the majority were predicted to bind to lipopolysaccharide (LPS), four phages were predicted to bind to Type IV pilus (T4P) by their inability to infect an unpilliated PAO1 host.

Table 2: The phages used in this study. Receptors were predicted by Wright *et al* (4).

Strain	Receptor	Morphotype	Reference
phiKZ	T4P	Myoviridae	(34)
PNM	T4P	Podoviridae	(34)
PT7	T4P	Myoviridae	(34)

Strain	Receptor	Morphotype	Reference
14/1	LPS	Myoviridae	(34)
PA1P1	LPS	-	(4)
PA1P2	LPS	-	(4)
PA1P3	LPS	-	(4)
PA1P4	LPS	-	(4)
PA1P5	LPS	-	(4)
PA2P1	LPS	-	(4)
PA4P2	LPS	-	(4)
PA5P1	LPS	-	(4)
PA5P2	T4P	-	(4)
PA7P1	LPS	-	(4)
PA7P2	LPS	-	(4)
PA8P1	LPS	-	(4)
PA8P2	LPS	-	(4)
PA10P1	LPS	-	(4)
PA10P2	LPS	-	(4)
PA10P3	LPS	-	(4)
PA11P1	LPS	-	(4)
PA11P2	LPS	-	(4)
PA12P1	LPS	-	(4)
PA12P2	LPS	-	(4)
PA13P1	LPS	-	(4)
PA13P2	LPS	-	(4)
PA14P2	LPS	-	(4)

2.2 Infectivity assay

While phage infectivity data is usually collected through spot tests, phage infectivity can also be tested by comparing absorbance readings in the presence and absence of phage in liquid media (4). This has the advantages of being easier to scale up when using many bacterial and phage strains. In addition, more detail can be obtained on the underlying growth dynamics than when observing spots on a soft overlay agar plate. Below, the methodology of the infectivity assay is described. Unless stated otherwise, all bacterial growth was performed at 37 °C, 200 RPM, in Synthetic Cystif Fibrosis Medium (SCFM) (35).

2.2.1 SCFM preparation

SCFM was created in two stages. First, a base media was prepared (“SCFM base”), which could be stored for two weeks. When SCFM was required, a set of nutrients were added to SCFM base on the day of the assay. The original recipe was prepared by Michael Bottery and the procedure was subsequently adapted for this work. To create the base media, the compounds in Table 3 were added to dH_2O. Additionally, the three nutrients in Table 4 were dissolved in sodium hydroxide prior to addition to the base. The solution was subsequently brought to pH 6.8 with HCl, filter sterilised, wrapped in foil and stored at 4 °C for up to two weeks.

Table 3: Nutrients that were added directly to dH_2O in creating SCFM base. Note that ‘mass to add (g)’ assumes 1000 mL SCFM base is being prepared.

chemical unit	name	Molecular weight (g/mol)	final concentration (mM)	mass to add (g)
NaH2PO4	sodium phospate monobasic	120.0	1.300	0.1560
Na2HPO4	sodium phospate dibasic	142.0	1.250	0.1770
KNO3	Potassium nitrate	101.0	0.348	0.0352
K2SO4	Potassium sulfate	174.0	0.271	0.0472
NH4Cl	Ammonium chloride	53.5	2.280	0.1220

chemical unit	name	Molecular weight (g/mol)	final concentration (mM)	mass to add (g)
KCl	Potassium chloride	74.6	14.900	1.1100
NaCl	Sodium chloride	58.4	51.800	3.0300
MOPS	MOPS buffer	209.0	10.000	2.0900
Ser	Serine	105.0	1.450	0.1520
Glu * HCl	Glutamic acid hydrochloride	184.0	1.550	0.2840
Pro	Proline	115.0	1.660	0.1910
Gly	Glycine	75.1	1.200	0.0903
Ala	Alanine	89.1	1.780	0.1590
Val	Valine	117.0	1.120	0.1310
Met	Methionine	149.0	0.633	0.0944
Ile	Isoleucine	131.0	1.120	0.1470
Leu	Leucine	131.0	1.610	0.2110
Orn * HCl	Ornothine hydrochloride	169.0	0.676	0.1140
Lys * HCl	Lysine hydrochloride	183.0	2.130	0.3890
Arg * HCl	Arginine hydrochloride	211.0	0.306	0.0645
Thr	Threonine	119.0	1.070	0.1280
Cys * HCl	Cysteine hydrochloride	158.0	0.160	0.0252
Phe	Phenylalanine	165.0	0.530	0.0876
His * HCl	Histidine hydrochloride	210.0	0.519	0.1090

Table 4: Nutrients that were dissolved in NaOH solution prior to addition to the SCFM base preparation. Note that ‘volume to add to beaker (mL)’ assumes 1000 mL SCFM base is being prepared.

chemical unit	name	NaOH concentration (M)	stock (M)	stock volume (mL)	Molecular weight (g/mol)	mass to add (g)	volume to add to beaker (mL)	final concentration (mM)
Asp	Aspartic acid	0.5	0.1	20	133	0.266	8.27	0.827
Trp	Tryptophan	0.2	0.1	10	204	0.204	0.13	0.013
Tyr	Tyrosine	1.0	0.1	20	181	0.362	8.02	0.802

To create SCFM on the day of the assay, solutions were prepared of the following nutrients, sterile filtered and added to an aliquot of SCFM base (Table 5). These solutions could be stored at 4 °C for up to 6 weeks, except for the Iron (II) sulfate heptahydrate solution which was always prepared on the day. In addition, frozen aliquots of the following solutions, which were filter sterilised prior to freezing at -20 °C, were thauwed and added to the base (Table 6).

Table 5: Nutrients that were added to an aliquot of SCFM base in creating SCFM. Note that ‘stock to add to SCFM base (uL)’ assumes 100 mL SCFM is being prepared. Also note that except for Iron (II) sulfate heptahydrate, all nutrients were stored as sterilised stocks for up to 6 weeks at 4 °C. The lactic acid solution was also brought to pH 7 with NaOH. A sterile Iron (II) sulfate heptahydrate solution was prepared on the same day as preparing the SCFM.

chemical unit	name	stock (M)	stock volume (mL)	Molecular weight (g/mol)	mass to add (g)	stock to add to SCFM base (uL)	final concentration (mM)	notes
G-glucose	Glucose	1.0000	5	180.0	0.901	300.0	3.0000	

chemical unit	name	stock (M)	stock volume (mL)	Molecular weight (g/mol)	mass to add (g)	stock to add to SCFM base (uL)	final concentration (mM)	notes
L-lactic acid	Lactic acid	1.0000	5	90.1	0.450	930.0	9.3000	pH stock to 7 with NaOH
CaCl2 * H2O	Clacium chloride	1.0000	5	147.0	0.735	175.0	1.7500	
MgCl2	Magnesium chloride	1.0000	5	95.2	0.476	60.6	0.6060	
FeSO4 * 7H2O	Iron (II) sulfate heptahydrate	0.0036	15	278.0	0.015	100.0	0.0036	Prepare freshly

Table 6: Nutrients that were thawed prior to addition to an aliquot of SCFM base in creating SCFM. Note that 'stock to add to SCFM base (uL)' assumes 100 mL SCFM is being prepared. Also note that in preparing the biotin solution, 1M NaOH was added until the biotin dissolved.

chemical unit	name	Stock concentration (mg/L)	Stock volume (mL)	mass to add (g)	volume to add to SCFM base (uL)	final concentration (mg/L)	notes
Thiamine * HCl	Thiamine hydrochloride (B1)	1000	30	0.03	100	1.000	
Niacin	Niacin acid (B3)	1000	30	0.03	120	1.200	
Calcium pantothenate	D-Pantothenic acid hemicalcium salt (B5)	1000	30	0.03	25	0.250	
Biotin	Biotin (B9)	500	40	0.02	1	0.005	1M NaOH added until biotin dissolved

2.2.2 Phage stock preparation

To create fridge stocks of the 27 phages, the phages were grown from frozen stock in the presence of PAO1. To this end, PAO1 was grown overnight in 50 mL Falcon tubes with 10 mL LB. The next day, the overnight cultures were diluted 1:10 in LB and allowed to grow for 60 minutes. After 60 minutes, each phage strain was inoculated into a separate host tube. These tubes were incubated overnight. The next day, the tubes were centrifuged at 5000 RPM for 5 minutes. The cultures were subsequently sterilised using 22 μm filters into separate hard plastic tubes and stored at 4 °C.

2.2.3 Overlay agar plate preparation

To determine phage titres throughout these assays, overlay agar plates were used. These were constructed as follows. PAO1 was grown overnight in LB. Additionally, soft LB agar was liquified and stored overnight in a 60 °C waterbath. The next day, PAO1 was diluted 1:10 in fresh LB and incubated for 3 hours. For the final hour of incubation, the soft LB agar was placed in a 48 °C waterbath. Following incubation, culture was added to the soft agar, at 10 mL PAO1 culture per 0.5 L soft agar. This suspension was gently swirled and poured onto LB hard square agar plates. These plates were allowed to dry for at least 30 minutes before use in phage titre determination.

2.2.4 Phage titre determination

To determine the titre of the phage stocks, plaque assays were performed. Phage fridge stocks were allowed to reach room temperature. Then, these stocks were serially diluted up to a dilution of 10^{-11}. Additionally, one plaque assay plate was prepared for each phage, as described above. Using a multichannel pipette, three 10 μL drops were introduced onto the plate for the dilutions $10^{-4} - 10^{-11}$. The spots were allowed to dry, after which the plates were incubated overnight at 37 °C. The next day, plaques were counted at each countable dilution. These were used to calculate the phage titre, as $\frac{\text{PFU}}{\text{mL}} = \text{mean(no. plaques)} \times 10 \times \text{dilution factor}$.

2.2.5 Phage master plate preparation

To limit variation in phage titre between phage treatments in the infectivity assay, two phage master plates were constructed. Using phage buffer, each phage was diluted to a titre of $7.66 \times 10^5 \frac{\text{PFU}}{\text{mL}}$, in a volume of 1 mL, inside a deep-well plate. Two plates were constructed with the following layout (Table 7). Wells labelled as ‘c’ only contained phage buffer (hereafter referred to as control wells).

Table 7: The phage master plate layout. Wells labelled with ‘c’ only contained phage buffer.

1	2	3	4	5	6	7	8	9	10	11	12
c	c	c	c	c	c	c	c	c	c	c	c
1	1	1	2	2	2	3	3	3	4	4	4
5	5	5	6	6	6	7	7	7	8	8	8
9	9	9	10	10	10	11	11	11	12	12	12
13	13	13	14	14	14	15	15	15	16	16	16
17	17	17	18	18	18	19	19	19	20	20	20
21	21	21	22	22	22	23	23	23	24	24	24
25	25	25	26	26	26	27	27	27	c	c	c

2.2.6 Running the infectivity assay

The infectivity assay was performed across six batches (i.e. separate days). Each bacterial isolate was replicated across two batches and twice within each batch. Additionally, PAO1 was included in all six batches as a positive control. To initiate a batch, the allocated bacterial isolates were grown overnight from frozen stock in 10 mL SCFM, in 50 mL Falcon tubes. The next day, the cultures were normalised to an OD600 (hereafter OD) of 0.2 with sterile water. The cultures were subsequently diluted 1:2000. Based on CFU counts with PAO1, this would bring the titres down to approximately $3.83 \times 10^5 \frac{\text{CFU}}{\text{mL}}$. These diluted cultures were used to prepare two assay plates per bacterial host (i.e. the two replicates per batch). Assay plates were prepared using 96-well plates with moats, to minimise evaporation during the assay. All wells within an assay plate were inoculated with 10 μL diluted host culture and 10 μL phage preparation in 180 μL SCFM. The assay plates were prepared using a Gilson PLATEMASTER 220μL. Absorbance readings were taken at 0, 8, 20 and 27 hours using a BMG Labtech SPECTROstar Nano. Plates were incubated at 37 °C between measurements.

2.2.7 Determining infectivity

To assess whether phages were able to infect bacterial hosts, the growth in the presence of phage was compared to control growth at the 20-hour time point. This time point was chosen because it was the earliest time at which bacteria had substantially grown. Using this data, the increase in OD (ΔOD) since the 0 time point was calculated for each well within a plate. These values were used to calculate relative OD as the ratio of growth in a phage well and the mean growth of control wells for a particular isolate. Overall, this resulted in 12 relative OD values for each phage-host pair, split across four assay plates in two batches. The raw infectivity data graphs were created using the `ggplot2` and `patchwork` packages (36, 37), and are available at (38). The standalone infectivity matrix was visualised using the `ComplexHeatmap` package (39). Rows were clustered by complete hierarchical clustering with `hclust` based on euclidean distances derived with `dist` (40).

2.2.8 Adsorption assay

To assess whether phage PNM was adsorbing and bursting out of hosts of interest, an adsorption assay was performed. This assay was performed in two batches, as measuring adsorption to all isolates in one assay was not feasible. The assay was performed once for isolates 2, 12, 15 and PAO1 and once for isolates 5, 9 and 10. Overnight cultures of the isolates of interest were set up in 15 mL SCFM. The next day, the phage PNM suspension was allowed to reach room temperature. Meanwhile, 18 eppendorf tubes were set aside per host. Three tubes were labelled per time point: 0, 10, 30, 60, 120 and 180 minutes. Disposable 5 mL syringes were attached to 0.45 μm filters and mounted onto these eppendorf tubes. Additionally, the overnight cultures were normalised to an approximate OD of 0.2 using SCFM. These cultures were then diluted 1:10 in SCFM and transferred into three 14 mL Falcon tubes (10 mL per tube). Then, 170 μL PNM phage suspension was added to each tube, for a final MOI of 0.2, and mixed by inversion. At the aforementioned time points, 400 μL samples were transferred into the designated syringes and filtered. Between time points, tubes were incubated in a 37 °C waterbath without shaking. Prior to sampling, tubes were mixed by inversion. Filtered suspensions were serially diluted (up to 10^{-9}) and the phage titre determined as above.

2.3 Bioinformatic analysis

2.3.1 Data source and pipeline construction

Genome sequences for the CF isolates were obtained from the RefSeq database (Bioproject PRJEB5438; (18)). The bioinformatics pipeline was constructed using Snakemake (41) and executed on the university's Viking Cluster.

2.3.2 Assembly and annotation of a PAO1 genome

To assemble a genome of the PAO1 isolate used to grow the phages in this assay, a hybrid assembly approach was taken. Firstly, genomic DNA was isolated using the Qiagen Genomic tip 100/G kit (Cat No. 10243). Sequencing was performed using an Oxford Nanopore Minion flowcell (R9.4.1) by the University of York's Technology Facility. Raw reads were demultiplexed, trimmed and basecalled using Guppy v6.3.9 with the super-accuracy model. Subsequently, reads were filtered using Filtlong (42), assembled using Flye (43) and polished using Medaka (44). Following filtering of trimmed Illumina reads (kindly provided by Michael Bottery) using Fastp (45), the output of Medaka was short-read polished using Polypolish (46). This pipeline is available online at `https://github.com/ezherman/ont-assembly-pipeline`. The assembly was then annotated with Bakta (47).

2.3.3 Determining genetic variants in receptor biosynthesis genes

To find genetic variants in the clinical CF isolates, bam files were created using snippy (48), with PAO1 as the reference strain. Variants were called using Freebayes (49) and annotated using SnpEff (50). Variants were filtered in two stages. First, only variants in LPS or T4P biosynthesis genes were retained using SnpSift (50), with pathway data taken from the Pseudomonas Genome Database (51). File conversion (vcf to tsv) was performed using the `vcf2tsv` from `vcflib` (52). Then, variants were filtered for quality. Firstly, an isolate's variant was required to have a read depth >= 5. Secondly, an overall variant quality of >= 690 was required (corresponding to a threshold of 30, times 23 isolates). Thirdly, variant allele reads were required to make up at least 50% of an isolate's reads for any given locus. Finally, synonymous variants were discarded.

2.3.4 Defence system prediction

Defence systems were predicted using DefenseFinder (31) and PADLOC (53), through the `find-defence-systems` pipeline (`https://github.com/ezherman/find-defence-systems`). This pipeline ensures that hits that require manual verification are excluded, as well as deduplicating and renaming hits. Additionally, abortive infection (abi) systems were excluded from the analysis. MOI values of 2 or lower have previously been associated with abi in abi systems (54–57), which would have prevented protection by abi systems from being observed in this assay.

2.3.5 Associations in pairwise distances

The associations in clusters based on resistance range and repertoires of relevant genetic features were assessed as follows. First, pairwise eucledian distances were calculated in relative OD. Then, pairwise euclidean distances in genetic variants and pairwise Bray-Curtis distances in defence system repertoire were calculated

(58). Whether associations between distances in relative OD and distances in genetic variants/defence system repertoire were positive was tested using one-sided mantel tests (59). Distances were calculated using `dist` in R (40). Heatmaps were visualised using `ComplexHeatmap` (39) and the associations were visualised using `ggplot2` (37).

2.3.6 Associations between phage resistance and the number of genetic variants or defence system families

To associate overall phage resistance with the number of genetic variants or defence system families, linear mixed effects models were fit (60). The models were fit on the raw data, thus there were 12 data points per phage-host combination. Separate models were fit for the number of genetic variants, splitting genetic variants and phages by receptor. The defense system models, as well as the LPS biosynthesis gene models, had three random intercept terms: phage, host and the interaction between phage and host. Instead, the T4P biosynthesis gene models had phage as a fixed effect, given the limited number of T4P phages. Non-parametric bootstrap confidence intervals were constructed using `confint` with 500 simulations (60). Parametric bootstrap confidence intervals for the regression lines were obtained using the `bootMer` function with 500 simulations (60). Parallelisation was achieved with the `parallel`, `doParallel` and `foreach` packages (40, 61, 62).

2.3.7 Determining important LPS variants using elastic net regularisation

To determine which LPS biosynthesis genes were contributing positively to the association between phage resistance and the number of LPS genetic variants, elastic net regularisation was performed using the `glmnet` package (63). Cross-validated elastic net regularisation was performed on 1000 data subsets. Each data subset contained 80% of the observations. The obtained coefficient estimates were visualised using the `ggplot2` package (37).

2.3.8 Associations between phage resistance and genetic variant or defence system presence

To associate genetic variant presence or defence system presence with phage resistance, a linear mixed effects model was fit for each pair of phage and genetic variant or defence system family. These models had relative OD as the outcome variable. The fixed effect was a binary variable, indicating presence of a genetic variant or a defence system family. A random effect for isolate ID was included. To find associations with potential biological significance, the slope parameter p-values were compared to a bonferroni-corrected threshold, based on an alpha of 0.05. Additionally, a minimum allele frequency of 3 was set for biosynthesis gene variants. Separate models were fit for LPS and T4P phages and genetic variants. Models were fit with the `lme4` package (60). P-values for t-tests on the slope parameters were obtained using the `lmerTest` package (64).

3 Results

3.1 Determining phage infectivity

To determine phage infectivity, isolates were cultured with each phage 12 times in SCFM, across two batches and two replicate plates per batch (i.e. three wells per plate). Infectivity was described through measuring relative OD in the presence and absence of phage. Examples of the raw infectivity data are shown for three hosts (Figure 2). These plots exemplify that infectivity patterns were largely consistent between batches, as plots within the same column showed the same pattern. Plots for all isolates are available at (38).

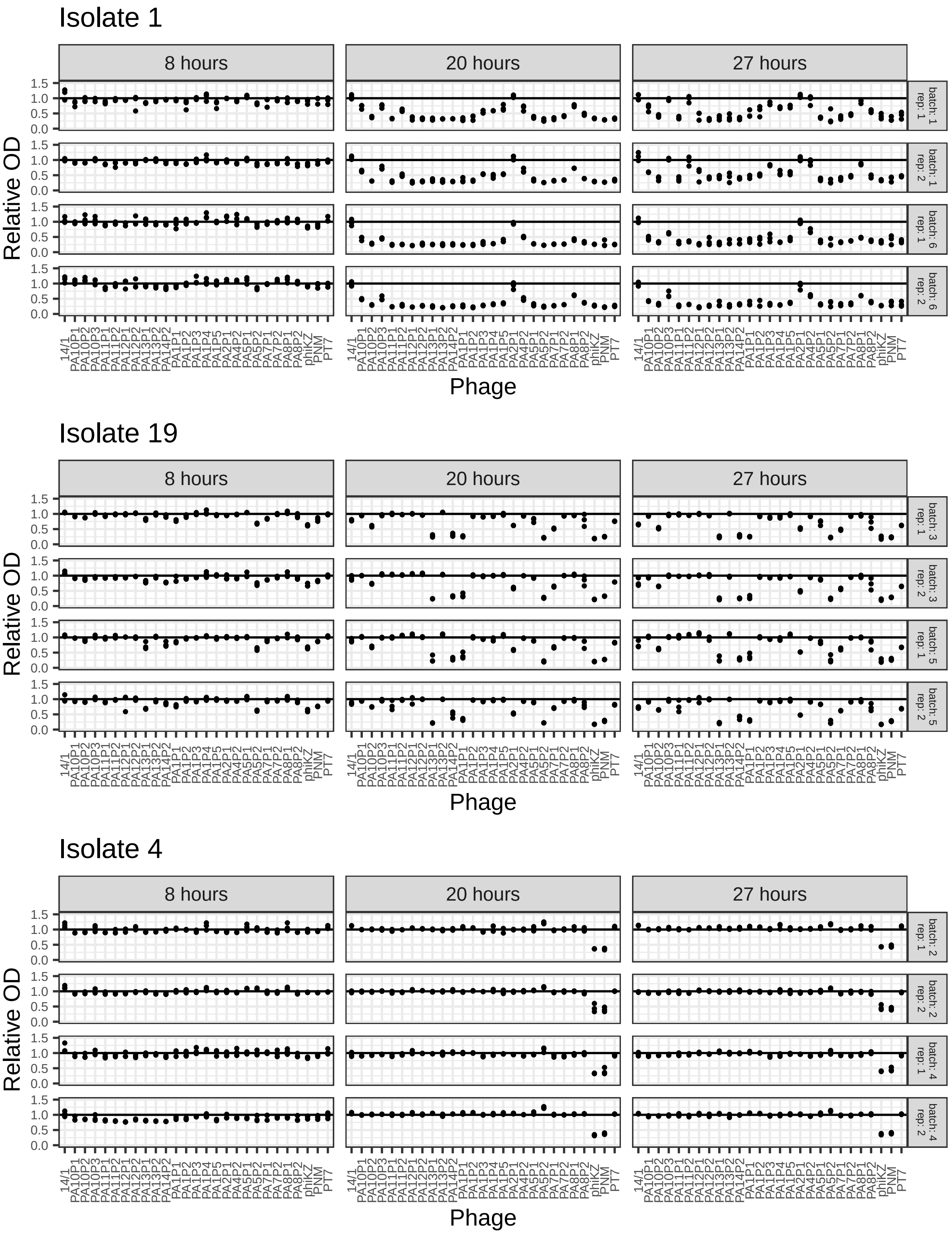

Figure 2: Examples of growth dynamics in the presence of phage, compared to control growth. Each data point shows the relative OD of one well compared to the mean OD of control wells. Qualitatively, isolate 1 was inhibited by most phages, isolate 19 was inhibited by some phages and isolate 4 was not inhibited by most phages. A relative OD below 1 suggests that growth was inhibited by presence of a particular phage. The four panels within each column show data from different replicate plates, where the first number represents batch and the second number represents replicate within that batch. The separate columns show data for different time points, as labelled above each column.

To study the range of isolates that each phage could infect, the mean relative OD was visualised for each phage-host pair (Figure 3). Additionally, isolates were clustered by their resistance range. Finally, phages were divided by their predicted host receptor type (4). The PBIN revealed a generalist to specialist range in resistance, i.e. some isolates resisted all phages, some resisted none and some isolates resisted some but not all phages. For example, the resistance range of isolate 3 appeared similar to that of PAO1, having a low relative OD across most phages. Isolate 21 showed high relative OD with some LPS phages, a lower reltive OD with others and a low relative OD with all T4P phages. Isolates 12, 10 and 15 appeared to be generalists, showing a high relative OD with all the phages used in this study. While T4P phages had a greater infectivity range than LPS phages, too few T4P phages were included in the assay for a formal comparison. Overall, this study included a gradient of generalist to specialist phage resistance.

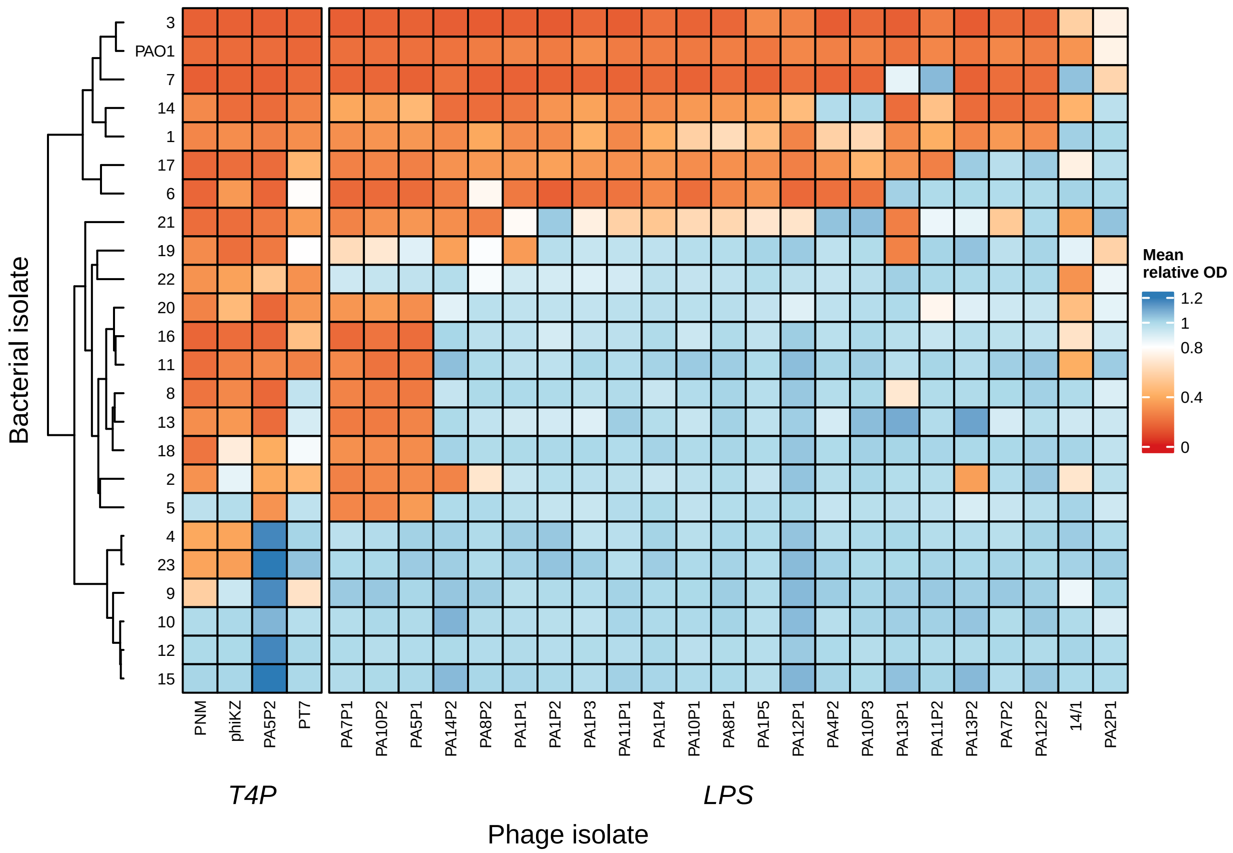

Figure 3: The mean relative OD of isolates in the presence and absence of phage, for each pair of phage and bacterial isolate. Bacterial isolates were clustered by their relative OD pattern. Phages were split by their predicted host receptor. The PBIN showed a gradient of generalist to specialist phage resistance.

3.2 No significant association between pairwise distances in receptor biosynthesis gene variants and resistance range

To ask whether mutations in phage receptor biosynthesis genes might have been driving phage resistance, mutations in these genes were inferred using Snippy (48), Freebayes (49) and SnpEff (50). On average, isolates carried variants in 15 LPS biosynthesis genes (IQR: (8.5; 21.5)) and 8 T4P biosynthesis genes (IQR: (5; 11)). In total, 48 LPS and 24 T4P biosynthesis genes were studied. Therefore, on average isolates carried variants in 32% of biosynthesis genes, compared to the PAO1 strain, which was sensitive to all phages used in this work.

To assess whether LPS biosynthesis variants may have been driving resistance specificity towards LPS phages, isolates were clustered by their variants. These clusters were then imposed on the subset of the PBIN with LPS phages (Figure 4A). To quantify the association, a mantel test was performed on the pairwise distances in relative OD with LPS phages and the pairwise distances in LPS biosynthesis variant counts. The pairwise distances were not significantly associated ($r = -0.085$, 95% CI = $(-0.151, 0.058)$, $p = 0.59$) (Figure 4). Additionally, a significant association could also not be concluded between distance in T4P variants and distance in resistance to T4P phages (Figure 13). Therefore, lack of an association between resistance distances and biosynthesis gene variant distances could not be rejected. Overall, this suggests that when studying divergent bacterial isolates, clustering based on receptor biosynthesis gene variants does not associate with clustering based on phage resistance.

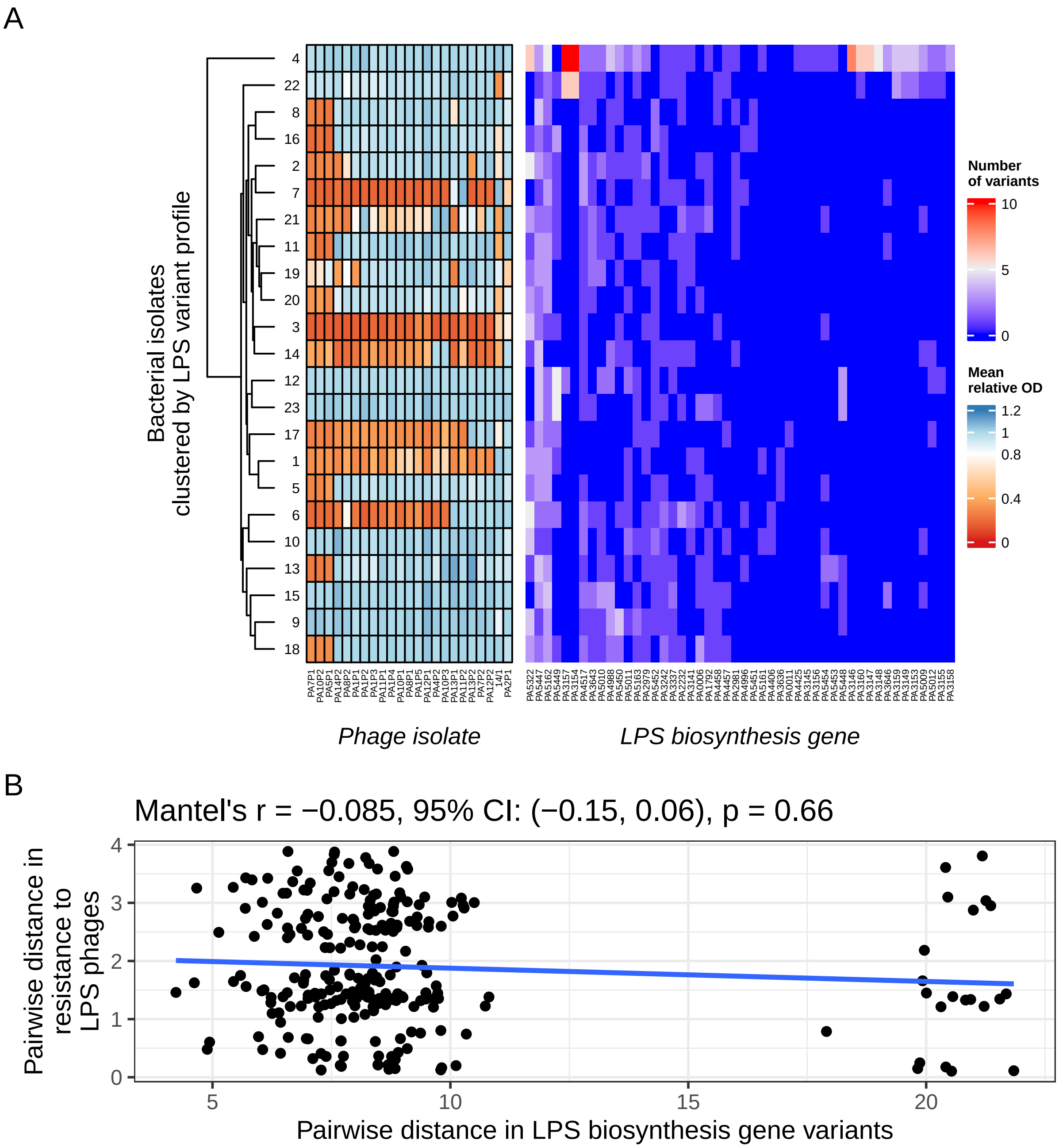

Figure 4: A matrix of LPS biosynthesis variants revealed that none of the isolates shared an identical variants profile. A) Isolates were clustered by their LPS biosynthesis variants, showing that isolates with relatively close variant profiles could differ greatly in their relative OD values with LPS phages. B) A one-sided mantel test did not suggest that a positive association exists between the pairwise distances.

3.3 Isolates with more variants in LPS biosynthesis genes were more resistant to LPS phages

To assess whether isolates with more biosynthesis gene variants tended to have greater phage resistance breadth, the association between relative OD (Figure 3) and the number of biosynthesis gene variants was estimated. This analysis was performed separately for the pairs of LPS/T4P phages and LPS/T4P biosynthesis genes, respectively. The model for T4P phages and genes suggested that the null hypothesis of no association could not be rejected ($\beta = 0.0088, 95\%\ \mathrm{CI} = (-0.00095, 0.018), \mathrm{p} = 0.061$). However, the model for LPS phages suggested that growth of hosts with a greater number of LPS biosynthesis gene variants was less affected by LPS phage presence (Figure 5; $\beta = 0.017, 95\%\ \mathrm{CI} = (0.0019, 0.031), \mathrm{p} = 0.04$). Overall, this data suggests that differences in LPS biosynthesis gene variants in CF isolates could drive differences in phage resistance breadth, with isolates carrying more variants showing on average less growth reduction in the presence of phage.

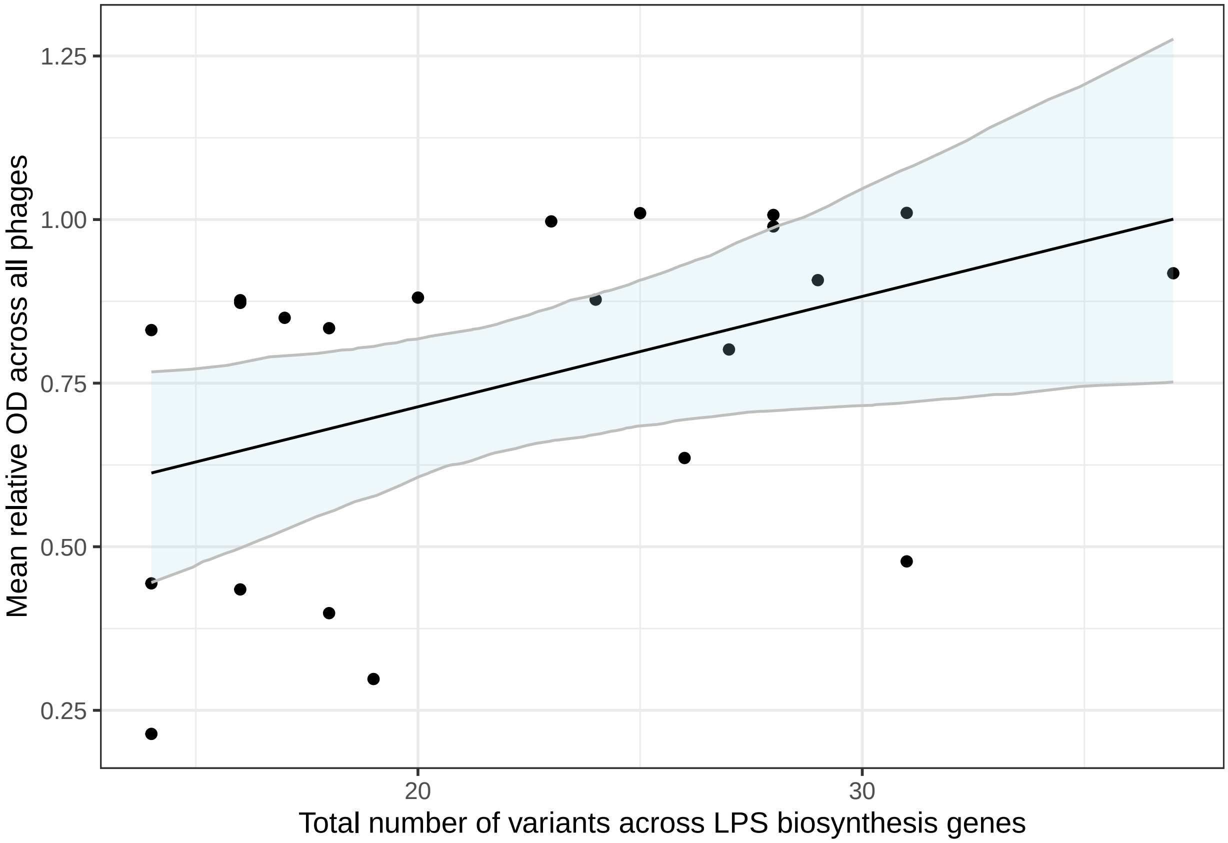

Figure 5: The mean relative OD across phages was significantly associated with the number of variants in LPS biosynthesis genes carried by isolates. Data points show the mean relative OD for each isolate, taken across all phages. The shaded area shows 95% parametric bootstrap intervals. Note that isolate 4 was excluded from this model due to being an outlier, carrying 117 LPS biosynthesis gene variants (Figure 4).

To ask which LPS genes may have driven the association between resistance and number of LPS variants, feature selection was performed using elastic net regularisation. This method aims to eliminate LPS genes from the model that do not associate with resistance, while improving the accuracy of parameter estimates for LPS genes that are associated with resistance. To assess the stability of this feature selection, regularisation was performed on 1000 random subsets of the data, each containing 80% of the data. Genes that were estimated to positively associate with phage resistance in more than 75% of data subsets were considered potentially important for resistance. In total, 18 out of 48 LPS genes met this threshold (Figure 14). Given their positive average coefficient estimates in models that associate LPS variant count with phage resistance, these genes may have been drivers of LPS phage resistance breadth in this assay (Table 8). The coefficient estimates in Table 8 indicate the estimated average increase in relative OD associated with a 1 unit standard deviation increase in the number of SNPs within any one locus tag, while accounting for the number of SNPs in all other locus tags. Therefore, variation in locus tags with greater parameter estimates was associated with greater variation in relative OD.

Table 8: The locus tags that were retained by regularisation with positive parameter estimates in more than 75% of data subsets.

Locus tag	Mean coefficient estimate	Proportion of data subsets in which this locus tag was retained	Product
PA3149	0.038	1	probable glycosyltransferase WbpH
PA3153	0.022	1	O-antigen translocase
PA3154	0.005	1	B-band O-antigen polymerase
PA3159	0.008	1	UDP-N-acetyl-d-glucosamine 6-Dehydrogenase
PA3160	0.019	1	O-antigen chain length regulator
PA3242	0.158	1	temperature-regulated acyltransferase HtrB1
PA3643	0.095	1	lipid A-disaccharide synthase
PA4406	0.138	1	UDP-3-O-acyl-N-acetylglucosamine deacetylase
PA4457	0.047	1	arabinose-5-phosphate isomerase KdsD
PA4517	0.020	1	conserved hypothetical protein
PA5010	0.096	1	UDP-glucose:(heptosyl) LPS alpha 1,3-glucosyltransferase WaaG
PA5011	0.106	1	heptosyltransferase I
PA5162	0.079	1	dTDP-4-dehydrorhamnose reductase
PA5163	0.129	1	glucose-1-phosphate thymidylyltransferase
PA5447	0.015	1	glycosyltransferase WbpZ
PA5451	0.067	1	membrane subunit of A-band LPS efflux transporter
PA5452	0.062	1	phosphomannose isomerase/GDP-mannose WbpW
PA5453	0.232	1	GDP-mannose 4,6-dehydratase

3.4 Genetic variation in the *rmlA* gene may have contributed to LPS phage resistance

To assess whether individual genetic variants in LPS biosynthesis genes may have been driving resistance to particular LPS phages, the association between relative growth and presence of a genetic variant was estimated for every combination of phage and LPS genetic variant. After correcting for multiple testing, none of the variants significantly associated with phage resistance (Figure 6A). However, closer inspection of the gene with the lowest p-value associations revealed a variant with potential importance. This concerned four isolates, which had a SNP in the *rmlA* gene (locus tag PA5163; amino acid position 101, Asn > Asp): 4, 9, 12 and 22. This gene is involved in the production of L-rhamnose, a component of the O-antigen of LPS (65, 66). These isolates tended to have high relative OD values across all phages (Figure 6B). Additionally, these SNP-carrying isolates did not uniquely share any other variants in LPS biosynthesis genes (Table 10). Inspection of the full set of variants within the *rmlA* gene revealed that except for this variant, only a frameshift mutation in the *rmlA* gene was associated with high relative OD across most LPS phages (Figure 6C). This frameshift is likely to have caused a loss of function, given its position midway in the amino acid sequence (amino acid 129 out of 293). Taken together, variants in the *rmlA* gene may have contributed to generalist LPS phage resistance in the present assay.

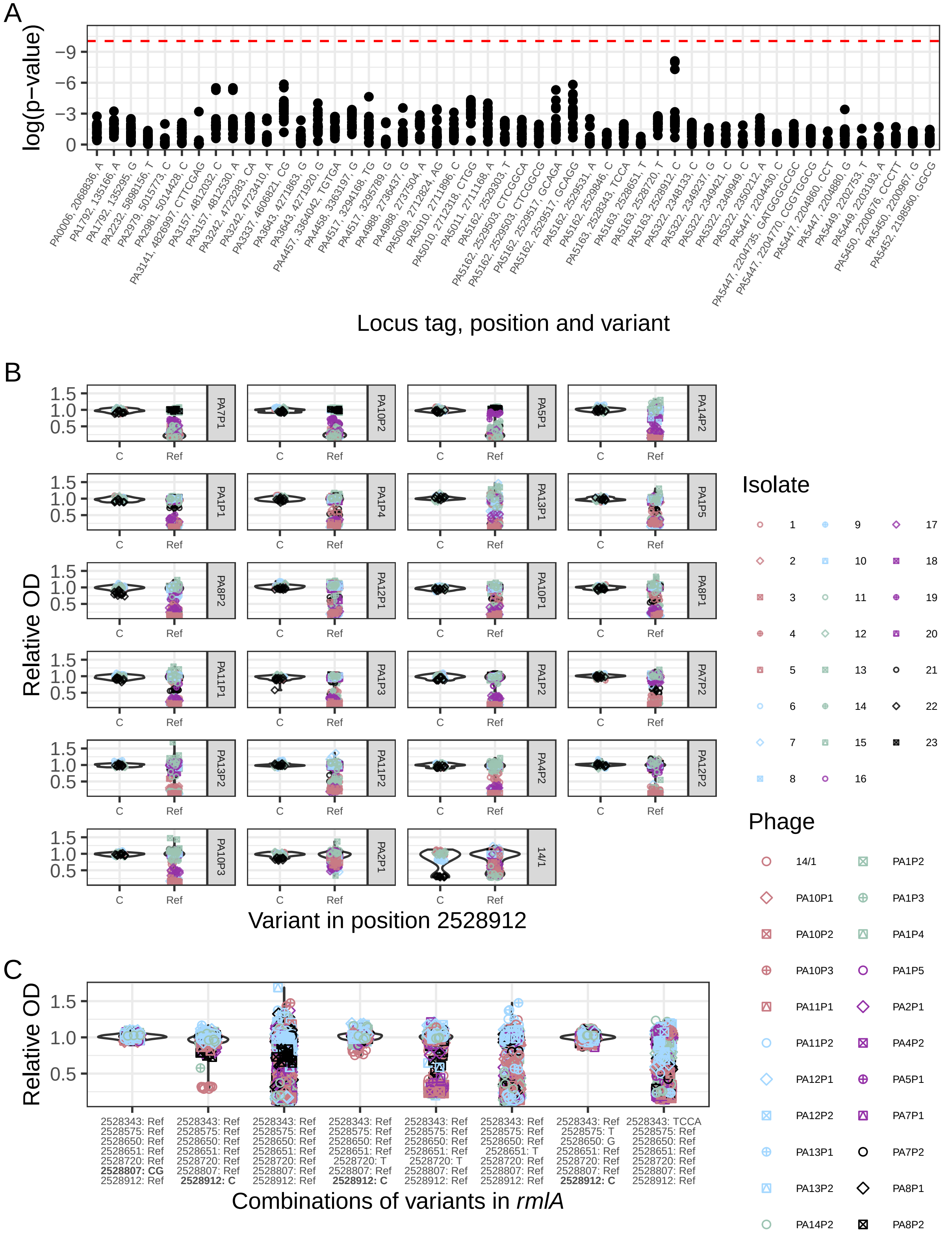

Figure 6: A variant in the *rmlA* gene may have been driving generalist LPS phage resistance. A) None of the variants in LPS genes significantly associated with resistance to individual phages. Each data point shows the p-value of a linear mixed model, estimating the association between the presence of a single genetic variant and relative OD in the presence of a specific phage. B) The lowest p-values in panel A were associated with a variant in the *rmlA* gene. It was found that isolates with this variant had high relative OD across most phages. Each facet shows the data for one phage. C) Only isolates with the variant shown in panel B (2528912: C) or a frameshift mutation (2528807: CG) had high relative OD across most phages.

3.5 A missense mutation in the *fimT* gene may have reduced phiKZ infectivity

To assess whether individual genetic variants in T4P biosynthesis genes may have been driving resistance to particular T4P phages, the association between relative growth and presence of a genetic variant was estimated for every combination of phage and T4P genetic variant. After correcting for multiple testing, one variant significantly associated with phage resistance (Figure 7A). Isolates 2, 9, 10, 15 and 18, with a Pro > Ala mutation in position 65 of the *fimT* gene, tended to have a high relative OD in the phiKZ treatment (Figure 7B). Previously, phage infectivity screens did not find *fimT* mutations to confer resistance to three T4P-binding phages (67, 68). In potential agreement with these findings, the *fimT* variant was associated with specialist resistance to phiKZ (Figure 7C). Finally, the isolates with this *fimT* variant did not uniquely share any other variants in T4P biosynthesis genes (Table 11).

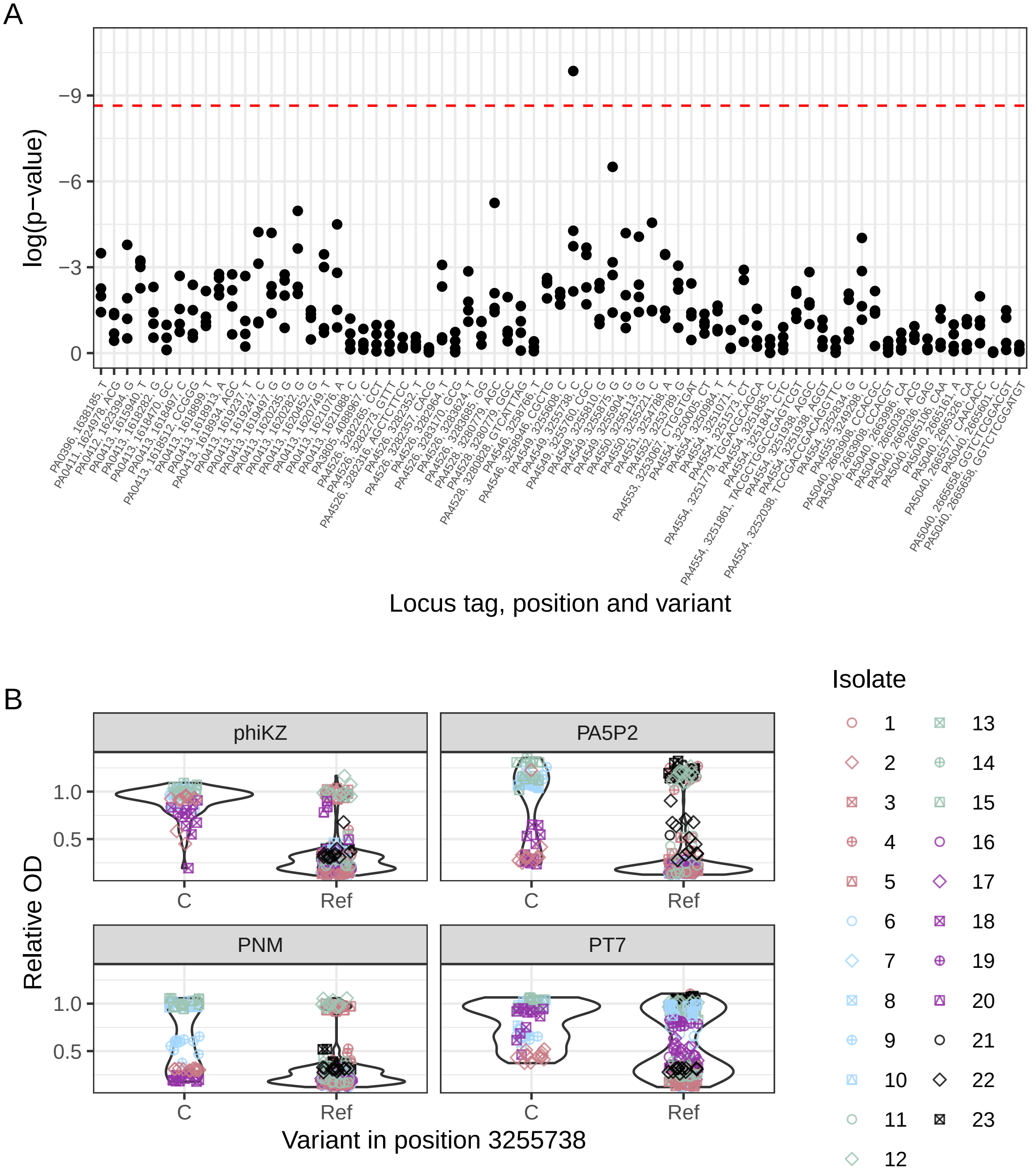

Figure 7: A variant in the *fimT* gene may have been driving phiKZ resistance. A) A variant in locus tag PA4549, position 3255738, significantly associated with resistance to one phage. Each data point shows the p-value of a linear mixed model, estimating the association between the presence of a single genetic variant and relative OD in the presence of a specific phage. B) Isolates with this variant only consistently had high average relative OD in the presence of phiKZ, suggesting that this variant was associated with specialist resistance.

3.6 Similarity in defence subsystem repertoire was positively associated with similarity in relative OD values

To ask whether defence systems might have been driving phage resistance, the defence subsystem repertoire of each bacterial isolate was predicted using DefenseFinder (31) and PADLOC (53). On average, isolates carried defence subsystems from 9 families (IQR: (4.75; 13.25)). A heatmap of this data showed a sparse matrix of defence system presence, with many system families only present in one or a few isolates (Figure 8). In total, 76 defence subsystem families were identified. These included restriction-modification (RM), CRISPR-Cas (Cas) and a wide range of recently discovered systems. Importantly, none of the isolates shared an identical defence subsystem repertoire. Note that abi systems were excluded from this analysis, as the high MOI used in the infectivity assay was unlikely to allow for resistance through abi to be detected (Figure 12, Table 9).

To assess whether overall defence subsystem repertoire was associated with resistance range, isolates were first clustered by their defence subsystem repertoire (Figure 8A). To quantify the association between similarity in defence system repertoire and resistance range, a mantel test was performed on the pairwise distances in resistance range (used to construct the clusters in Figure 3) and the pairwise distances in defence subsystem repertoire (Figure 8). The pairwise distances were significantly, albeit weakly, associated ($\mathrm{r} = 0.187, 95\%\ \mathrm{CI} = (0.02, 0.34), \mathrm{p} = 0.03$, Figure 8B). Overall, this suggests that similarity in defence subsystem repertoire contributes to predicting phage resistance specificity in these isolates.

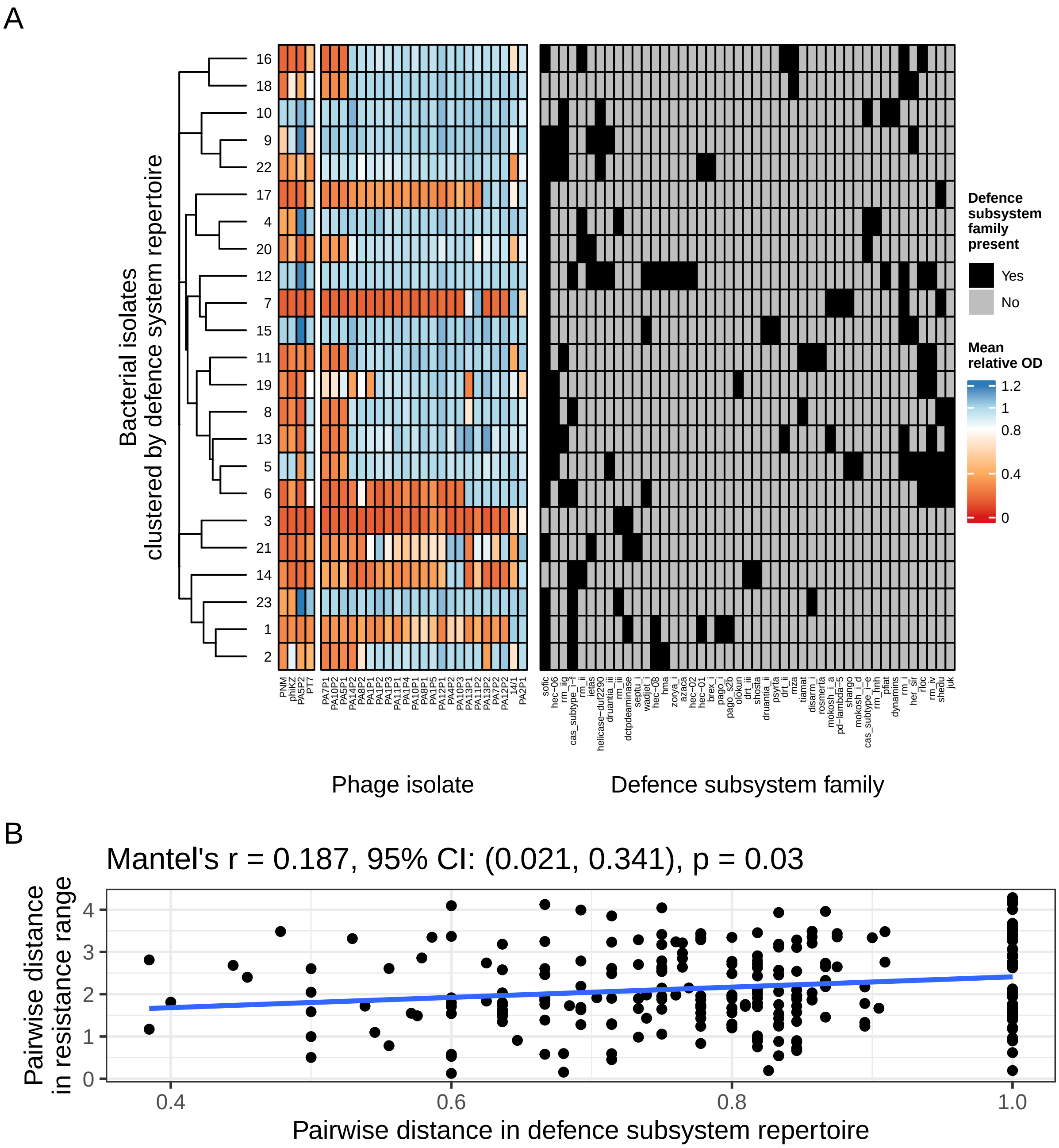

Figure 8: The association between pairwise distances in defence system repertoire and resistance range in terms of relative OD. A) A sparse matrix of defence system families revealed that none of the isolates shared an identical defence system repertoire. Isolates were clustered by their defence system repertoire. B) There was a weak positive association in pairwise distances.

3.7 Defence system repertoire size and phage resistance breadth were not significantly associated

To assess whether isolates with more defence systems tended to have greater phage resistance, the association between relative OD (Figure 3) and the number of defence subsystems was estimated. After accounting for non-independence due to experimental design, the effect of the number of subsystems was not significant

($\beta = 0.019, 95\%\ \mathrm{CI} = (-0.006, 0.046), \mathrm{p} = 0.18$) (Figure 9). Overall, this suggests that isolates with more defence subsystems were not on average more phage-resistant.

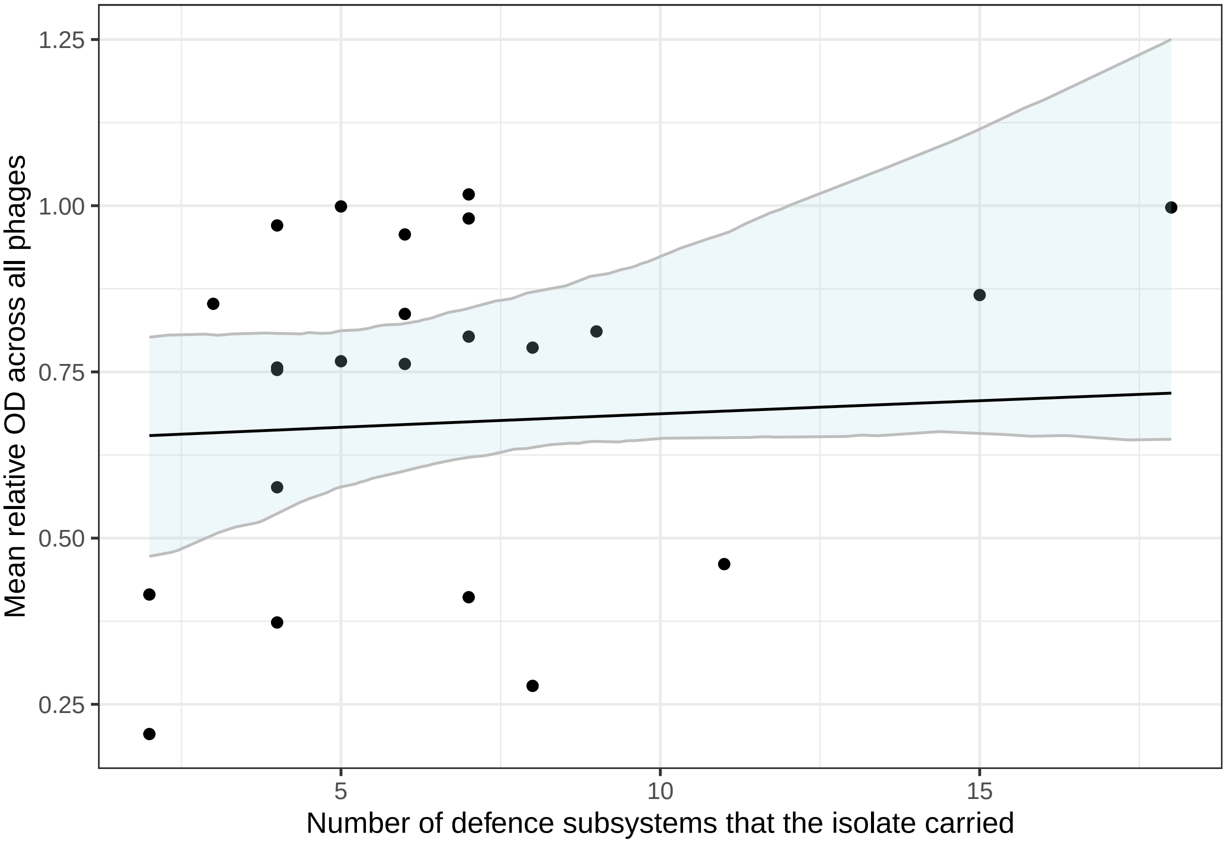

Figure 9: The mean relative OD of isolates in the presence of phage was not significantly associated with the number of defence subsystems carried by isolates. Data points show the mean relative OD for an isolate across all phages.

3.8 An adsorption assay suggested that the druantia defence system was not driving resistance to phage PNM

To assess whether individual defence systems may have been driving resistance to particular phages, the association between relative growth and the presence of a defence system was estimated for every combination of phage and defence system. This analysis suggested that the druantia defence system was providing protection against the PNM and phiKZ phages (Figure 10). The druantia system contains a large gene with a domain of unknown function, accompanied by 1-4 accessory genes, with some subtypes relying on nuclease activity for phage resistance (69, 70). Isolates 2, 5 and 12 carried druantia type II, while isolate 15 carried druantia type III. Despite this classification, the exact composition of the druantia system differed between all isolates (Figure 15). This suggested that these different compositions shared the ability to recognise and terminate PNM and phiKZ infections.

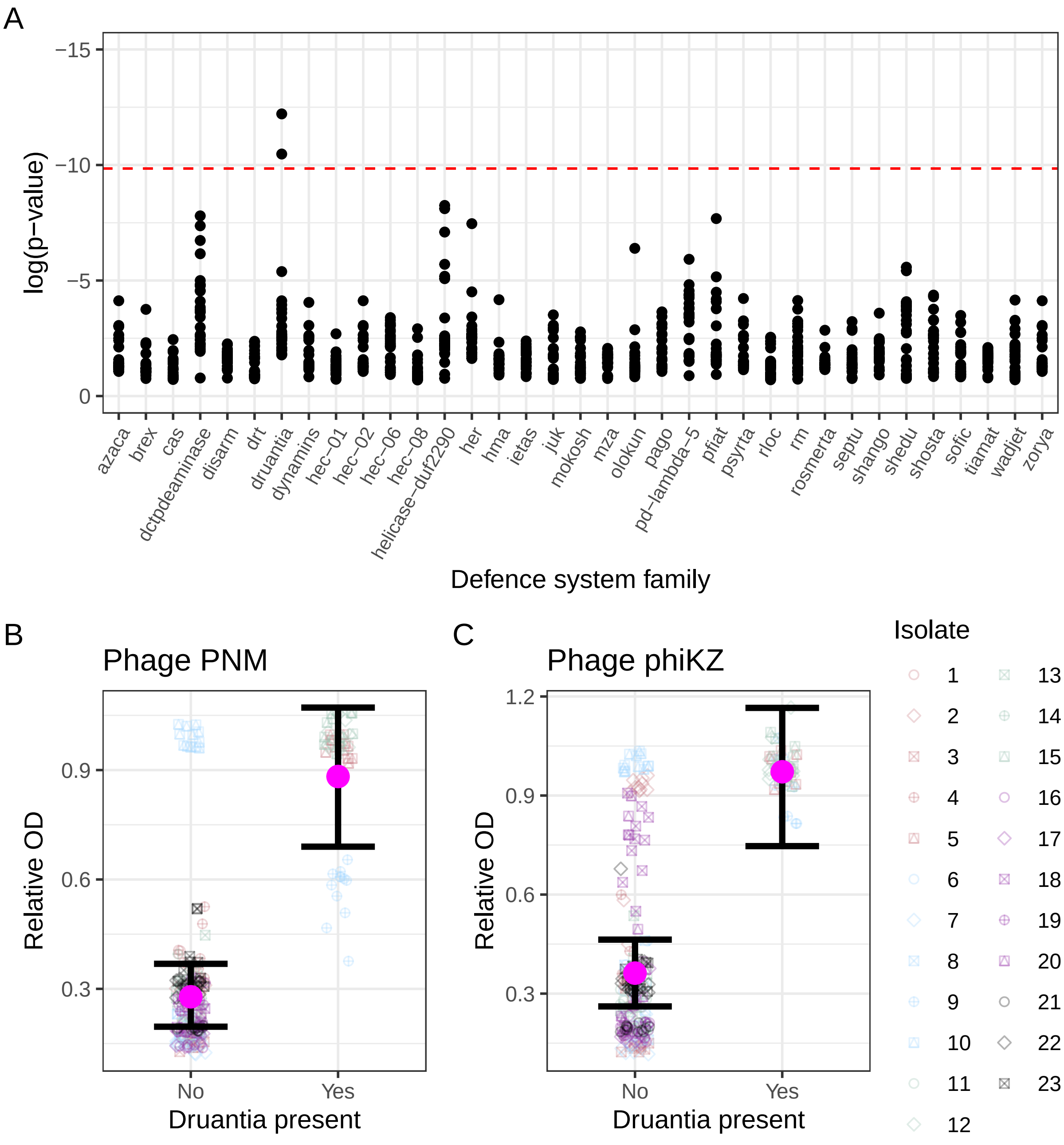

Figure 10: After correcting for multiple testing, hosts with druantia were found to have higher average relative OD than hosts without druantia in the presence of phages PNM or phiKZ. A) Each data point shows the log(p-value) for a linear mixed effects model estimating the association between presence of one defence system family and relative OD in the presence of one phage. B,C) The raw data for phages PNM and phiKZ, showing higher mean relative OD in hosts that carry druantia.

While receptor modification prevents a phage from binding to a host, resistance through a defence system still permits the phage to enter the host. Therefore, an adsorption assay was performed, to test whether phage PNM was entering hosts carrying the druantia defence system. An adsorption assay involved adding phage to a growing bacterial culture and frequently sampling the culture to determine phage titre. If the phage enters the host, an early drop in phage titre is observed. If the phage is able to replicate, a subsequent spike in phage titre is observed. All isolates carrying druantia, as well as PAO1 and non-druantia isolates with resistance to phiKZ/PNM were included in the assay (Figure 11). Note that while growth of isolate 9 was inhibited by phage PNM, this inhibition was only partial (Figure 10B). Therefore, it was expected that isolate 9 would still show a drop in phage titre, followed by a reduced increase in phage titre compared to susceptible strains, suggestive of phage entry followed by partial inhibition by druantia.

Firstly, isolates 24 and 2, which showed susceptibility to PNM in the infectivity assay, showed an increase in phage titre over the course of the assay. This was in line with the expectation that PNM would enter these hosts and replicate. Host 10, which showed resistance to PNM but did not carry druantia, showed a persistent drop in phage titre, suggestive of resistance through a non-druantia defence system. A BLAST search of isolate 10's CRISPR spacers against the PNM genome did not reveal significant hits. The only systems that were unique to isolate 10 were PifA and dynamins (Figure 8A, Figure 12). Since isolate 10 showed no growth inhibition in the presence of phage PNM in the infectivity assay, where the MOI was approximately 2, it is unlikely that isolate 10 resisted PNM through PifA, which is an abi system (71). In contrast, isolate 10 may have resisted PNM with the help of dynamins, which was shown to mediate phage resistance through the delay of cell lysis (72).

Surprisingly, none of the druantia-carrying strains showed a drop in phage titre, suggesting that phage PNM was not entering the hosts. Therefore, these hosts may have resisted PNM through a receptor modification. Unfortunately this modification could not be identified, as the isolates did not uniquely share any genetic variants in LPS biosynthesis genes (Table 12). Overall, this data suggests that despite druantia presence significantly associating with PNM resistance, the mode of resistance was more likely to be receptor-mediated.

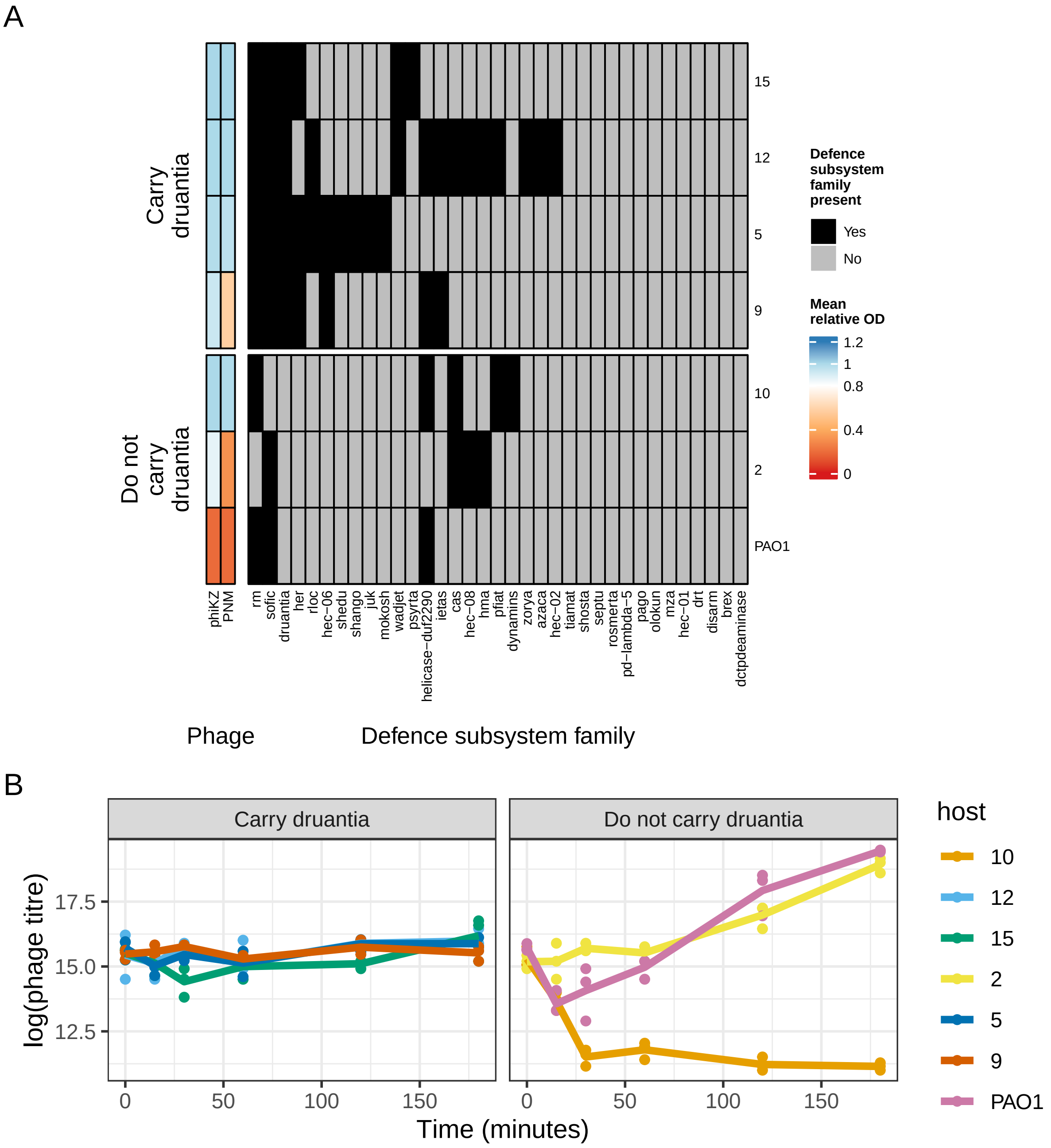

Figure 11: Druantia did not appear to be the driver of resistance to PNM in druantia-carrying isolates. A) All isolates carrying druantia, as well as PAO1 and non-druantia carrying isolates with resistance to PNM/phiKZ were included in the assay. B) Isolates carrying druantia displayed a constant phage titre, suggestive of receptor modification being the driver of phage resistance in these isolates. Phage titre in the absence of druantia depended on phage susceptibility: PAO1 and isolate 2 displayed an increase in phage titre, while isolate 10 showed a persistent drop in phage titre, suggestive of resistance through a defence system.

4 Discussion

For phage resistance in the presence of an appropriate phage receptor, bacterial hosts require a receptor modification or an appropriate defence system. This report examined the associations between growth under phage presence and these two modes of phage resistance. The associations between pairwise distances in host resistance range and receptor variant ranges were not significant. In contrast, the association between pairwise distances in host resistance and defence system repertoire was positive. In terms of the number of variants or systems, isolates with more LPS variants were found to have greater average resistance to LPS phages. Out of the 48 LPS genes, the number of variants in 18 genes were predicted to positively associate with resistance. Conversely, isolates with more defence system families were not found to have greater average resistance to phages. This report also showed that it is possible to identify potential drivers of resistance to specific phages. A variant in the *rmlA* gene may be driving generalist LPS phage resistance. Additionally, a variant in the *fimT* gene may be driving specialist phiKZ resistance. While presence of the druantia defence system was found to associate with resistance to phages PNM and phiKZ, an adsorption assay revealed that the druantia-carrying isolates were more likely to resist PNM through receptor modification. Overall, this report suggests that in the present isolate collection, LPS modifications were associated with resistance breadth, while defence system carriage was associated with resistance specificity.

The infection network suggested that on average, T4P phages were infective to a greater proportion of isolates than LPS phages. In the transition from acute to chronic CF infection, T4P-mediated twitching motility is thought to be lost (73). The isolates studied here were all the earliest representatives of their clone types within the local isolate collection. The earliest isolates were chosen to represent the scenario wherein an early infection is treated with phage therapy. Future work using isolates from later time points could ask whether these isolates show greater resistance to T4P phages. Additionally, future work could include a broader set of T4P phages, as only four were included in the present study. Nevertheless, this work highlights the importance of considering the changes that receptors undergo through the transition from acute to chronic infection, as these may inform the phages applied in therapy.

This work identified that CF isolates of *P. aeruginosa* carry genetic variants in many of the LPS and T4P biosynthesis genes, compared to the phage-sensitive PAO1. Nevertheless, there was no association between pairwise distances in variant counts and pairwise distances in relative OD. However, isolates with more LPS variants were found on average to have greater phage resistance. Out of the 48 studied LPS biosynthesis genes, 18 consistently showed positive parameter estimates through elastic net regularisation. The analogous association was not estimated to be significant for T4P biosynthesis genes, however this needs to be considered in light of the limited number of T4P phages included in this study. Notably, previous analyses of the drivers of resistance in PBINs have not described variants in receptor biosynthesis genes (8), described a sample in which isolates did not vary in those genes (9), tested for adsorption and concluded that a defence system must be at play (12, 74) or focused on uncovering the resistance drivers amongst phage-host pairs that were more likely to show defence system resistance (32, 75). The present analysis suggests that in non-clonal isolates, variation in receptor biosynthesis genes can associate with resistance breadth, highlighting the importance of considering defence systems in the context of isolate's receptor profile. The infectivity network in Piel *et al.* suggested that phage adsorption is largely clade-specific, and downstream characterisation suggested that differences in resistance within clades was driven by defence systems (75). Future work could expand the present PBIN, using both clonal and non-clonal isolates from the collection (18), to test whether resistance specificity is associated with receptor similarity between non-clonal isolates, and defence system similarity between clonal isolates.

In an attempt to associate specific variants with phage resistance, two genes with potential importance were identified: *rmlA* and *fimT*. A variant in the *rmlA* gene appeared to associate with generalist LPS phage resistance, while a variant in the *fimT* gene appeared to associate with specialised phiKZ resistance. Although the *rmlA* variant was not associated with a p-value below the adjusted significance threshold, the raw data provided a compelling case for its importance.

RmlA catalyses the first step in the production of L-rhamnose, a component of the O-antigen of LPS (65, 66). This fits with the broad resistance suggested by the present data. Additionally, variants in the *rmlA* gene (also known as *rfbD*) were previously found to allow *P. aeruginosa*, *Pseudomonas syringae* and *Listeria monocytogenes* to resist phage (6, 76–78). Moreover, an *rmlA* SNP was previously found to confer PA10P2 resistance, without conferring complete resistance to phage 14/1 (5). This pattern was also observed in this data (Figure 6B, panels PA10P2 and 14/1). Nevertheless, the protective effect of this *rmlA* variant (N101D) could be confirmed in future work, by treating the *rmlA* frameshift mutation identified in this study as a de-facto knockout and complementing it with N101D or WT *rmlA*. Interestingly, an *rmlA* variant was found to evolve *in vitro* under phage pressure in a *P. aeruginosa* isolate obtained from CF patients, and associated with reduced ceftazidime resistance in addition to increased phage resistance (6). Given that the isolates with this *rmlA* variant were isolated from different patients across a period of approximately five years (isolates 4, 9, 12 and 22 in Table 1), the *rmlA*-ceftazidime induced fitness trade-off would be worth exploring in future phage therapy trials (79). Overall, variants in the *rmlA* gene may have contributed to generalist LPS phage resistance in this study.

In contrast to *rmlA*, the association between the *fimT* variant and phage resistance was less clear based on previous work. The *fimT* gene is found adjacent to the *fimU* gene in the prepilin gene cluster, required for T4P biogenesis. FimU is known as a minor pilin which is incorporated into the external fibre of the T4P (80). Unfortunately, the incorportation of FimT was not assessed in this previous study (80). While *fimT* mutation has not been shown to affect twitching motility or phage sensitivity in *P. aeruginosa* (67, 68), *fimT* overexpression could restore twitching motility and phage sensitivity in a *fimU* mutant (68). Moreover, FimT shares important predicted features with FimU: a hydrophobic N-terminal region, a motif that is involved in leader sequence cleavage by the endoprotease PilD and a C-terminal disulphide-bonded domain (68). Nevertheless, the function of FimT is unknown (81), besides a potential role in DNA binding (82). However, given that phiKZ requires the T4P for adsorption (4, 83), it is possible that FimT is involved in the construction of the T4P and that the genetic variant observed here impairs phiKZ binding.

Notably, models were fit for each identified variant, rather than fitting a model for each biosynthesis gene, precluding the possibility of estimating variant effects given other variants carried by isolates within the same gene. Such an approach may be possible in the future with a network that involves more isolates. Additionally, given that variants in global regulators can also drive resistance, albeit at a greater fitness cost (4), future work could assess whether variants in regulatory genes associate with phage resistance. This would be especially important in the present CF isolates, as *P. aeruginosa* is known to accumulate variants in global regulators within the CF lung (84).

The bacteria in this study were predicted to carry defence systems from 59 different families. Many isolates were found to carry an RM system, which agrees with a previous analysis of all complete *P. aeruginosa* genomes in RefSeq (which does not include the CF isolates studied here) (31). Additionally, the sparse distribution of most identified defence systems is in line with the patchy distribution of the ten most common defence systems found in the same analysis (31). The association between defence subsystem repertoire and resistance range was weak (r = 0.187) suggesting that defence subsystem repertoire is not the only driver of phage resistance specificty in these isolates. Nevertheless, given that the range of known defence systems is continually expanding, alongside a continually changing naming system, it will be important for future work to re-assess the association between defence system similarity and phage resistance specificity.

Importantly, the number of defence systems was not found to significantly associate with phage resistance breadth. This result contrasts previous findings, where resistance breadth positively associated with the number of defence systems (32, 75). Future work could consider the diversity of the phage panel, as the panels used in previous work may have been more diverse than the panel used here. Overall, the present data suggests that the presence of specific defence systems, rather than the number of systems, may have been more important in determining phage resistance in the present isolate collection.

This analysis excluded abi system families (Figure 12), due to the setup of the infectivity assay. Under the high MOI used, it was deemed unlikely for the effects of abi to be detected, as the entire bacterial population would be expected to abort. Note however that defence system exclusion may have been too stringent, with the binarisation of defence systems on abi activity having been called into question (85). Assays which lead to abi phenotypes being called may have actually been driven by phage-induced damage to the host, leading to death by ‘mutual destruction’ (85). For example, while the DarTG system inhibits phage DNA synthesis, the resulting host death is thought to be driven by phage-induced host chromosome degradation (86). Additionally, there are known examples of context-dependence in abi phenotypes. For example, RM presence can rescue the growth arrest induced by Type IV CRISPR-Cas (87). Nevertheless, given that most of the excluded systems have not (yet) been found to act conditionally or to have phenotypes that were actually driven by mutual destruction, it was deemed beneficial for statistical power to exclude them from the analysis. However, in light of the high MOI complicating the analysis, future work could be performed at a lower MOI.

An analysis of pairwise associations between defence system presence and phage resistance suggested that the druantia system was associated with resistance to phages PNM and phiKZ. To rule out an undetected effect of receptor modification, an adsorption assay was performed with phage PNM. Unfortunately the phage did not appear to enter isolates carrying druantia, suggesting that these isolates were protected by a receptor modification. Additionally, assuming that the recent discovery of an abi phenotype for druantia Type III applies to the present isolates (32), it would have been unlikely to detect an effect of this subsystem in the infectivity assay with an MOI of 2. Finally, given the proposed nuclease-dependence of druantia type III (70), this system may have been unable to mediate resistance against phiKZ, which protects its nucleic acid through a nucleus-like compartment (88). Even if the isolates were carrying a defensive receptor modification, the data would still be puzzling: isolate 9, which was inhibited by PNM in the infectivity assay, did not show the expected decrease in phage titre early in the adsorption assay. In contrast, isolate 10, which was assayed in the same batch as isolate 9, did show the expected drop in phage titre. Overall, the adsorption assay did not provide evidence in favour of druantia providing the isolates with protection. Nevertheless, the positive control data suggested that a repeat of the assay would be necessary in order to confidently conclude receptor-mediated protection.

In conclusion, this study suggested that out of the factors studied, only the number of modifications in LPS biosynthesis genes was associated with phage resistance breadth. In contrast, the range of defence subsystems appeared to associate with phage resistance specificity. Nevertheless, given the low number of T4P phages studied, and the infrequent overlap in defence system repertoires between isolates, future work including more T4P phages, as well as clonal isolates in addition to non-clonal isolates, could more comprehensively assess the importance of these modes of resistance. Finally, this work highlights the importance of considering naturally-occuring LPS modifications in the development of phage therapeutics for CF *P. aeruginosa* infections. The use of phage-antibiotic synergy, as well as the combination of T4P and LPS phages, may avoid such modifications from protecting CF isolates from phage therapeutics.

5 References

1. Westra ER, et al. (2015) Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Current Biology* 25(8):1043–1049.

2. Betts A, Gifford DR, MacLean RC, King KC (2016) Parasite diversity drives rapid host dynamics and evolution of resistance in a bacteria-phage system. *Evolution* 70(5):969–978.

3. Betts A, Gray C, Zelek M, MacLean RC, King KC (2018) High parasite diversity accelerates host adaptation and diversification. *Science* 360(6391):907–911.

4. Wright RCT, Friman V-P, Smith MCM, Brockhurst MA (2018) Cross-resistance is modular in bacteria–phage interactions. *PLOS Biology* 16(10):e2006057.

5. Wright RCT, Friman V-P, Smith MCM, Brockhurst MA (2019) Resistance Evolution against Phage Combinations Depends on the Timing and Order of Exposure. *mBio* 10(5). doi:10.1128/mBio.01652-19.

6. Nordstrom HR, et al. (2022) Genomic characterization of lytic bacteriophages targeting genetically diverse Pseudomonas aeruginosa clinical isolates. *iScience* 25(6). doi:10.1016/j.isci.2022.104372.

7. Georjon H, Bernheim A (2023) The highly diverse antiphage defence systems of bacteria. *Nat Rev Microbiol*:1–15.

8. LeGault KN, et al. (2021) Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. *Science*. doi:10.1126/science.abg2166.

9. Hussain FA, et al. (2021) Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science* 374(6566):488–492.

10. Alseth EO, et al. (2019) Bacterial biodiversity drives the evolution of CRISPR-based phage resistance. *Nature* 574(7779):549–552.

11. Weitz JS, et al. (2013) Phage–bacteria infection networks. *Trends in Microbiology* 21(2):82–91.

12. Laanto E, Hoikkala V, Ravantti J, Sundberg L-R (2017) Long-term genomic coevolution of host-parasite interaction in the natural environment. *Nat Commun* 8(1):111.

13. Kortright KE, Chan BK, Koff JL, Turner PE (2019) Phage Therapy: A Renewed Approach to Combat Antibiotic-Resistant Bacteria. *Cell Host & Microbe* 25(2):219–232.

14. Thornton CS, Parkins MD (2023) Microbial Epidemiology of the Cystic Fibrosis Airways: Past, Present, and Future. *Semin Respir Crit Care Med* 44(02):269–286.

15. Adler FR, Liou TG (2016) The Dynamics of Disease Progression in Cystic Fibrosis. *PLOS ONE* 11(6):e0156752.

16. Salsgiver EL, et al. (2016) Changing Epidemiology of the Respiratory Bacteriology of Patients With Cystic Fibrosis. *Chest* 149(2):390–400.

17. Rossi E, et al. (2020) Pseudomonas aeruginosa adaptation and evolution in patients with cystic fibrosis. *Nature Reviews Microbiology*:1–12.

18. Marvig RL, Sommer LM, Molin S, Johansen HK (2015) Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. *Nature Genetics* 47(1):57–64.

19. Mangalea MR, Duerkop BA (2020) Fitness Trade-Offs Resulting from Bacteriophage Resistance Potentiate Synergistic Antibacterial Strategies. *Infection and Immunity* 88(7). doi:10.1128/IAI.00926-19.

20. Chan BK, Stanley G, Modak M, Koff JL, Turner PE (2021) Bacteriophage therapy for infections in CF. *Pediatric Pulmonology* 56(S1):S4–S9.

21. Trend S, Fonceca AM, Ditcham WG, Kicic A, Cf A (2017) The potential of phage therapy in cystic fibrosis: Essential human-bacterial-phage interactions and delivery considerations for use in Pseudomonas aeruginosa-infected airways. *Journal of Cystic Fibrosis* 16(6):663–670.

22. El Didamony G, Askora A, Shehata AA (2015) Isolation and Characterization of T7-Like Lytic Bacteriophages Infecting Multidrug Resistant Pseudomonas aeruginosa Isolated from Egypt. *CURRENT MICROBIOLOGY* 70(6):786–791.

23. Furusawa T, et al. (2016) Complete Genome Sequences of Broad-Host-Range Pseudomonas aeruginosa Bacteriophages Phi R18 and Phi S12-1. *MICROBIOLOGY RESOURCE ANNOUNCEMENTS* 4(3). doi:10.1128/genomeA.00041-16.

24. Issa R, et al. (2019) Antibiofilm potential of purified environmental bacteriophage preparations against early stage Pseudomonas aeruginosa biofilms. *JOURNAL OF APPLIED MICROBIOLOGY* 126(6):1657–1667.

25. Li L, Yang H, Lin S, Jia S (2010) Classification of 17 newly isolated virulent bacteriophages of Pseudomonas aeruginosa. *CANADIAN JOURNAL OF MICROBIOLOGY* 56(11):925–933.

26. Cullen L, et al. (2015) Phenotypic characterization of an international Pseudomonas aeruginosa reference panel: Strains of cystic fibrosis (CF) origin show less in vivo virulence than non-CF strains. *MICROBIOLOGY-SGM* 161(10):1961–1977.

27. Essoh C, et al. (2013) The Susceptibility of Pseudomonas aeruginosa Strains from Cystic Fibrosis Patients to Bacteriophages. *PLOS ONE* 8(4):e60575.

28. Forti F, et al. (2018) Design of a Broad-Range Bacteriophage Cocktail That Reduces Pseudomonas aeruginosa Biofilms and Treats Acute Infections in Two Animal Models. *ANTIMICROBIAL AGENTS AND CHEMOTHERAPY* 62(6). doi:10.1128/AAC.02573-17.

29. Kwiatek M, et al. (2015) Characterization of five newly isolated bacteriophages active against Pseudomonas aeruginosa clinical strains. *FOLIA MICROBIOLOGICA* 60(1):7–14.

30. Bondy-Denomy J, et al. (2016) Prophages mediate defense against phage infection through diverse mechanisms. *The ISME Journal* 10(12):2854–2866.

31. Tesson F, et al. (2022) Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun* 13(1):2561.

32. Costa AR, et al. (2023) Accumulation of defense systems in phage resistant strains of Pseudomonas aeruginosa. doi:10.1101/2022.08.12.503731.

33. Johnson MC, et al. (2023) Core defense hotspots within Pseudomonas aeruginosa are a consistent and rich source of anti-phage defense systems. *Nucleic Acids Research*:gkad317.

34. Merabishvili M, et al. (2007) Digitized fluorescent RFLP analysis (fRFLP) as a universal method for comparing genomes of culturable dsDNA viruses: Application to bacteriophages. *Research in Microbiology* 158(7):572–581.

35. Palmer KL, Aye LM, Whiteley M (2007) Nutritional Cues Control Pseudomonas aeruginosa Multicellular Behavior in Cystic Fibrosis Sputum. *Journal of Bacteriology* 189(22):8079–8087.

36. Pedersen TL (2020) *Patchwork: The Composer of Plots* Available at: `https://CRAN.R-project.org/package=patchwork`.

37. Wickham H (2016) *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York) Available at: `https://ggplot2.tidyverse.org`.

38. Herman E (2024) Raw infectivity data plots. doi:10.5281/zenodo.11217198.

39. Gu Z, Eils R, Schlesner M (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics.*

40. R Core Team (2022) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria) Available at: `https://www.R-project.org/`.

41. Mölder F, et al. (2021) *Sustainable data analysis with Snakemake* (F1000Research) doi:10.12688/f1000research.29032.2.

42. Wick R (2022) Rrwick/Filtlong. Available at: `https://github.com/rrwick/Filtlong` [Accessed December 21, 2022].

43. Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 37(5):540–546.

44. Nanoporetech (2022) Medaka. Available at: `https://github.com/nanoporetech/medaka` [Accessed December 21, 2022].

45. Chen S, Zhou Y, Chen Y, Gu J (2018) Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34(17):i884–i890.

46. Wick RR, Holt KE (2022) Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLOS Computational Biology* 18(1):e1009802.

47. Schwengers O, et al. (2021) Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics* 7(11):000685.

48. Seemann T Snippy: Fast bacterial variant calling from NGS reads. Available at: `https://github.com/tseemann/snippy`.

49. Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. doi:10.48550/arXiv.1207.3907.

50. Cingolani P, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 6(2):80–92.

51. Winsor GL, et al. (2016) Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. *Nucleic Acids Research* 44(D1):D646–D653.

52. Garrison E, Kronenberg ZN, Dawson ET, Pedersen BS, Prins P (2022) A spectrum of free software tools for processing the VCF variant call format: Vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLOS Computational Biology* 18(5):e1009123.

53. Payne LJ, et al. (2021) Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Research* (gkab883). doi:10.1093/nar/gkab883.

54. Millman A, et al. (2020) Bacterial Retrons Function In Anti-Phage Defense. *Cell* 0(0). doi:10.1016/j.cell.2020.09.065.

55. Gao L, et al. (2020) Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science* 369(6507):1077–1084.

56. Tal N, et al. (2021) Cyclic CMP and cyclic UMP mediate bacterial immunity against phages. *Cell* 184(23):5728–5739.e16.

57. Cohen D, et al. (2019) Cyclic GMP–AMP signalling protects bacteria against viral infection. *Nature* 574(7780):691–695.

58. Oksanen J, et al. (2020) Vegan: Community ecology package. Available at: `https://CRAN.R-project.org/package=vegan`.

59. Goslee SC, Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software* 22(7):1–19.

60. Bates D, Maechler M, Bolker B, Walker S (2015) *Fitting Linear Mixed-Effects Models Using lme4* (Journal of Statistical Software).

61. Corporation microsoft, Weston S (2022) *doParallel: Foreach parallel adaptor for the 'parallel' package* Available at: `https://CRAN.R-project.org/package=doParallel`.

62. Microsoft, Weston S (2022) *Foreach: Provides foreach looping construct* Available at: `https://CRAN.R-project.org/package=foreach`.

63. Friedman JH, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33:1–22.

64. Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13):1–26.

65. Blankenfeldt W, Asuncion M, Lam JS, Naismith JH (2000) The structural basis of the catalytic mechanism and regulation of glucose-1-phosphate thymidylyltransferase (RmlA). *The EMBO Journal* 19(24):6652–6663.

66. Xiao G, et al. (2021) Next generation Glucose-1-phosphate thymidylyltransferase (RmlA) inhibitors: An extended SAR study to direct future design. *Bioorganic & Medicinal Chemistry* 50:116477.

67. Belete B, Lu H, Wozniak DJ (2008) *Pseudomonas aeruginosa* AlgR Regulates Type IV Pilus Biosynthesis by Activating Transcription of the *fimU-pilVWXY1Y2E* Operon. *J Bacteriol* 190(6):2023–2030.

68. Alm RA, Mattick JS (1996) Identification of two genes with prepilin-like leader sequences involved in type 4 fimbrial biogenesis in Pseudomonas aeruginosa. *J Bacteriol* 178(13):3809–3817.

69. Doron S, et al. (2018) Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* 359(6379). doi:10.1126/science.aar4120.

70. Wang S, et al. (2023) Landscape of New Nuclease-Containing Antiphage Systems in Escherichia coli and the Counterdefense Roles of Bacteriophage T4 Genome Modifications. *Journal of Virology* 97(6):e00599–23.

71. Lopatina A, Tal N, Sorek R (2020) Abortive Infection: Bacterial Suicide as an Antiviral Immune Strategy. *Annual Review of Virology* 7(1):null.

72. Guo L, Sattler L, Shafqat S, Graumann PL, Bramkamp M (2022) A Bacterial Dynamin-Like Protein Confers a Novel Phage Resistance Strategy on the Population Level in Bacillus subtilis. *mBio* 13(1):e03753–21.

73. Kus JV, Tullis E, Cvitkovitch DG, Burrows LL (2004) Significant differences in type IV pilin allele distribution among Pseudomonas aeruginosa isolates from cystic fibrosis (CF) versus non-CF patients. *Microbiology* 150(5):1315–1326.

74. Cahier K, et al. (2023) Environmental vibrio phage–bacteria interaction networks reflect the genetic structure of host populations. *Environmental Microbiology.* doi:10.1111/1462-2920.16366.

75. Piel D, et al. (2022) Phage–host coevolution in natural populations. *Nat Microbiol*:1–12.

76. Garbe J, Bunk B, Rohde M, Schobert M (2011) Sequencing and Characterization of Pseudomonas aeruginosa phage JG004. *BMC Microbiology* 11(1):102.

77. Meaden S, Paszkiewicz K, Koskella B (2015) The cost of phage resistance in a plant pathogenic bacterium is context-dependent. *Evolution* 69(5):1321–1328.

78. Eugster MR, et al. (2015) Bacteriophage predation promotes serovar diversification in Listeria monocytogenes. *Molecular Microbiology* 97(1):33–46.

79. Engeman E, et al. (2021) Synergistic Killing and Re-Sensitization of Pseudomonas aeruginosa to Antibiotics by Phage-Antibiotic Combination Treatment. *Pharmaceuticals* 14(3):184.

80. Giltner CL, Habash M, Burrows LL (2010) Pseudomonas aeruginosa Minor Pilins Are Incorporated into Type IV Pili. *Journal of Molecular Biology* 398(3):444–461.

81. Burrows LL (2012) Pseudomonas aeruginosa Twitching Motility: Type IV Pili in Action. *Annual Review of Microbiology* 66(1):493–520.

82. Braus SAG, et al. (2022) The molecular basis of FimT-mediated DNA uptake during bacterial natural transformation. *Nat Commun* 13(1):1065.

83. Danis-Wlodarczyk K, et al. (2016) A proposed integrated approach for the preclinical evaluation of phage therapy in Pseudomonas infections. *Sci Rep* 6(1):28115.

84. Camus L, Vandenesch F, Moreau K (2021) From genotype to phenotype: Adaptations of Pseudomonas aeruginosa to the cystic fibrosis environment. *Microbial Genomics* 7(3):000513.

85. Aframian N, Eldar A (2023) Abortive infection antiphage defense systems: Separating mechanism and phenotype. *Trends in Microbiology.* doi:10.1016/j.tim.2023.05.002.

86. LeRoux M, et al. (2022) The DarTG toxin-antitoxin system provides phage defence by ADP-ribosylating viral DNA. *Nat Microbiol* 7(7):1028–1040.

87. Williams MC, et al. (2023) Restriction endonuclease cleavage of phage DNA enables resuscitation from Cas13-induced bacterial dormancy. *Nat Microbiol* 8(3):400–409.

88. Mendoza SD, et al. (2020) A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature* 577(7789):244–248.

89. Chopin M-C, Chopin A, Bidnenko E (2005) Phage abortive infection in lactococci: Variations on a theme. *Current Opinion in Microbiology* 8(4):473–479.

90. Gao LA, et al. (2022) Prokaryotic innate immunity through pattern recognition of conserved viral proteins. *Science* 377(6607):eabm4096.

91. Béchon N, et al. (2024) Diversification of molecular pattern recognition in bacterial NLR-like proteins. doi:10.1101/2024.01.31.578182.

92. Zhang T, et al. (2022) Direct activation of a bacterial innate immune system by a viral capsid protein. *Nature* 612(7938):132–140.

93. Duncan-Lowey B, Kranzusch PJ (2022) CBASS phage defense and evolution of antiviral nucleotide signaling. *Current Opinion in Immunology* 74:156–163.

94. LeRoux M, et al. (2022) The DarTG toxin-antitoxin system provides phage defence by ADP-ribosylating viral DNA. *Nat Microbiol* 7(7):1028–1040.

95. Stokar-Avihail A, et al. (2023) Discovery of phage determinants that confer sensitivity to bacterial immune systems. *Cell* 186(9):1863–1876.e16.

96. Garb J, et al. (2022) Multiple phage resistance systems inhibit infection via SIR2-dependent NAD+ depletion. *Nat Microbiol* 7(11):1849–1856.

97. Cheng R, et al. (2023) Prokaryotic Gabija complex senses and executes nucleotide depletion and DNA cleavage for antiviral defense. *Cell Host & Microbe* 31(8):1331–1344.e5.

98. Millman A, et al. (2022) An expanded arsenal of immune systems that protect bacteria from phages. *Cell Host & Microbe* 30(11):1556–1569.e5.

99. Rousset F, et al. (2022) Phages and their satellites encode hotspots of antiviral systems. *Cell Host & Microbe* 30(5):740–753.e5.

100. Vassallo C, Doering C, Littlehale ML, Teodoro G, Laub MT (2022) Mapping the landscape of anti-phage defense mechanisms in the E. Coli pangenome. doi:10.1101/2022.05.12.491691.

101. Tal N, et al. (2021) Cyclic CMP and cyclic UMP mediate bacterial immunity against phages. *Cell* 184(23):5728–5739.e16.

102. Millman A, et al. (2020) Bacterial Retrons Function In Anti-Phage Defense. *Cell* 183(6):1551–1561.e12.

103. Sberro H, et al. (2013) Discovery of Functional Toxin/Antitoxin Systems in Bacteria by Shotgun Cloning. *Molecular Cell* 50(1):136–148.

104. Ofir G, et al. (2021) Antiviral activity of bacterial TIR domains via immune signalling molecules. *Nature* 600(7887):116–120.

6 Supplementary materials

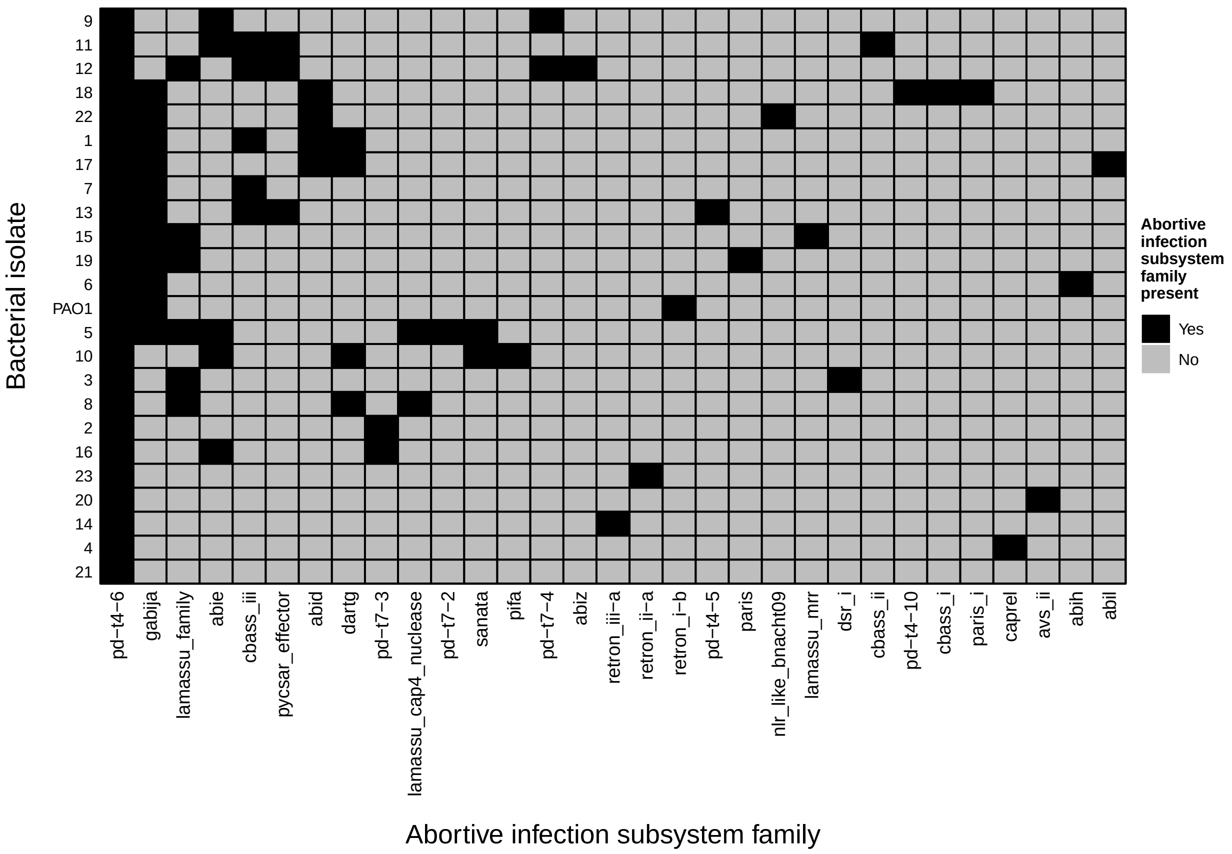

Figure 12: The abortive infection subsystems that were detected in this work. These systems were excluded from the analysis, due to low chance of detecting the effects of abortive infection in the infectivity assay with an MOI of 2.

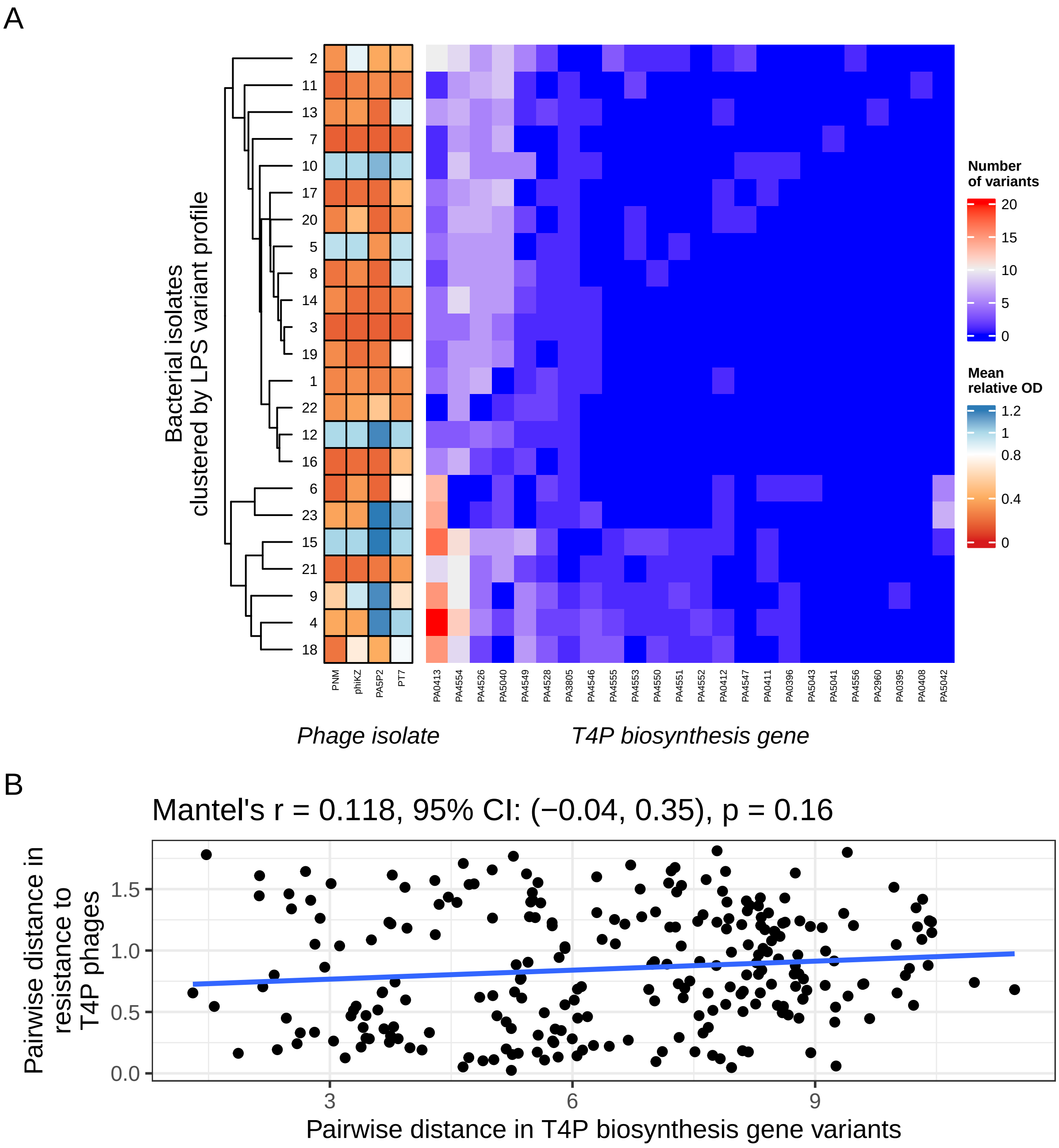

Figure 13: A matrix of T4P biosynthesis variants revealed that none of the isolates shared an identical variants profile. A) Isolates were clustered by their T4P biosynthesis variants, showing that isolates with relatively close variants profiles could differ greatly in their relative OD values with T4P phages B) A one-sided mantel test did not suggest that a positive association exists between the pairwise distances.

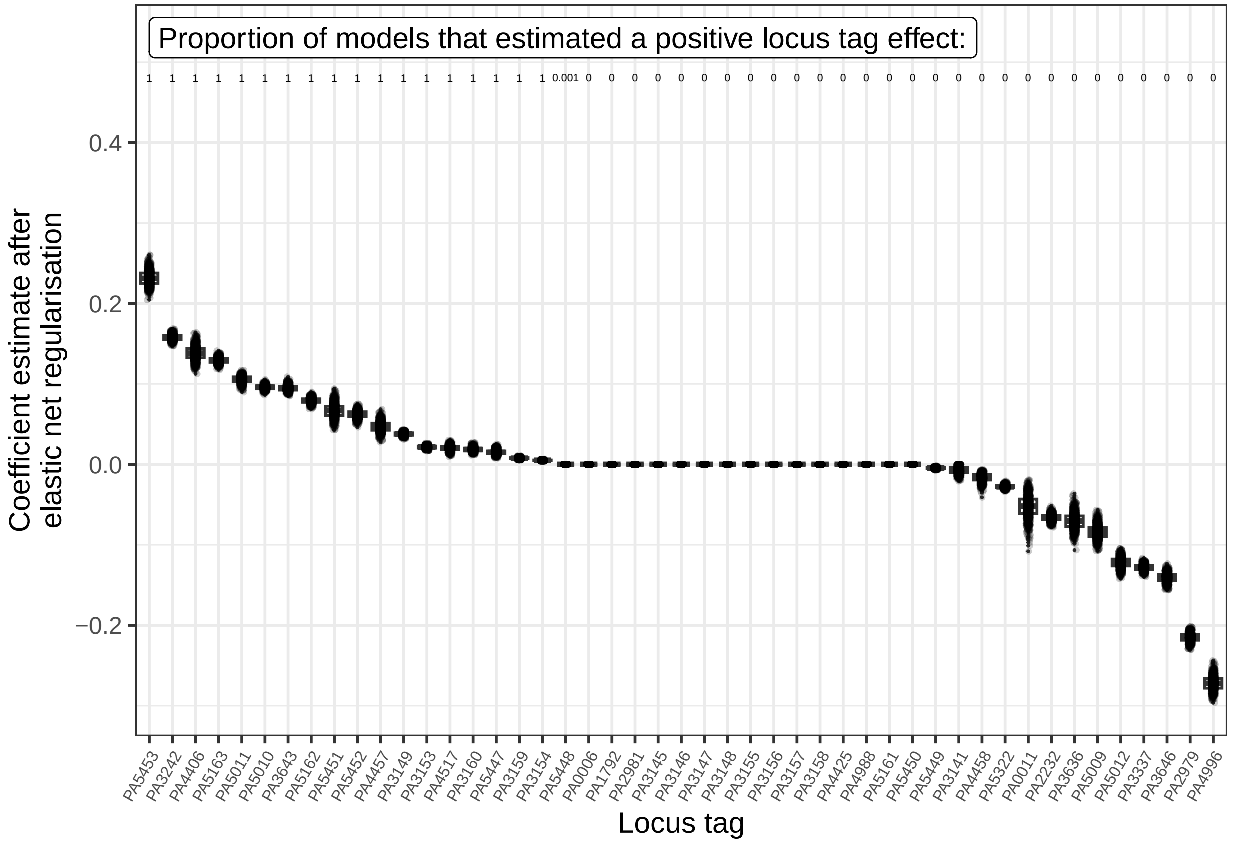

Figure 14: The coefficient estimates associated with each locus tag across data subsets. The proportion of data subsets in which regularisation retained a locus tag with a positive parameter estimate is shown at the top of the figure. Note that variant counts were standardised prior to regularisation, to avoid locus tags with higher average variant counts being assigned greater importance in regularisation.

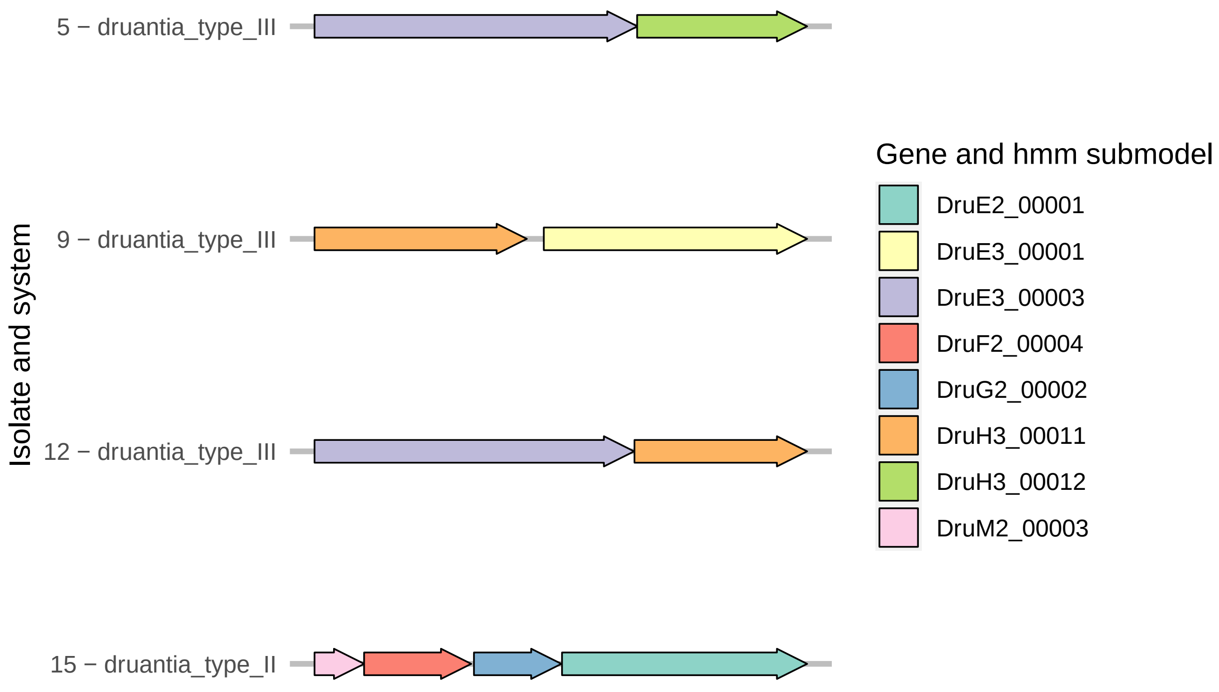

Figure 15: The four identified druantia systems differed in composition. While three systems were of type II, none shared the exact same set of Hidden Markov Model (hmm) models that identified the genes. The different hmm models for one gene identify different variations on that gene.

Table 9: The abortive infection systems that were identified in this work and excluded from the analysis.

system	ref
Abi	(89)
Avast	(90)
Avs	(90)
Avs	(91)
CapRel	(92)
CBASS	(93)
DarTG	(94)
Dazbog	(95)
DSR	(96)
Gabija	(97)
Lamassu	(98)
nlr	(90)
PARIS	(99)
PD-T4-10	(100)
PD-T4-5	(100)
PD-T4-6	(100)
PD-T7-2	(100)
PD-T7-3	(100)

system	ref
PD-T7-4	(100)
PifA	(71)
Pycsar	(101)
Retron	(102)
sanaTA	(103)
Thoeris	(104)

Table 10: The number of isolates which carry a variant which is found in all isolates with the *rmlA* variant of interest (4, 9, 12 and 22). This table shows that there were no other variants, besides the variant at position 2528912, which were shared by all four isolates and were unique to them across all bacteria.

Position	Variant	Locus Tag	Number of occurences in other isolates
2204880	G	PA5447	16
2528912	C	PA5163	0
4606821	CG	PA3337	4

Table 11: The number of isolates which carry a variant which is found in all isolates with the *fimT* variant of interest (2, 9, 10, 15 and 18). This table shows that there were no other variants, besides the variant at position 3255738, which were unique to them across all bacteria.

Position	Variant	Locus Tag	Number of occurences in other isolates
3250984	T	PA4554	14
3251841	CTC	PA4554	13
3255608	C	PA4549	10
3255738	C	PA4549	0
3255810	G	PA4549	6
3255875	G	PA4549	1

Table 12: The number of isolates which carry a variant which is found in all isolates carrying druantia (5, 9, 12 and 15). This table shows that there were no variants which were shared by all four isolates and were unique to them across all bacteria.

Position	Variant	Locus Tag	Number of occurences in other isolates
2204880	G	PA5447	16
3294168	TG	PA4517	15

Chapter 3: Clinical Cystic Fibrosis *P. aeruginosa* isolates show a broad but stable defence system repertoire

Abstract

Defence systems are thought to be in constant flux, including in *Pseudomonas aeruginosa* which is known to have a broad pan-immune system. This study investigated the defense system repertoire of clinical cystic fibrosis *P. aeruginosa* isolates, following the idea that bacteria carry different segments of a population's pan-immune system over time. This theory suggests that distantly-related isolates should show signs of horizontal gene transfer, and that the defence system repertoire of lineages should be changing over time. To assess whether these patterns held in the clinical isolates, the defense system repertoire was described, alongside its genomic localisation, mobility, and changes within patients. The clinical isolates possessed a diverse and open pan-immune system, hinting at the potential discovery of more system families beyond the 82 studied here. Larger genomes tended to carry more defense systems, however their number did not positively associate with prophage count after accounting for genome size. Most defense systems localized to regions of genomic plasticity. While many variable regions could be grouped into spots of putative insertion, most spots did not appear to be defence system hotspots. These spots also revealed that sharing multiple defense systems in the same location among distant isolates was rare, and closely related isolates typically shared the same systems in any one location, aligning with the stable defense system repertoires identified within patients over time. Overall, clinical CF *P. aeruginosa* isolates showed a broad pan-immune system with minimal signs of horizontal defence system movement.

1 Introduction

Over the past years, the field of phage-bacteria interactions has seen a major influx of novel defence system discoveries (1). Defence systems protect bacteria from phage infection by degrading phage nucleic acid or by inducing programmed cell death or dormancy (1). These defence systems have taken our understanding of bacterial immunity from reliant on RM and CRISPR-Cas to a model of pan-immunity, where the defence system repertoire of a bacterial community is far greater than any one of its inhabitants (2). With the availability of tools to detect defence systems in bacterial genomes (3, 4), there is a growing interest in defence system dynamics in natural populations (5–8). Nevertheless, studies of natural population-wide defence system repertoires are rare, and the study of changes over time has been limited to individual genomic regions (5–7). As of yet, our understanding of defence system evolutionary dynamics in natural populations is limited.

These previous studies of defence systems in natural environments have focused on the *Vibrio* genus: *Vibrio cholerae*, the cause of cholera, and marine *Vibrio* species (5–8). These studies suggested that defence system repertoires change over time, with closely-related isolates carrying variable genomic regions that differ in defence systems and associated phage resistance range (5–7). While these are clear examples of the importance of defence system movement in natural environments, their results may be specific to *Vibrio* species. Firstly, *Vibrio* may not be very separated spatially, with sewage systems (*V. cholerae*) and seawater (marine *Vibrio*) potentially allowing for frequent encounters between non-clonal isolates. In line with this, *Vibrio* are known for their high rates of horizontal gene transfer (HGT) (9). Additionally, Hussain *et al.* found little variation in phage receptors and Piel *et al.* found phage adsorption to be clade-specific, suggesting that defence systems could be particularly important in *Vibrio*'s phage defence (5, 7). It is possible that defence system dynamics will be different in a study system with increased spatial structure between hosts and/or greater variation in phage receptors.

P. aeruginosa, an opportunistic human pathogen that is often found in the lungs of patients with Cystic

Fibrosis (CF), is a major cause of morbidity and mortality (10). Crucially, CF patients are isolated when visiting the clinic, to minimise patient to patient transmission (11). Additionally, previous work has found receptor mutations in clinical isolates as well as in isolates evolved *in vitro*, suggesting that receptor-mediated phage resistance is important for this species (12–18). Nevertheless, *P. aeruginosa* is known to have a broad pan-immune system, with isolates and the species as a whole carrying many different systems (4, 19). However, the patterns of defence system gain and loss have not been studied in natural or clinical environments. Moreover, adaptation of early *P. aeruginsa* isolates to the CF lung is characterised by genome reduction (20), suggesting that acquisition of novel defence systems could be rare. These patterns are important, as frequent transfer may have major implications for phage therapy, where resistance could be quickly shared between isolates within a patient. Moreover, the CF pathogen presents a novel opportunity to study the dynamics of defence systems in a natural environment, with the ability to characterise the defence systems at clinically-relevant evolutionary timescales.

Across a *Vibrio* pangenome, 92% of defence systems were localised to variable genomic regions, and defence systems were enriched in putatively mobile variable regions (7). Additionally, 91% of defence systems in a set of 190 complete *E. coli* genomes were localised to variable genomic regions, with many regions classified as putative mobile genetic elements (MGEs) (21). These anlyses suggested that out of all the MGE types, prophages carried the most defence systems (7, 21). In *P. aeruginosa*, two variable genomic regions were recently discovered that contain defence systems in approximately 84% (cDHS1) and 15% (cDHS2) of isolates (19). In addition, examples were presented of highly similar cDHS1 regions between distant isolates and highly diverse cDHS1 regions between near-clonal isolates. Nevertheless, Johnson *et al.* did not characterise the overall proportion of systems localised to variable regions (19). Additionally, population-wide similarities and differences in all the defence systems that are localised to variable regions are yet to be studied in clinical populations. It is therefore unknown at what rate changes occur within such regions, and to what extent near-clonal isolates generally differ in these regions.

Marvig *et al.* previously sequenced a longitudinal collection of 474 CF isolates from 34 patients over the course of 1-10 years, providing the opportunity to study defence system dynamics in a natural and clinically important environment (22). In this work, the defence system repertoire of these isolates is described. Additionally, the dynamics of defence system gain and loss is studied within patients. Overall, this work suggests that the isolates carry a broad and open pan-immune system, with the number of defence systems associating with genome size. The number of defence systems was not found to positively associate with the number of prophages after accounting for genome size, which is in line with the discovery of infrequent localisation of defence systems to prophages. Nevertheless, the vast majority of defence systems are found in variable genomic regions. Grouping the variable regions into putative spots of insertion suggests that most regions are not hotspots specifically for defence systems. Additionally, distant isolates rarely have exactly matching defence systems within spots, and closely related isolates rarely differ in their systems within spots. Finally, the defence system repertoire appears to be stable through time within patients, suggesting that the clinical isolates do not undergo high turnover or recombination of defence systems. Overall, this work sheds light on the defence system dynamics in clinical CF *P. aeruginosa*, suggesting that in contrast to previous findings, the defence system repertoire is not in major flux.

2 Methods

2.1 Genomic data source

Nucleotide assembly, predicted protein sequence and annotation files were obtained for the collection of clinical CF isolates from NCBI's BioProject PRJEB5438. This project involved the longitudinal sampling of *P. aeruginosa* isolates from sputum samples of 34 CF patients, over the course of 1-10 years per patient (22). While the original sample included 474 isolates, only the 456 samples that were of sufficiently high quality to be included in RefSeq were analysed. Marvig *et al.* grouped the isolates into "Clone Types", representing groups of isolates with less than 10,000 SNPs between each pair (22). This data was obtained from the publication's supplementary materials (22).

2.2 Detecting putative defence systems

Putative defence systems were predicted using `PADLOC` and `DefenseFinder` (3, 4). Their outputs were wrangled and merged using `pandas` as follows (23). Firstly, predicted defence system genes were associated with their original RefSeq locus tags. Since DefenseFinder identifies hits by their position in the provided predicted protein sequence (`.faa`) file, the predicted protein sequence was used to pull the associated locus tag from an isolate's `gbff` annotation file using `seqtk` and `biopython` (24, 25). Since PADLOC identifies hits by their start and end in the assembly, these values were used to pull the associated locus tag from an isolate's `gff` annotation file using `gffutils` (26). Then, systems were renamed to ensure consistency between the programs. Commands were written to rename systems for cases where the two programs differed in naming conventions, given the present output. For example, it was found that DefenseFinder returned capitalised CRISPR-Cas system names with class and type/subtype designation, while PADLOC returned lowercase names with a type annotation, even if the system was in fact a substype (e.g. CAS_Class1-Subtype-I-F from DefenseFinder vs cas_type_I-F1 from PADLOC were renamed to cas_subtype_i-f). Subsequently, redundant and duplicated system annotations were removed. Systems were removed if their locus tags were included in another system annotation with the same system name and the same or additional locus tags. For example, isolate GCF_900143905.1 was assigned 'cas_subtype_i-f' by DefenseFinder, with genes Cas___cas8f_I-F_4 (locus tag BUR30_RS26530) and Cas___cas3f_I-F_1 (locus tag BUR30_RS26525). Since PADLOC identified cas_subtype_i-f with four additional genes/locus tags (Cas1f, Cas23f, Cas8f, Cas5f, Cas7f, Cas6f with locus tags BUR30_RS26520, BUR30_RS26525, BUR30_RS26530, BUR30_RS26535, BUR30_RS26540, BUR30_RS26545), the DefenseFinder annotation was discarded for the isolate. Family-level hits were also removed if a subfamily hit was available with the same or additional locus tags. For example, DefenseFinder defined subsystems for PARIS (paris_i and paris_ii), while PADLOC returned a family-level classification (paris). If a paris_i hit covered the same locus tags as a paris hit, the paris hit was discarded. Moreover, PADLOC assigned "other" classifications to some systems, e.g. 'DMS_other'. These represented fragmented putative systems, which would require manual verification (3), and were therefore excluded. Finally, PADLOC "phage defence candidate" (PDC) hits were also excluded due to their unverified nature. The pipeline developed for calling, renaming and filtering defence systems can be found at `https://github.com/ezherman/find-defence-systems`.

2.3 Collector's curve analysis

To determine whether the pan-immune system was open or closed, a collector's curve was estimated. To estimate the curve, showing the estimated number of defence system families identified for any number of bacterial samples, the rows in the presence-absence table of defence system families were randomly reshuffled 100 times. Then, the cumulative number of defence system families was calculcated for each reshuffled table as a function of row index. This data, as well as the mean cumulative number of defence system families per row index, was visualised using `ggplot2` in R (27, 28).

2.4 Associating the number of defence systems with genomic features

The number of defence system families were associated with genome size, number of contigs and number of prophage sequences using `R`'s `glm` function with the Poisson family (28). The regression lines were visualised using the `jtools` package (29). Summary statistics were obtained using the `summ` function from `jtools`. Genome size and the number of contigs were calculated using `quast` (30).

2.5 Detecting prophage sequences

Putative prophages were inferred using virsorter2, with a minimum length of 1500 bp and a minimum score of 0.8 (31). To find defence systems located in prophage sequences, locus tags were extracted from an isolate's gff file using the start and end coordinates of prophage sequences. Given that prophage coordinates could fall within CDS features, the final locus tag with start coordinates smaller or equal to that of the prophage sequence was taken as the start. Likewise, the first locus tag with end coordinates greater or equal to that of the prophage sequence was taken as the end. If the prophage start coordinates went beyond the first locus tag of the contig, the first locus tag was taken as the start. Similarly, if the prophage end coordinates were ahead of the last locus tag of the contig, the last locus tag was taken as the end. Locus tags in prophage sequences were then matched up with defence system locus tags, obtained as described above. If a defence system's locus tags matched the contents of multiple prophage hits, the prophage which contained the greatest proportion of the system's locus tag was taken forward.

2.6 Visualising prophage features

To construct a persistent genome phylogeny, multiple sequence alignments were obtained for the persistent portion of the pangenome obtained with PPanGoLIN (32). Using IQ-TREE, a phylogeny was constructed with the NONREV+FO model from nQMaker, 1000 bootstraps with UFBoot2 and the nuclear option for msub (33–35). The non-reversible model allowed the phylogeny to be rooted without the need for an outgroup. The phylogeny was visualised using the ggtree package for visualising the relatedness between isolates sharing the same prophage (36), alongside the prophage genes which were visualised with gggenes (37).

2.7 Localising defence systems to regions of genomic plasticity

To determine which defence systems were localised to variable genomic regions, suggestive of movement by horizontal gene transfer, the panRGP module of PPanGoLIN was used (38). This module identifies sequences of shell and/or cloud genes surrounded by persistent genes. Shell and cloud genes are found at intermediate and low frequencies within the pangenome, respectively, with genes assigned to partitions through a multivariate Bernoulli Mixture Model (32). PanRGP returns the locus tags associated with genes in RGPs. The module also returns spots into which the RGPs have been clustered, where possible. The locus tags associated with RGPs were paired with the locus tags of defence system genes, identified as described above.

To estimate the difference in the number of systems within RGPs that were or were not on contig borders, a generalised linear mixed model of the truncated negative binomial family was fit using glmmTMB (39). This model had a random intercept for isolate and a fixed effect for contig border. The truncated negative binomial family was chosen because the outcome variable, number of systems, had a minimum value of one. In addition, a model of the truncated poisson family showed overdispersion with the testDispersion function of the DHARMa package (40). The data was plotted using ggplot2 (27).

2.8 Identifying putative defence system hotspots

To identify putative hotspots, the number of unique defence system repertoires (ignoring surrounding genes) was graphed against the proportion of isolates that carried any genes in the spot. Then, spots were sized by the proportion of RGPs that contain at least one defence system. The graph was generated with ggplot2 (27). The spots corresponding to cDHS1 and cDHS2 regions were identified by the bordering MerR family transcriptional regulator and the bordering alpha/beta hydrolase, respectively (19), after parsing the isolate's gff files with ape (41). Manual inspection of the remaining bordering genes of the spots that were flanked by one of the two aforementioned genes revealed in both cases that the most abundant spot matched the hotspot identified by Johnson *et al.* (19).

2.9 Characterising defence system variation within spots

To find occurences of identical systems within spots between non-clonal isolates, Bray-Curtis distances within spots were calculated using the vegdist function from the vegan package (42). The output values were matched with their spots using the dist_groups function of the usedist package (43). Then isolate pairs were filtered for a Bray-Curtis distance of 0 and a phylogenetic distance greater than 0.001, based on the observed bimodal distribution of phylogenetic distances. The frequency of identical pairs within spots was visualised using ggplot2 (27). To find occurences of non-identical systems in near-clonal isolates, the same distance data was filtered for a Bray-Curtis distance greater than 0 and a phylogenetic distance smaller or equal to 0.001. To find cases where only one member of a near-clonal pair carried systems within a spot, variable regions with 0 defence systems were included in the `vegdist` computation, followed by exclusion of cases where both pair members carried 0 systems. Since there were no cases of near-clonal isolates with completely different defence systems within spots, a Bray-Curtis value of 1 could be used to identify cases where only one isolate carried systems within a spot. The relevant spots were visualised using the `ComplexHeatmap` package (44).

2.10 Estimating changes in defence system repertoire size over time within clone types and patients

To estimate whether the number of systems changed on average over time within closely-related isolates obtained from the same patient, a generalised linear mixed model of the generalised Poisson family was fit with `glmmTMB` (39). The generalised Poisson was chosen because a model of the Poisson family showed underdispersion with the `testDispersion` function of `DHARMa` (40). Closely-related isolates were grouped together using the "Clone Type" classification developed by Marvig *et al.*, described under the Genomic data source section above. The model had number of systems as the outcome variable and number of days since the first isolation of an isolate of the same clone type from the same patient as the fixed effect. The model also contained a random intercept for clone type, in addition to a random intercept and a random slope for clone type within patient. If less than 5 isolates were available for a clone type within a patient, isolates of that patient-clone type combination were excluded. In addition, Bray-Curtis distances were calculated between the first isolate of a clone-type within a patient and subsequent isolates using the `vegan` package (42). The number of systems over time was visualised using `ggplot2`, with generalised linear models of the Poisson family generated by the same package to show the marginal trends within clone types (i.e. not the exact marginal trends predicted by the model). Additionally, the Bray-Curtis distances were shown, with straight lines connecting data of isolates of the same clone-type over time.

3 Results

3.1 Clinical isolates carried many different defence systems

To study the defence system repertoires of clinical CF *P. aeruginosa* isolates, defence system presence was predicted using PADLOC and DefenseFinder (3, 4). Following deduplication of hits and renaming of systems, the isolates were predicted to carry defense systems belonging to 82 different families (Figure 1A). Unsurprisingly, RM and Cas were found amongst the most abundant system families (82% RM and 56% Cas). Interestingly, many of the most abundant system families were abortive infection systems (100% PD-T4-6; 48% CBASS; 41% Gabija). The two most prominent system families were single-gene systems: PD-T4-6 (100%) and SoFIC (89%). Most isolates appeared to be carrying between 4 and 16 defence system families (Figure 1B), between 4 and 18 defence system subfamilies (Figure 1C) and between 5 and 22 defence systems (Figure 1D). This suggests that these clinical isolates carried many different means of protection. Despite the wide range of system families, a collector's curve suggests that the pan-immune system of clinical CF *P. aeruginosa* is broader than described here (i.e. an open pan-immune system), as the curve does not appear to reach an asymptote (Figure 1E). Overall, while a few abundant systems stand out, the repertoire was mostly made up of systems at low frequency, and the data suggests that more families are left to be discovered in this bacterial population.

3.2 Genome fragmentation may have negatively biased defence system detection

Given the fragmented nature of some of the assemblies in this work, the association between the number of systems and the number of contigs was estimated. If fragmentation was resulting in systems being missed, there may be a negative association between the number of systems detected and the number of contigs. Across the full data, and when excluding assemblies composed of more than 500 contigs, there were positive associations between the number of contigs and the number of defence systems (Figure 1F,G) ($\beta_{\text{number of contigs}} = 3.3 \times 10^{-4}$, 95% CI: ($1.3 \times 10^{-4}$, 5.3×10^{-4}), p-value: 0.001; $\beta_{\text{number of contigs, filtered}} = 5.5 \times 10^{-4}$, 95% CI: ($2.7 \times 10^{-4}$, 8.3×10^{-4}), p-value: 9.2×10^{-5}). This suggested that fragmentation did not negatively bias the number of defence systems identified in this work.

3.3 Clinical isolates with larger genomes carried more systems

Previous work suggested that the number of systems is associated with genome size when considering all species with complete genomes in RefSeq (4). Additionally, the report found isolates with more prophages on average to carry more defence systems (4). While prophage presence suggests that the host could not protect itself from infection, prophages could also provide the host with new defence systems (45). To ask whether the number of defence systems carried by the clinical isolates was associated with genome size and/or the number of prophages, firstly univariate poisson GLMs were fit to the data. At this level, both variables were positively associated with the number of systems (Figure 1H,I). However, given the positive association between genome size and the number of prophages (Figure 1J), it was important to account for the effect of genome size.

After accounting for genome size, the number of prophages was assigned a small negative effect ($\beta_{\text{scaled size}} = 0.26$, 95% CI: (0.22, 0.29), p-value: 10^{-46}; $\beta_{\text{scaled number of prophages}} = -0.051$, 95% CI: (-0.085, -0.017), p-value: 0.0035). This suggests that after accounting for genome size, isolates with more prophages tended to have slightly fewer defence systems. Importantly, the variance inflation factor (although operating on linear correlation) was not sufficiently large to suggest that the effect of the number of prophages changed due to collinearity (VIF: 1.9). Overall, this analysis suggested that isolates with larger genomes tended to carry more defence systems, while the number of prophages was not positively associated with the number of systems after accounting for genome size.

3.4 Defence systems were infrequently found in prophage sequences

Given the lack of a positive association between the number of prophages and the number of defence systems after accounting for genome size, the defence system content of prophage sequences was investigated. Out of the 2106 prophage sequences identified, 168 contained at least one defence system and a further 9 contained defence systems partially. Out of the 5265 identified defence systems, 187 were localised to prophage sequences and 10 had a subset of genes that were localised to prophage sequences. While Lamassu systems were the most frequently found in prophage sequences, other systems were found in multiple prophages too (Figure 2A). The prophage sequences did not appear to be hotspots for defence systems, as most carried only one defence

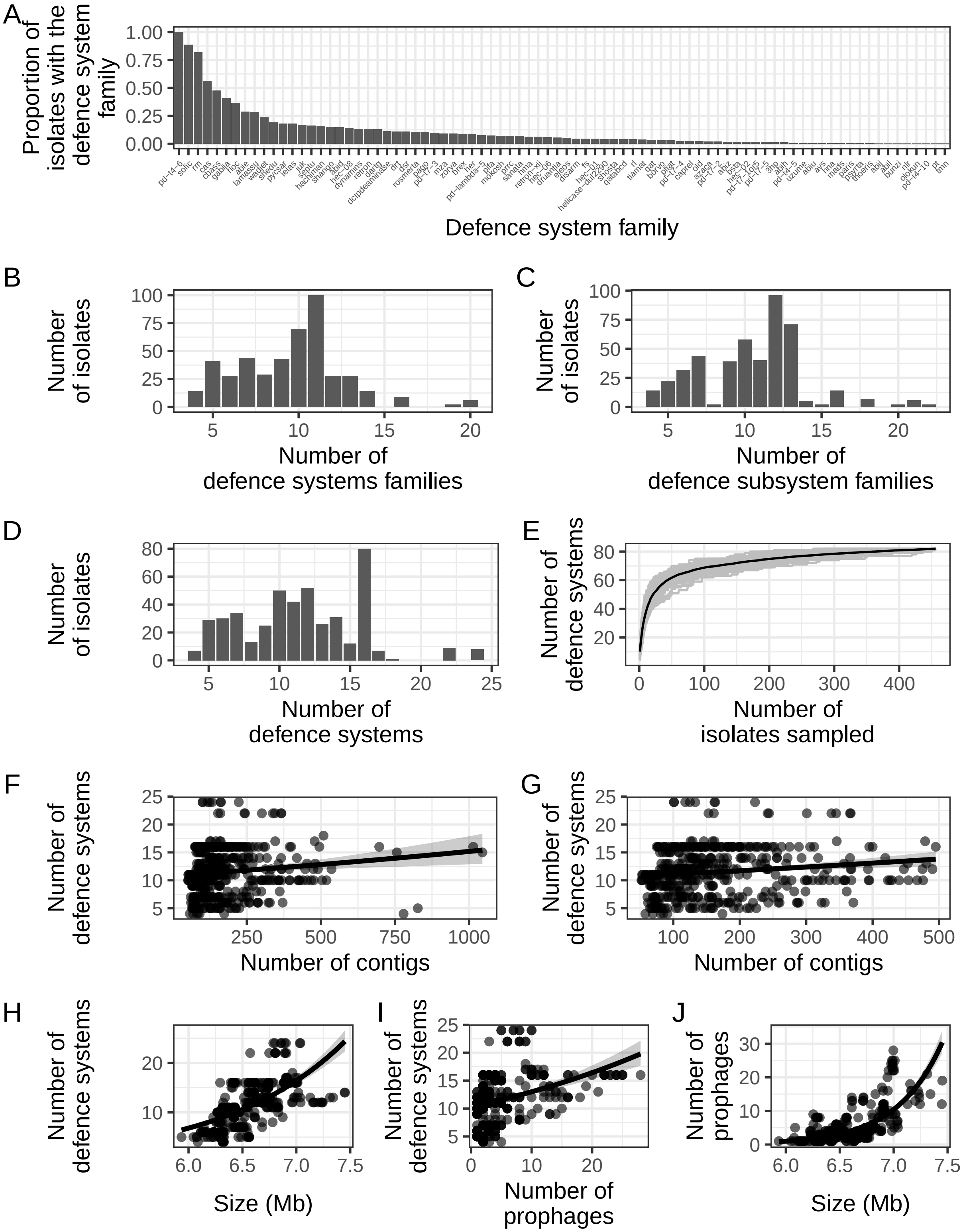

Figure 1: Describing the defence system repertoire of clinical *P. aeruginosa* isolates from CF patients. A) The frequency of defence system families varied widely within this group of isolates, with PD-T4-6, SoFIC and RM being the most frequently found. B,C,D) Isolates generally carried many different defence system families, subfamilies and individual systems. E) A collectors curve suggests that more families are left to be discovered in this population. F,G) There did not appear to be a negative association between the number of systems and genome fragmentation. H,I) Both assembly size and the number of prophages appeared to associate positively with the number of defence systems in univariate graphs. J) The number of prophages was associated with genome size through a poisson GLM.

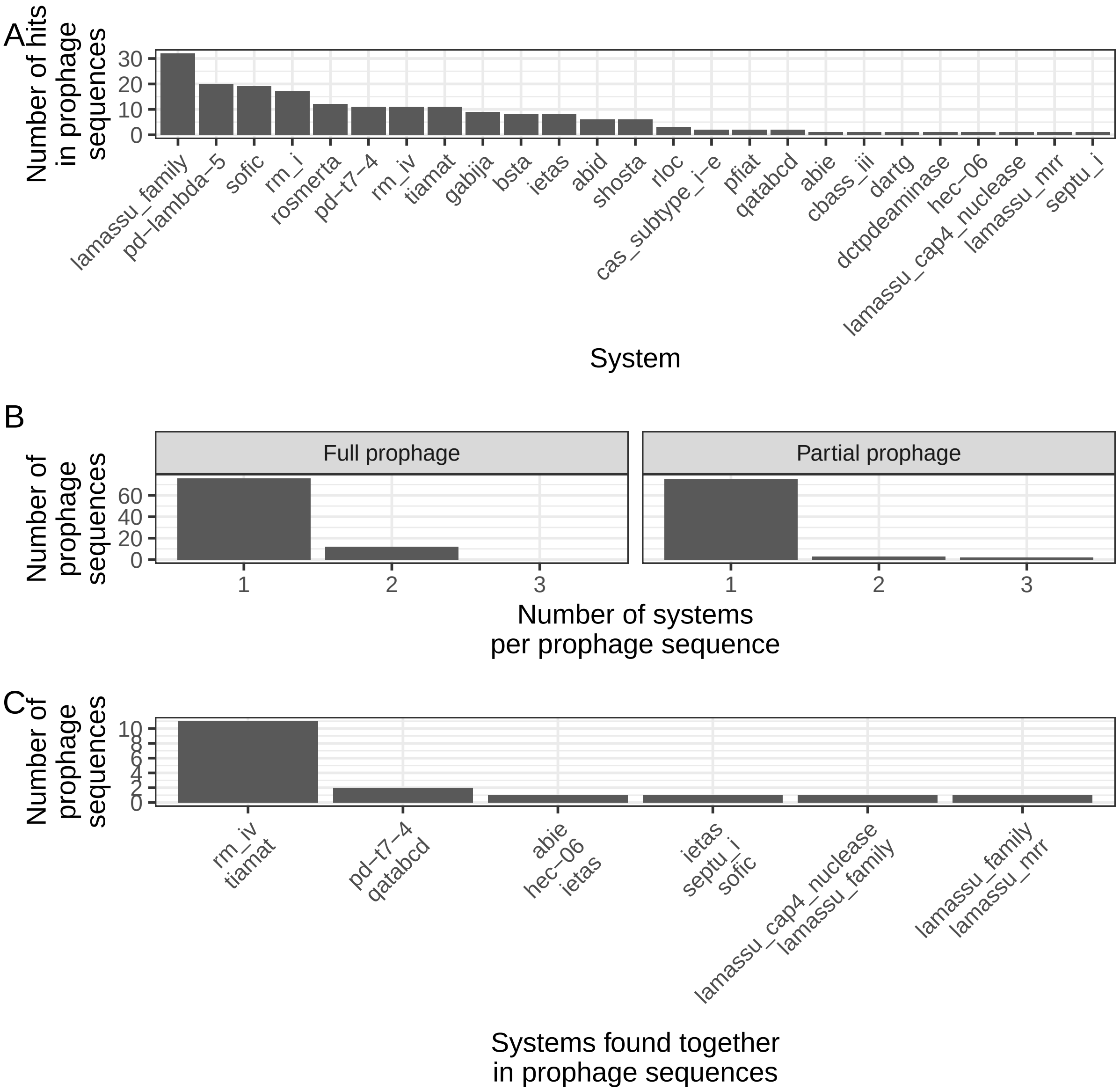

Figure 2: Describing the minority of defence systems found in prophage sequences. A) Lamassu family systems were most frequently found in prophages. B) Regardless of whether prophage sequences were full or partial, sequences mostly carried no more than one system. C) Although prophage sequences with multiple systems were rare, the combination of Tiamat and RM Type IV was most frequently found.

system (Figure 2B). This was irrespective of partial or full prophage classification. Finally, although prophage sequences with multiple systems were rare, out of the detected combinations of systems the pair of Tiamat and RM Type IV was most frequently found (Figure 2C). This combination was found in a group of closely-related isolates (Figure 3A,B). Alongside the defence system genes, an example of one of these prophages showed many genes of unknown function and mobile genetic element genes (Figure 3C). Not much is known about Tiamat, besides it being a one-gene system with an HSp90 domain alongside two domains of unknown funciton (DUF3883) (46). Overall, some examples of system-containing prophages could be found, however the low frequency of defence systems in prophage sequences was in line with the lack of a positive association between the number of defence systems and the number of prophages after accounting for genome size.

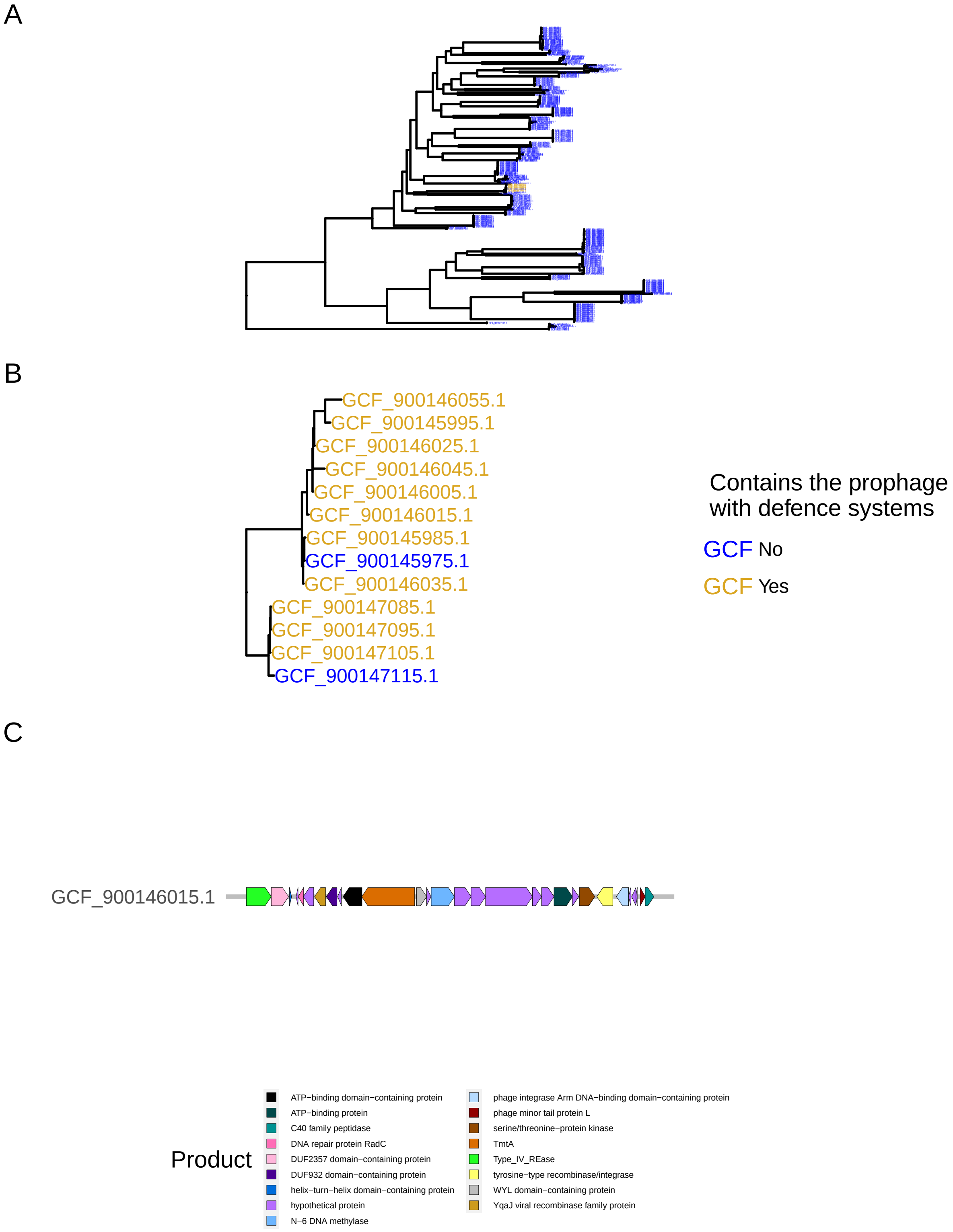
A
B
GCF_900146055.1
GCF_900145995.1
GCF_900146025.1
GCF_900146045.1
GCF_900146005.1
GCF_900146015.1
GCF_900145985.1
GCF_900145975.1
GCF_900146035.1
GCF_900147085.1
GCF_900147095.1
GCF_900147105.1
GCF_900147115.1
Contains the prophage with defence systems
GCF No
GCF Yes
C
GCF_900146015.1
Product
ATP−binding domain−containing protein
ATP−binding protein
C40 family peptidase
DNA repair protein RadC
DUF2357 domain−containing protein
DUF932 domain−containing protein
helix−turn−helix domain−containing protein
hypothetical protein
N−6 DNA methylase
phage integrase Arm DNA−binding domain−containing protein
phage minor tail protein L
serine/threonine−protein kinase
TmtA
Type_IV_REase
tyrosine−type recombinase/integrase
WYL domain−containing protein
YqaJ viral recombinase family protein

Figure 3: The most frequent combination of defence systems in prophages contained RM Type IV and Tiamat. A) This combination of systems inside prophages was only found in a group of closely-related isolates. B) Nevertheless, not all isolates from this clade carried the defence systems. C) An example of the prophage showed co-localisation of the defence systems with mobile genetic element genes and many genes of unkown function

3.5 Defence systems in regions of genomic plasticity

Regions of genomic plasticity (RGPs) are defined as stretches of shell and/or cloud genes, flanked by stretches of persistent genes (38). Since many defence systems were originally identified in variable genomic regions, and defence systems are thought to spread by horizontal gene transfer, the localisation of defence systems to variable regions was studied in the present clinical isolates (1). Firstly, the majority of defence systems were found fully within RGPs, with a minority being found completely outside of RGPs, partially within RGPs or fully across multiple RGPs (Figure 4A). Note that 58 system hits were not considered by panRGP since the program excludes genes labeled as pseudogenes (47). Out of 23675 RGPs, 2908 contained at least one defence system. While most of those regions contained a single system, larger collections of systems within RGPs were also identified (Figure 4B). Note that approximately 52% of these RGPs were located on contig borders, suggesting they may contain more systems. Moreover, the difference in mean number of systems was significantly different between RGPs that were and were not on contig borders, suggesting the number of systems per RGP may have been negatively biased by genome fragmentation (estimated mean number of systems in RGPs not on contig borders: 1.62; estimated mean number of systems in RGPs on contig borders: 1.46; p-value: 2.1×10^{-5})(Figure 4C). Overall, most defence systems were localised to variable genomic regions, most commonly in isolation from other defence systems.

3.6 Defence system hotspots

RGPs across isolates can be grouped into regions of putative insertion, known as “spots”, by clustering RGPs with similar flanking persistent genes together (38). Out of 2908 RGPs that contained at least one defence system, 1305 could be assigned to a spot (45%). Note that RGPs on contig borders by definition could not be assigned a spot, as they were only bound by persistent genes on one side. Therefore, almost all RGPs that were not located on a contig border could be assigned a spot. To ask whether spots may represent defence system hotspots, first the number of distinct RGPs per spot was studied. Spots varied in the number of unique RGPs that contained defence systems (Figure 4D). Interestingly, the proportion of unique RGPs that contained defence systems decreased with the number of unique identified RGPs, suggesting that most spots were not specific insertion sites for defence systems (β: -0.08, 95% CI: (-0.1, -0.06), p-value: 1.6×10^{-13}) (Figure 4E).

The recently discovered core Defence system HotSpots cDHS1 and cDHS2 (19) were matched to the spots identified by panRGP, to ask whether the hotspots were any different from the other spots. cDHS1 and cDHS2 were identified as genomic regions flanked by core genes, within which isolates varied in defence system repertoire (19). While no official definition of a hotspot was found in the literature, it may be defined as a location that is occupied in many isolates with variable genes, usually with at least one defence system, and with relatively high variability in the defence system repertoires across isolates. Such spots can be identified by graphing the number of unique defence system repertoires against the proportion of isolates which carry genes in a region, with points sized by the proportion of variable regions that carry at least one defence system (Figure 4F). This graph suggested that cDHS1 was a genuine hotpot in these isolates (proportion of variable regions that contained defence systems: 0.79, proportion of isolates that carried genes in this spot: 0.29, number of unique defence system repertoires: 12). In contrast, very few isolates that showed cDHS2 carried defence systems in the spot (proportion of variable regions that contained defence systems: 0.091, proportion of isolates that carried genes in this spot: 0.7, number of unique defence system repertoires: 7). The low proportion of variable regions that contained defence systems in cDHS2 was in line with the 15% figure identified originally (19), and suggested that this may be an “ordinary” variable region, albeit one that was present in many isolates. Interestingly, Figure 4F highlighted a spot with 11 different defence system repertoires, found in 35% of isolates and with 35% of regions containing at least one defence system. The proportion of RGPs within this spot that contained defence systems was between the proportions of cDHS1 and cDHS2. The spot may or may not be considered a novel, previously unidentified hotspot. Nevertheless, this analysis suggested that cDHS1 was the only spot that could be called a defence system hotspot with high confidence.

3.7 Spots suggested infrequent sharing of large defence system modules between non-clonal isolates

The spots highlighted regions in which defence systems, transferred horizontally, may integrate. Under HGT, some distantly related isolates would be expected to show the same defence systems within the same spot. Overall, amongst all shared occurences of a spot in non-clonal isolates where both isolates carried system(s) in the spot, 36% shared an identical repertoire, 28% shared a partially identical repertoire and 35% shared

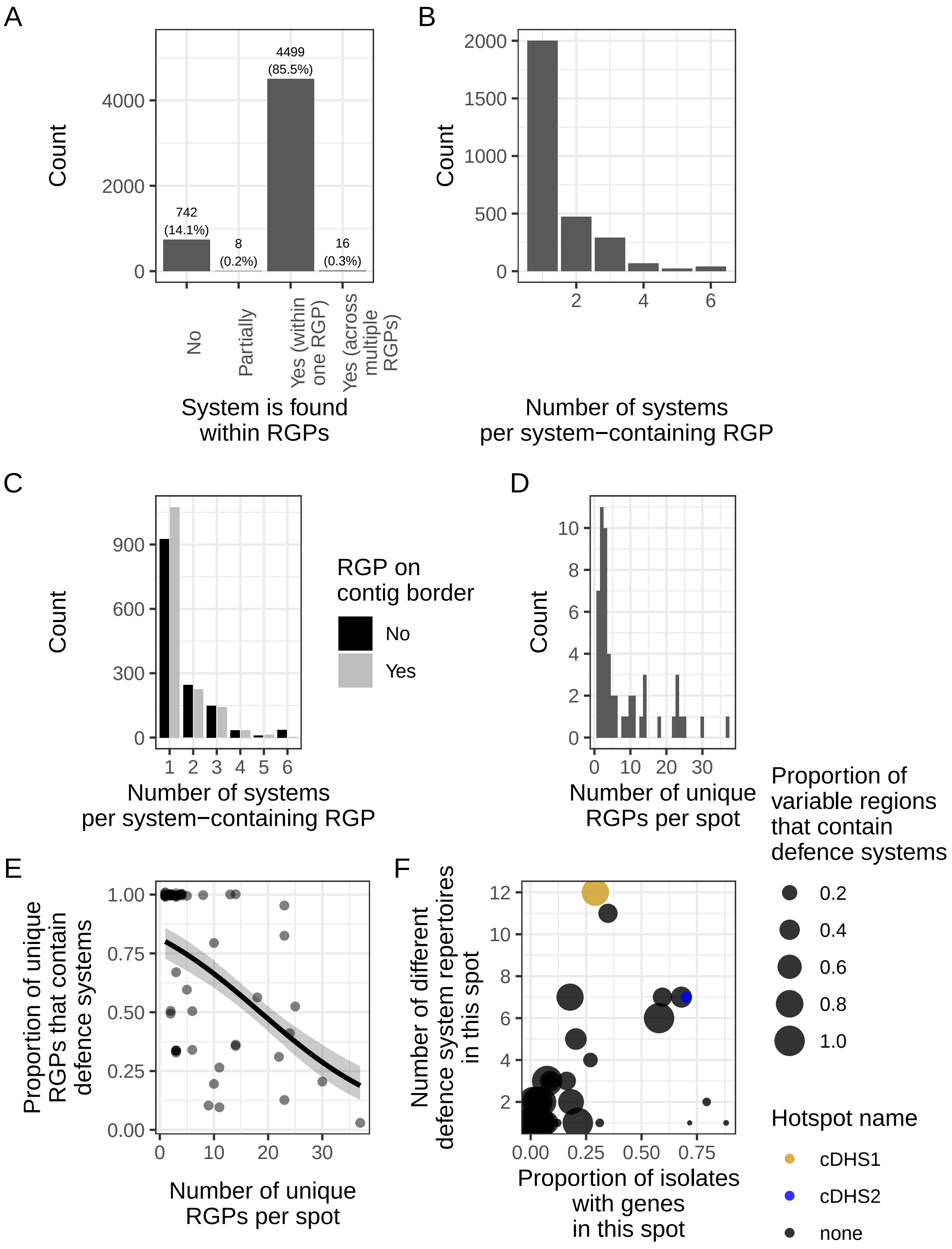

Figure 4: Defence systems in RGPs. A) The majority of defence systems were found in RGPs. B) The majority of RGPs that contained defence systems, only contained one system. C) RGPs on contig borders on average contained fewer defence systems, potentially driven by a greater frequency of single-system occurrences. D) When considering RGPs that could be clustered into spots, spots varied in the number of unique RGPs that they grouped together. E) However, the proportion of RGPs that contained defence systems appeared to decrease with the number of unique RGPs. F) When considering abundance of spots across isolates, the number of different system repertoires within spots and the proportion of RGPs that contained defence systems, only the recently discovered cDHS1 appeared to be a genuine defence system hotspot.

no defence systems at all. Shared repertoires were mostly limited to two single-system spots (Figure 5A). Spot 1 accounted for most of the cases of identical systems amongst non-clonal isolates (70.5%) and partial overlap (84.1%), with the vast majority of spots carrying SoFIC and some carrying one or two additional systems (abiD, Lamassu, RM Type III or Septu + ietAS). Spot 58 accounted for a further 18.7% of pairs and exclusively contained CRISPR I-F (i.e. there were no cases of partially shared defence system repertoires). Excluding these spots, most cases showed no shared defence system repertoire (identical repertoire: 10%; partially shared repertoire: 12%; non-identical repertoire: 78%). Overall, this data suggests that besides two spots that often contained one or more shared systems, most spots showed non-identical defence system repertoires amongst non-clonal isolate pairs.

3.8 Closely-related isolates mostly carried the same systems within spots, except for a few signals of potential recent insertion

The low frequency of shared defence systems within spots between non-clonal isolates, described above, could be a consequence of stringent segregation of patients with CF. Under HGT, (near)-clonal isolates would be expected to show variation in defence systems within spots, suggestive of recent defence system movement. When focusing on pairs of clonal isolates in which both carried defence systems within a spot, the vast majority of pairs carried exactly the same systems (exactly the same systems: number of pairs = 9207, 98%; some shared systems: number of pairs = 209, 2%; completely different systems: number of pairs = 0, 0%). This suggested that recent defence system mobility was extremely rare in this sample set, spanning multiple years per clone type. However, these statistics excluded cases where only one pair member carried systems within the spot. Such cases may represent recent acquisition of defence systems, rather than recent changes in the within-spot repertoire. In total, there were 317 pairs across six spots within which only one member of the pair carried defence systems (Figure 5B). Within most spots, the frequency of defence system regions was far lower than the frequency of regions without systems, suggesting that systems were recently acquired in these spots. Overall, while clonal isolates with systems in spots usually shared their systems with other clonal isolates, a few spots showed signs of possible recent acquisition of defence systems via horizontal gene transfer.

3.9 Systems were rarely gained or lost within patients over time

Given the possibility that isolates were under phage selection within patients, the dynamics of system gain and loss were investigated at the patient level. Marvig *et al.* previously grouped the present isolates into 'clone types', which were defined as groups of isolates that differed in less than 10,000 SNPs (22). Using this data, the number of systems within closely related isolates was studied within patients over time. Absence of a significant change in the the number of systems carried by clone types within patients over time could not be rejected (β_{days} = -0.0046, 95% CI: (-0.013, 0.0035), p-value: 0.27). The raw data also showed that for the majority of clone type and patient combinations, the average number of systems was stable over time (Figure 6). Additionally, Bray-Curtis distances between the first isolate of a clone type within a patient and subsequent isolates were mostly 0, indicating that the stable number of systems arose from a stable system repertoire (Figure 6). Interestingly, Figure 6 also suggested that some within-clone type variation in defence system repertoire size was explained by patient origin. Overall, this analysis suggested that the defence system repertoire of *P. aeruginosa* isolates within patients was largely stable through time.

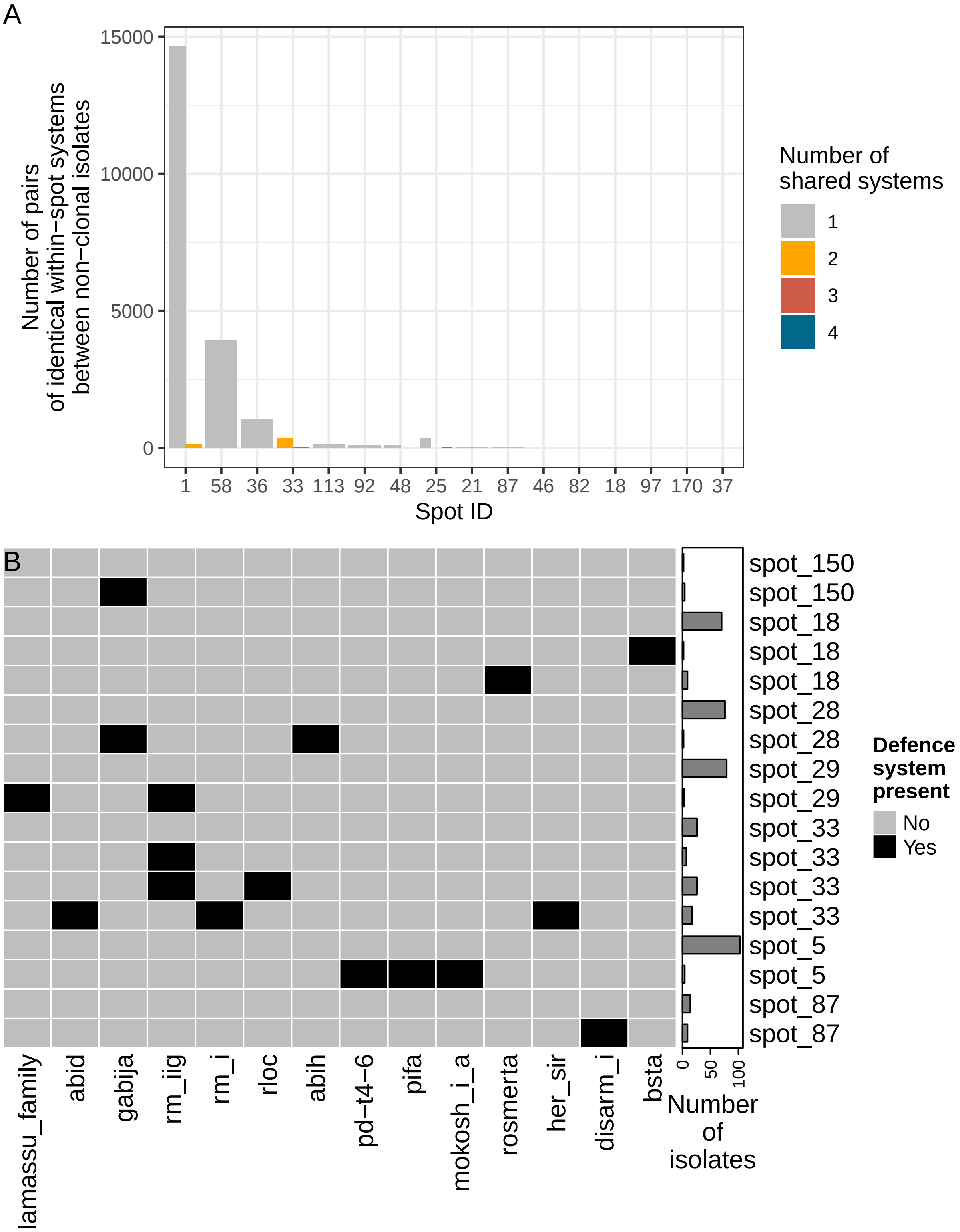

Figure 5: Defence systems similarities and differences in spots. A) Spots which carried identical defence system repertoires across non-clonal isolates. In most cases in which non-clonal isolates had identical within-spot system repertoires, only a single defence system was shared. Note that spot 82 corresponded to cDHS1, spot 46 corresponded to the potential new hotspot, described above, and cDHS2 (spot 29) did not show identical system repertoires amongst non-clonal isolates. B) Spots amongst which some near-clonal isolates carried no defence systems, while others carried some defence systems. Where a spot is represented by more than two rows, the system-containing rows may refer to system carriage by separate groups of near-clonal isolates.

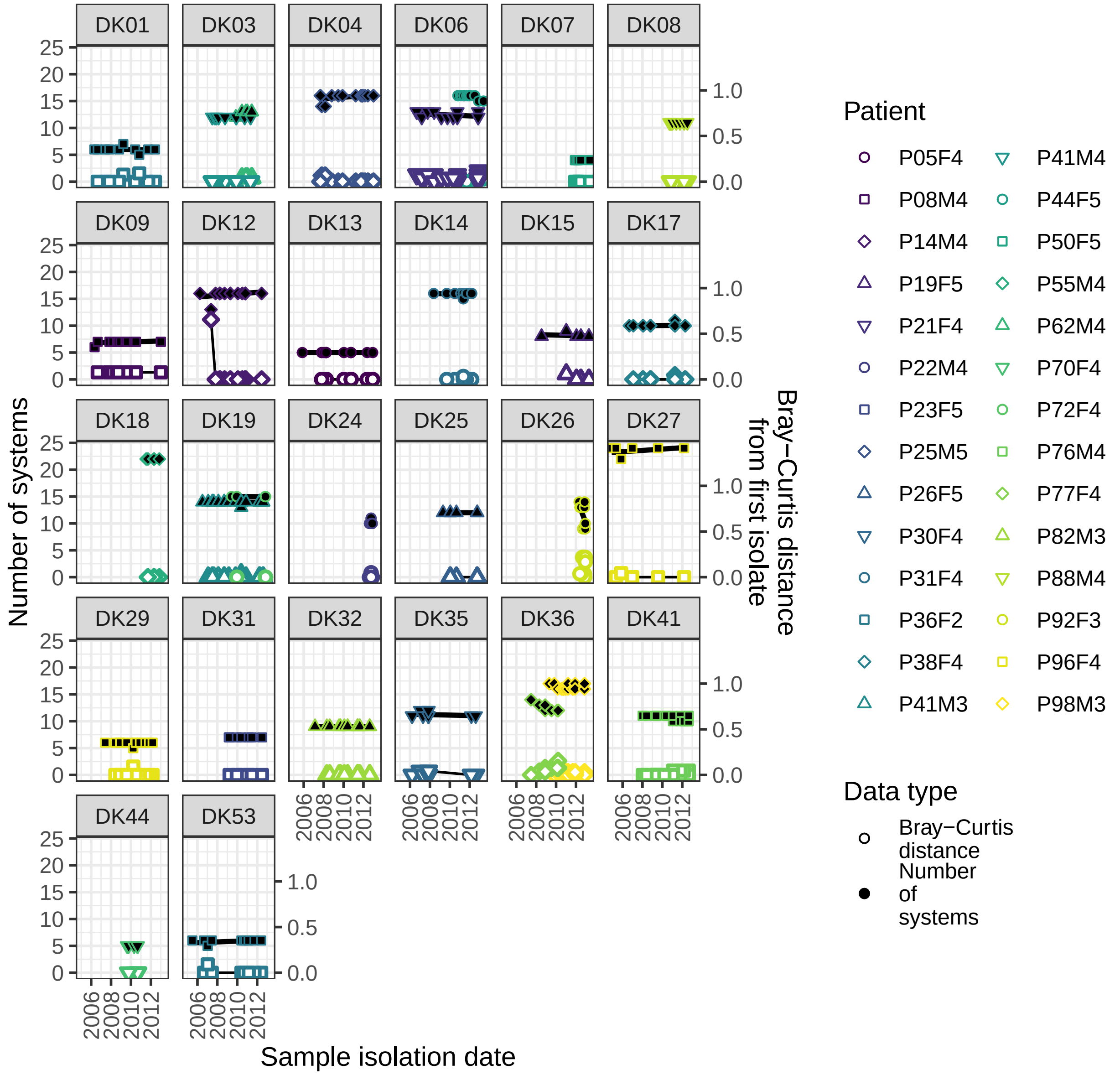

Figure 6: The number of systems stayed stable over time in most combinations of clone types and patients. Additionally, Bray-Curtis distances between the first and subsequent isolates were mostly 0. Note that this plot excluded any clone type and patient combinations for which less than 5 isolates were available.

4 Discussion

Defence systems are thought to be in constant flux, with bacteria harbouring different subsets of a population's pan-immune system over time (2). This study set out to characterise the defence system repertoire of clinical CF *P. aeruginosa* isolates, its localisation to prophages and variable genomic regions, its mobility and its changes within patients over time. The clinical isolates showed signs of a broad and open pan-immune system, with more system families set to be detected despite 456 isolates studied here. Isolates with larger genomes tended to carry more defence systems. In contrast, the number of defence systems did not positively associate with the number of prophages after accounting for genome size. This was likely explained by the small proportion of prophage sequences carrying defence systems, and the small proportion of defence systems found in prophage sequences. The majority of systems were localised to regions of genomic plasticity (RGPs), with most RGPs containing only one system. Grouping of RGPs into spots suggested that most spots were not defence system hotspots, except for cDHS1 (with high confidence) and one newly identified spot (with less confidence). Additionally, the spots suggested that sharing of more than one defence system between distant isolates was rare within spots, and that closely related isolates almost always carried the same systems within spots. This was in line with most clone types appearing to have a stable defence system repertoire through time within patients. Overall, the defence system arsenal of clinical CF *P. aeruginsa* isolates appeared mostly clade-specific, with little evidence for changes within patients or horizontal movement between distantly-related isolates.

The clinical isolates were found to carry systems from 82 different families. Notably, the collector's curve did not show signs of reaching an asymptote, suggesting that further systems are to be identified in clinical CF isolates. This is in line with the collector's curve generated previously across isolates of the *Pseudomonas* genus (19). While data from Tesson *et al.* suggested that the number of defence systems was greater in 6 Mb genomes than in 5Mb genomes, with no difference between genomes of 6 Mb and those greater than 6Mb, their model was fit across all RefSeq species in their sample set. In the present analysis, the number of systems appeared to associate with genome size, given a range of approximately 6 to 7.5 Mb. While this might suggest that *P. aeruginsa* are particularly good at accumulating defence systems, in *Vibrio* the association between the number of defence systems and genome size vanished after accounting for strain phylogenetic relatedness (7). If genome size variation is low between closely-related CF *P. aeruginosa* isolates, future work accounting for phylogenetic relatedness may find that the effect of genome size disappears.

The present analysis detected numerous defence systems at high abundance. Of note, all isolates were estimated to carry the PD-T4-6 system, a single-gene abortive infection system with unknown function (48). Additionally, approximately 89% of isolates were estimated to carry SoFIC, which also is a single-gene system with unknown function (46). Strikingly, both systems were found at higher abundance than the RM, CRISPR-Cas, Gabija and CBASS systems, which were previously considered the most abundant systems in *P. aeruginosa* (4, 19). PD-T4-6 and SoFIC were recently found in similar abundances across a 100-isolate panel of clinical *P. aeruginosa* isolates (49). However, given the risk of false-positives when identifying single-gene systems, future work could assess the ability of these hits to provide phage resistance, as well as their variation within *P. aeruginosa.* Also of note, the RloC system was discovered 15 years ago (50) and was identified in approximately 36% of isolates, however it has not yet been reported in *P. aeruginsa.* This system is activated through detection of double-stranded breaks, after which tRNA molecules are cleaved to halt translation (51). Its prominence in the clinical isolates suggests that this system is important for abortive infection in the clinical isolates, however this too is a single-gene system which may suffer from a risk of false-positive identification.

Previous work has suggested that across bacterial species, isolates with more prophages carry more defence systems (4). However, this association appeared to be limited to isolates with up to 6 prophages, with isolates with more prophages grouped together (4). In this work, where many isolates were estimated to carry more than 6 prophages, the number of systems were not positively associated with the number of prophages after accounting for genome size. In contrast, a small negative association was found. While this may suggest that isolates with fewer defence systems are less able to protect themselves from prophages, the small effect size warrants further analysis on a larger data set. In line with the lack of a positive association, very few prophages carried defence systems, and only a small proportion of defence systems were found in prophages. This is in contrast with previous *Vibrio* and *E. coli* pangenomes, where prophages frequently carried defence systems and where they were the MGE type most enriched in defence systems (7, 21). Notably, 70% of prophage genes were found to be lost in the present isolate collection over time within patients (20). Gabrielaite *et al.* also found plasmids to be rare in the isolate collection, which appeared to be a property of *P. aeruginsa* more generally (20). Future work could estimate the proportion of systems localised to integrative and conjugative elements (ICEs) and Integrative Mobilizable Elements (IMEs), which have previously been found to carry defence systems (6, 21), as well as transposons (21), which are prominent in *P. aeruginosa*

(52).

To ask whether defence systems tended to be located in variable genomic regions, RGPs were identified using the panRGP module of PPanGGOLiN (32). While most defence systems were localised to RGPs, RGPs containing more than one or two systems were rare. Unfortunately the average number of systems in RGPs was significantly associated with whether the RGP was on a contig border, suggesting that genome fragmentation negatively biased the number of systems that were found within RGPs. Additionally, it is possible that larger RGPs, carrying more systems, were more likely to be on contig borders. Future work could ask whether larger defence system RGPs are found when using less fragmented long-read assemblies. Additionally, future work could apply the panModule function of PPanGGOLiN, to ask whether systems tend to co-occur in variable regions, suggestive of synergism (53, 54). Either way, the localisation of most defence systems to variable regions was in line with the same finding in *Vibrio* and *E. coli* (7, 21). However, given the small proportion of RGPs that contained defence systems, this data suggests that defence systems made up a small part of the variable genome of clinical *P. aeruginsa*, unlike one *Vibrio* species where defence systems made up a large part (5).

To ask whether RGPs may have been localised to hotspots, RGPs were grouped into putative spots of insertion using the panRGP module of PPanGGOLiN (32). The fragmented nature of the assemblies precluded many RGPs from being clustered into spots, as the panRGP module required matching persistent genes on both RGP borders for clustering. Future work using long-read assemblies may assign more RGPs to any one spot. Nevertheless, most RGPs that were not on contig borders could be clustered into a spot. The data suggested that defence system hotspots were rare. Firstly, the proportion of RGPs containing defence systems decreased with the number of different RGPs. Secondly, when considering the proportion of isolates carrying a spot, the number of different defence system repertoires within the spot, and the proportion of RGPs that contained defence systems, only the recently discovered cDHS1 appeared to be a hotspot (19). The present analysis probably identified more spots than Johnson *et al.* for two reasons. Firstly, Johnson *et al.* selected putative flanking genes by manual inspection of a gene co-occurence network, while panRGP automated this process. Secondly, Johnson *et al.* grouped variable regions which had a single matching gene on each border together, while panRGP clustered RGPs together if either their first flanking genes were of the same gene family or two out of their first three flanking genes were of the same gene family. Crucially, the present analysis highlights the need to evaluate the definition of a 'hotspot'. Johnson *et al.* defined two regions as hotspots, although only one of the regions contained defence systems in most of their isolates (cDHS1: 84%, cDHS2: 15%) (19). The present analysis suggests that cDHS2, with its low proportion of RGPs containing defence systems and its total number of unique defence system repertoires, was more likely to be a "regular" spot. Future work could consider the variables proposed here on a larger genomic data collection, to develop a formal definition of a defence system hotspot.

The finding that defence systems within spots varied little between closely-related isolates, and rarely fully overlapped between distant isolates (with the exception of two spots that mostly carried single systems), suggests that the clinical isolates were not frequently exchanging defence system modules during chronic infection of the CF lung. However, this finding needs to be considered given that the majority of spots only carried single systems. Since some of the isolates of different clone types were isolated from the same patient, future work could assess whether the few occurences of identical spots between distant isolates correspond to isolates that were obtained from the same patient.

A unique feature of the present isolate collection was that closely related isolates were obtained from the same patient over time, allowing for defence system dynamics to be studied at a microevolutionary scale. Surprisingly, in contrast to previous findings of defence systems in flux in *Vibrio* species (5–7), the defence system repertoire of clone types within patients was on average stable through time. Both in terms of the number of systems, as well as Bray-Curtis distance from the first detected isolate. Future work characterising the phage diversity of sputum samples over time could reveal whether this diversity is low, which would suggest that selection for novel defence systems is low. Additionally, given that *P. aeruginosa* undergoes many changes in phage receptor genes within the CF lung (55), it is possible that receptor modifications are driving phage resistance in this environment. Finally, clinical CF isolates are possibly exposed to fewer defence systems, given the relatively closed nature of the human lung in comparison to the sewage and ocean environments of *Vibrio* (5–7). Nevertheless, given the diversity of defence system repertoires between more distantly-related isolates, it is possible that defence systems are more frequently gained and lost in the natural environment (56), from which patients are likely to acquire the pathogen (57). Future work could longitudinally sample *P. aeruginsa* from soil or water reservoirs, to ask whether the defence system repertoire changes over time in these non-clinical environments.

5 References

1. Georjon H, Bernheim A (2023) The highly diverse antiphage defence systems of bacteria. *Nat Rev Microbiol*:1–15.
2. Bernheim A, Sorek R (2020) The pan-immune system of bacteria: Antiviral defence as a community resource. *Nature Reviews Microbiology* 18(2):113–119.
3. Payne LJ, et al. (2021) Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Research* (gkab883). doi:10.1093/nar/gkab883.
4. Tesson F, et al. (2022) Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun* 13(1):2561.
5. Hussain FA, et al. (2021) Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science* 374(6566):488–492.
6. LeGault KN, et al. (2021) Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. *Science*. doi:10.1126/science.abg2166.
7. Piel D, et al. (2022) Phage–host coevolution in natural populations. *Nat Microbiol*:1–12.
8. Cahier K, et al. (2023) Environmental vibrio phage–bacteria interaction networks reflect the genetic structure of host populations. *Environmental Microbiology*. doi:10.1111/1462-2920.16366.
9. Le Roux F, Blokesch M (2018) Eco-evolutionary Dynamics Linked to Horizontal Gene Transfer in Vibrios. *Annual Review of Microbiology* 72(1):89–110.
10. Thornton CS, Parkins MD (2023) Microbial Epidemiology of the Cystic Fibrosis Airways: Past, Present, and Future. *Semin Respir Crit Care Med* 44(02):269–286.
11. Rowbotham NJ, Palser SC, Smith SJ, Smyth AR (2019) Infection prevention and control in cystic fibrosis: A systematic review of interventions. *Expert Review of Respiratory Medicine* 13(5):425–434.
12. Westra ER, et al. (2015) Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Current Biology* 25(8):1043–1049.
13. Betts A, Gifford DR, MacLean RC, King KC (2016) Parasite diversity drives rapid host dynamics and evolution of resistance in a bacteria-phage system. *Evolution* 70(5):969–978.
14. Betts A, Gray C, Zelek M, MacLean RC, King KC (2018) High parasite diversity accelerates host adaptation and diversification. *Science* 360(6391):907–911.
15. Wright RCT, Friman V-P, Smith MCM, Brockhurst MA (2018) Cross-resistance is modular in bacteria–phage interactions. *PLOS Biology* 16(10):e2006057.
16. Wright RCT, Friman V-P, Smith MCM, Brockhurst MA (2019) Resistance Evolution against Phage Combinations Depends on the Timing and Order of Exposure. *mBio* 10(5). doi:10.1128/mBio.01652-19.
17. Nordstrom HR, et al. (2022) Genomic characterization of lytic bacteriophages targeting genetically diverse Pseudomonas aeruginosa clinical isolates. *iScience* 25(6). doi:10.1016/j.isci.2022.104372.
18. Castledine M, et al. (2022) Parallel evolution of Pseudomonas aeruginosa phage resistance and virulence loss in response to phage treatment in vivo and in vitro. *eLife* 11:e73679.
19. Johnson MC, et al. (2023) Core defense hotspots within Pseudomonas aeruginosa are a consistent and rich source of anti-phage defense systems. *Nucleic Acids Research*:gkad317.
20. Gabrielaite M, Johansen HK, Molin S, Nielsen FC, Marvig RL (2020) Gene Loss and Acquisition in Lineages of Pseudomonas aeruginosa Evolving in Cystic Fibrosis Patient Airways. *mBio* 11(5). doi:10.1128/mBio.02359-20.
21. Hochhauser D, Millman A, Sorek R (2023) The defense island repertoire of the Escherichia coli pan-genome. *PLOS Genetics* 19(4):e1010694.
22. Marvig RL, Sommer LM, Molin S, Johansen HK (2015) Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. *Nature Genetics* 47(1):57–64.
23. McKinney W (2010) Data Structures for Statistical Computing in Python (Austin, Texas), pp 56–61.
24. Li H lh3/seqtk: Toolkit for processing sequences in FASTA/Q formats. *seqtk*. Available at: `https://github.com/lh3/seqtk` [Accessed August 12, 2023].

25. Cock PJA, et al. (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422–1423.

26. Dale R Daler/gffutils: GFF and GTF file manipulation and interconversion. *gffutils*. Available at: `https://github.com/daler/gffutils` [Accessed August 12, 2023].

27. Wickham H (2016) *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York) Available at: `https://ggplot2.tidyverse.org`.

28. R Core Team (2023) *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria) Available at: `https://www.R-project.org/`.

29. Long JA (2023) Jtools: Analysis and Presentation of Social Scientific Data. Available at: `https://cran.r-project.org/web/packages/jtools/index.html` [Accessed July 22, 2023].

30. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 29(8):1072–1075.

31. Guo J, et al. (2021) VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 9(1):37.

32. Gautreau G, et al. (2020) PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLOS Computational Biology* 16(3):e1007732.

33. Minh BQ, et al. (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* 37(5):1530–1534.

34. Dang CC, et al. (2022) nQMaker: Estimating Time Nonreversible Amino Acid Substitution Models. *Systematic Biology* 71(5):1110–1123.

35. Hoang DT, Chernomor O, Haeseler A von, Minh BQ, Vinh LS (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* 35(2):518–522.

36. Yu G, Smith D, Zhu H, Guan Y, Lam TT-Y (2017) Ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8(1):28–36.

37. Wilkins D (2020) *Gggenes: Draw gene arrow maps in 'ggplot2'* Available at: `https://CRAN.R-project.org/package=gggenes`.

38. Bazin A, Gautreau G, Médigue C, Vallenet D, Calteau A (2020) panRGP: A pangenome-based method to predict genomic islands and explore their diversity. *Bioinformatics* 36(Supplement_2):i651–i658.

39. Brooks ME, et al. (2017) glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal* 9(2):378–400.

40. Hartig F (2022) *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models* Available at: `https://CRAN.R-project.org/package=DHARMa`.

41. Paradis E, Schliep K (2019) Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:526–528.

42. Oksanen J, et al. (2020) Vegan: Community ecology package. Available at: `https://CRAN.R-project.org/package=vegan`.

43. Bittinger K (2020) *Usedist: Distance matrix utilities* Available at: `https://CRAN.R-project.org/package=usedist`.

44. Gu Z, Eils R, Schlesner M (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*.

45. Rocha EPC, Bikard D (2022) Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLOS Biology* 20(1):e3001514.

46. Millman A, et al. (2022) An expanded arsenal of immune systems that protect bacteria from phages. *Cell Host & Microbe* 30(11):1556–1569.e5.

47. Issues · labgem/PPanGGOLiN *GitHub*. Available at: `https://github.com/labgem/PPanGGOLiN/issues/126` [Accessed September 18, 2023].

48. Vassallo CN, Doering CR, Littlehale ML, Teodoro GIC, Laub MT (2022) A functional selection reveals previously undetected anti-phage defence systems in the E. Coli pangenome. *Nat Microbiol* 7(10):1568–1579.

49. Burke KA, Urick CD, Mzhavia N, Nikolich MP, Filippov AA (2024) Correlation of Pseudomonas aeruginosa Phage Resistance with the Numbers and Types of Antiphage Systems. *International Journal of Molecular Sciences* 25(3):1424.

50. Davidov E, Kaufmann G (2008) RloC: A wobble nucleotide-excising and zinc-responsive bacterial tRNase. *Molecular Microbiology* 69(6):1560–1574.

51. Bitton L, Klaiman D, Kaufmann G (2015) Phage T4-induced DNA breaks activate a tRNA repair-defying anticodon nuclease. *Molecular Microbiology* 97(5):898–910.

52. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS (2020) A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation. *Cell Host & Microbe* 27(1):140–153.e9.

53. Bazin A, Medigue C, Vallenet D, Calteau A (2021) panModule: Detecting conserved modules in the variable regions of a pangenome graph. doi:10.1101/2021.12.06.471380.

54. Tesson F, Bernheim A (2023) Synergy and regulation of antiphage systems: Toward the existence of a bacterial immune system? *Current Opinion in Microbiology* 71:102238.

55. Camus L, Vandenesch F, Moreau K (2021) From genotype to phenotype: Adaptations of Pseudomonas aeruginosa to the cystic fibrosis environment. *Microbial Genomics* 7(3):000513.

56. Crone S, et al. (2020) The environmental occurrence of Pseudomonas aeruginosa. *APMIS* 128(3):220–231.

57. Balfour-Lynn IM (2021) Environmental risks of Pseudomonas aeruginosa–What to advise patients and parents. *Journal of Cystic Fibrosis* 20(1):17–24.

Chapter 4: Robust statistical methods suggest that a minority of empirical phage-bacteria infection networks are nested

Abstract

The interactions between bacteria and phages are often studied at a network level, generating Phage-Bacteria Infection Networks (PBINs). These PBINs could reveal how the bacteria and phages are interacting and co-evolving, in part through the nestedness of the network. In a nested network, only generalist phages infect bacteria with broad resistance profiles. Additionally, specialist phages only infect bacteria with narrow resistance profiles. This is thought to arise from arms-race dynamics (ARD) co-evolution, in which bacteria and phages acquire increasing levels of resistance and infectivity, respectively. While nestedness analyses have often been performed with the Temperature metric (T), alongside the Equiprobable-Equiprobable null model (EE), this combination does not allow for an informative or reliable test. The effect of applying more robust combinations of nestednesness metrics and null models is yet to be studied in the context of PBINs. In this work, the performance of the T and NODF metrics is characterised, alongside the EE and qFF null models, on simulated networks. The NODF and qFF combination is shown to more accurately classify simulated networks as nested or non-nested than the T and EE combination. Additionally, the NODF and qFF combination is used to suggest that the majority of empirical PBINs from a previous meta-analysis are not significantly nested. Overall, this work calls for the adoption of NODF and qFF in the analysis of PBINs. This work also suggests that empirical PBINs should no longer be assumed to generally contain a nested pattern. These findings are important for the understanding of phage-bacteria co-evolution: this work suggests that NODF and qFF are preferable for inference of ARD in PBINs, and that many PBINs that were previously inferred to have ARD dynamics instead have generalists and specialists without a strict overlap in interactions.

1 Introduction

Ever since the initial reports of co-evolution between bacteria and phages, their interactions have been studied at the level of single bacterium-phage pairs. This has led to important discoveries, such as the phenomenon of directional selection for increased resistance and infectivity (1). Increasingly, individual bacterial and phage strains are also considered as members of a larger network of bacteria and phages, known as phage-bacteria infection networks (PBINs) (2). Studying interactions at the network level allows for processes such as local adaptation (3, 4), clade-specificity (5) and fluctuating selection (6) to be discovered in natural and clinical environments. In experimental systems, networks have allowed the dependence of co-evolutionary dynamics on resource availablity to be uncovered (7), as well as suggesting that phage-bacteria interactions over time are not limited to one of arms-race or fluctuating selection dynamics (8).

Given the complex patterns that can emerge in PBINs, their study relies on network metrics to distinguish pattern from noise (2). Nestedness quantifies the degree to which a network contains a spectrum of generalists to specialists, with the interactions of specialists forming subsets of the interactions of generalists (2, 9) (Figure 1A). The metric has been used to describe interactions between plants and pollinators or seed dispersers, predators and prey, species and habitats and host and parasites (10). Within the context of phage-bacteria interactions, nestedness is thought to be the consequence of arms-race dynamics (ARD) co-evolution (2). In this mode, bacteria continuously evolve novel ways to prevent phage infection, resulting in a gradient from generalists (infected by most phages) to specialists (infected by few phages). Phages continuously evolve ways to overcome the new defences, resulting in a gradient of specialists (infecting few hosts) to generalists (infecting most hosts). There is interest in uncovering the conditions that promote ARD, as this mode could be used to cultivate generalist phages for use as therapeutics, in what is known as "phage therapy" (11). The overlapping resistance and infectivity ranges, arising from the escalation in resistance and infectivity, is what distinguishes an ARD network from a network that simply contains a range of generalists and specialists. This latter network is known as having "heterogeneous marginal totals" (12) (Figure 1B). Nestedness has been reported in oceanic phage-bacteria communities (3), in soil phage-*Rhizobium* networks (4) and in *Vibrio* and their temperate phages isolated from fish and oysters (5). Additionally, a meta-analysis of 38 PBINs of various sizes and sources suggested 27 of these networks to be nested (13).

A commonality between all these applications is the need for a nestedness metric and a null model with which to assess statistical significance. While many nestedness metrics exist (14), the nestedness Temperature (T) and the Nestedness Metric based on Overlap and Decreasing Fill (NODF) have commonly been used with PBINs (2–4, 7, 13, 15–20). To determine statistical significance, the null model first generates randomised versions of the original network under certain constraints (9, 21). Then the nestedness of the original network is standardised using the null nestedness values, returning a standardised effect size (SES), with which statistical significance can be determined.

Choosing the right constraints is imperative for meaningful biological inference (22). For example, the commonly used Equiprobable-Equiprobable (EE) null model reshuffles interactions in the network without preserving host and infectivity range sizes (9, 21). Recall that nestedness consists of a generalist to specialist pattern, as well as overlapping resistance and infectivity ranges. Using EE, the hypothesis test asks: "is the PBIN significantly more nested than similar networks without any generalist to specialist ranges in resistance and infectivity?". Instead, consider the Fixed-Fixed (FF) null model (9, 21). This model retains host and phage range sizes in the null matrices. Therefore, only the degree to which ranges overlap vary in the null matrices. The hypothesis test therefore asks: "does the PBIN have significantly more overlap in resistance and infectivity range sizes than null matrices with the same range of generalists and specialists?".

Since overlapping fill is a key expected consequence of ARD, the FF null model allows for more informative biological inference. Nevertheless, many previous PBIN nestedness analyses have relied on the EE null model (3, 4, 15–19), including the widely cited notion that most PBINs are nested (13). As a consequence, it is unknown whether their significant results arose from a truely nested pattern, or only from having ranges of generalists and specialists. Given the breadth of resistance and counter-resistance mechanisms that bacteria and phages can employ (23), respectively, a non-nested generalist to specialist pattern would generally be an unsurprising result.

Despite allowing for a more informative hypothesis test, the FF null model is not perfect. Namely, a truely nested network is likely to be classified as non-nested using this null model (i.e., it has a high Type II error rate) (21). In an attempt to move beyond the use of single null models, Strona *et al.* developed the "Tuning Peg" (TP) algorithm (12). This method generates a variety of null models, which differ in the degree to which they preserve the distribution of host and resistance range sizes of the original network. The null matrices range from the aforementioned EE to FF in terms of restrictiveness. Interestingly, null matrices with near-fixed row and column totals, coined quasi-FF (qFF) matrices, were found to distinguish nested networks and networks with heterogeneous marginal totals (12).

Nevertheless, the applicability of qFF to PBIN nestedness analysis is still to be determined. Firstly, Strona *et al.* did not quantify error rates for this null model. Additionally, the method was tested on square networks (i.e. the same number of phages and hosts) with column and row numbers varying between 15 and 100. Empirical PBINs are often smaller, and rarely square (13). Therefore, while qFF promises to be more robust than EE, its applicability to PBIN nestedness analysis remains to be verified. If qFF was found to distinguish nested and non-nested PBINs well, it would allow researchers to ask: "does the PBIN have significantly more overlap in resistance and infectivity range sizes than null matrices with a very similar range of generalists and specialists?". This, combined with the potential of good statistical properties, would make qFF a strong contender for the main null model to use in PBIN nestedness analysis.

While the objective is often to determine whether a PBIN is significantly more nested than null matrices, it is also possible for PBINs to be significantly less nested than null matrices. This is known as anti-nestedness (24). There are various patterns that can underly anti-nestedness. One such pattern is "non-inclusive sets", in which bacteria resist phages with infectivity ranges greater than the phage which infects them (24, 25). Another such pattern is the "perfect checkerboard", which can be rearranged to show two compartments (24, 25). In analyses of PBINs, compartmentalisation can be of interest, as it can be a marker of local adaptation or fluctuating selection dynamics (FSD) (2, 3) (Figure 1C). In local adaptation, phages can evolve to be more infective on local hosts than on hosts from other populations (26). In FSD, phage selection drives cycling in the abundance of host genotypes, which drives a lagged cycle of abundance in phage genotypes (27). Crucially, FSD is non-directional, meaning that resistance and infectivity ranges are not expected to grow over time. While Almeida-Neto *et al.* described anti-nestedness for specific types of compartmentalisation (24), it is currently unknown whether anti-nestedness is a general property of compartmentalised networks. Additionally, the behaviour of qFF on compartmentalised matrices has not been characterised. Since compartmentalisation is expected to occur under local adaptation or FSD, it is important to understand whether anti-nestedness can be used as one of the markers for these dynamics.

In this work, the performance of T and NODF are compared using the EE and qFF null models. Random, nested, heterogeneous and compartmentalised networks are simulated, with size and fill appropriate to empirical PBINs. To ask which of the four metric-null model combinations can best characterise PBINs, these combinations are applied to simulated networks. Then, the PBINs from Flores *et al.* (13) are re-analysed using these four method combinations. Overall, this work shows that the NODF and qFF combination is the most reliable. Additionally, this work suggests that the majority of empirical PBINs in the Flores collection cannot be considered nested, when using robust methodology. Finally, this work highlights the need to reconsider the conditions under which ARD are expected, and whether nestedness would be a good indicator of this dynamic.

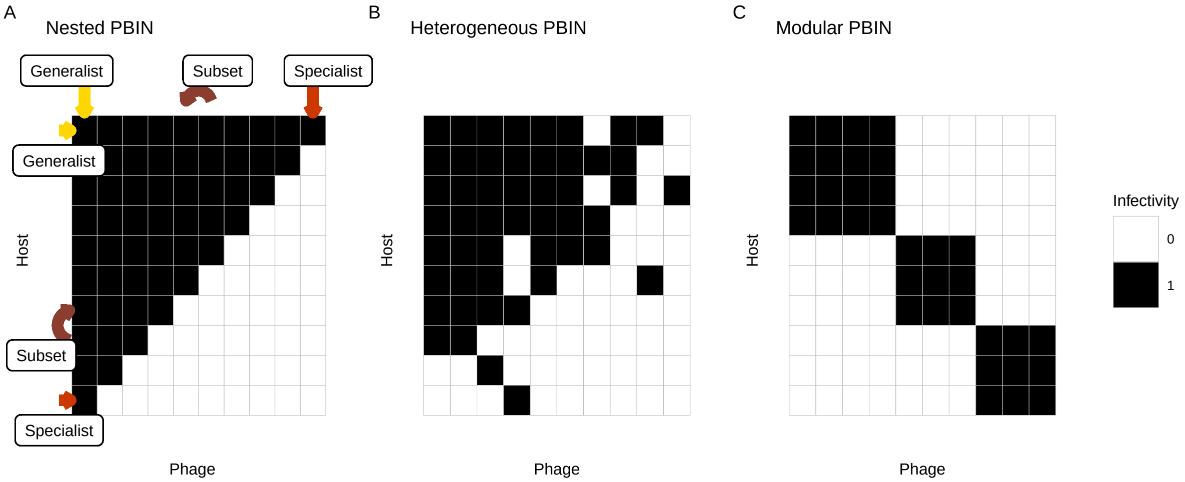

Figure 1: Patterns of interest in PBINs. A) In the nested pattern, hosts and phages exist along a continuum of generalists to specialists. In addition, the interactions of specialists form subsets of the interactions of generalists. B) In a PBIN with heterogeneous marginal totals, hosts and phages exist along a continuum of generalists to specialists without the interactions of specialists forming nested subsets of the interactions of generalists. C) In a modular PBIN, phage-bacteria interactions are clustered.

2 Methods

2.1 Overview

Broadly speaking, this work consisted of five stages (Figure 2). A brief overview is given here, before further details are outlined below.

In the first stage, 100 PBINs of four types were simulated: uniform (completely random, given network size and fill), heterogeneous (with generalist to specialist ranges in hosts and phages, but no nested subsets), nested and compartmentalised (Figure 2A). In addition, the PBINs analysed by Flores *et al.* were obtained from the publication's supplementary files (13).

In the second stage, null networks were obtained for each of the 437 PBINs (Figure 2B). Since the TP algorithm returned 5 qFF null matrices per run, the algorithm was run 50 times per PBIN to generate 250 qFF null matrices (Figure 2B1). Additionally, the EE function was run 250 times for each PBIN (Figure 2B2).

In the third stage, standardised effect sizes (SESs) were calculated for each PBIN (Figure 2C). First, raw nestedness was calculated with NODF and NT (Figure 2C1). Then, four SES values were calculated, following all possible combinations of nestedness metric and null model (Figure 2C2). An SES adjusted the raw nestedness value (e.g. NODF) by the nestedness values of all null networks (e.g. 250 qFF networks), thereby measuring how different the nestedness of the PBIN was from the nestedness of the null networks. An SES above 1.64 was taken as significant nestedness and below -1.64 as sigificant anti-nestedness, given the 95th and 5th percentiles of the standard normal distribution, respectively.

In the fourth stage, the proportion of nested, non-nested and anti-nested PBINs were visualised for each of the five PBIN types, in addition to estimating an average effect (Figure 2D).

Finally, two quality control analyses were performed (Figure 2E). First, the association between SES, network size and network fill was measured across the different types of simulations (Figure 2E1). Second, the proportion of unique qFF null networks was calculated for each PBIN type (Figure 2E2).

All PBINs used in this work, alongside their nestedness metrics and null nestedness distributions, can be viewed online (28). PBINs were simulated, analysed and visualised in R and the workflow was managed using Snakemake (29, 30).

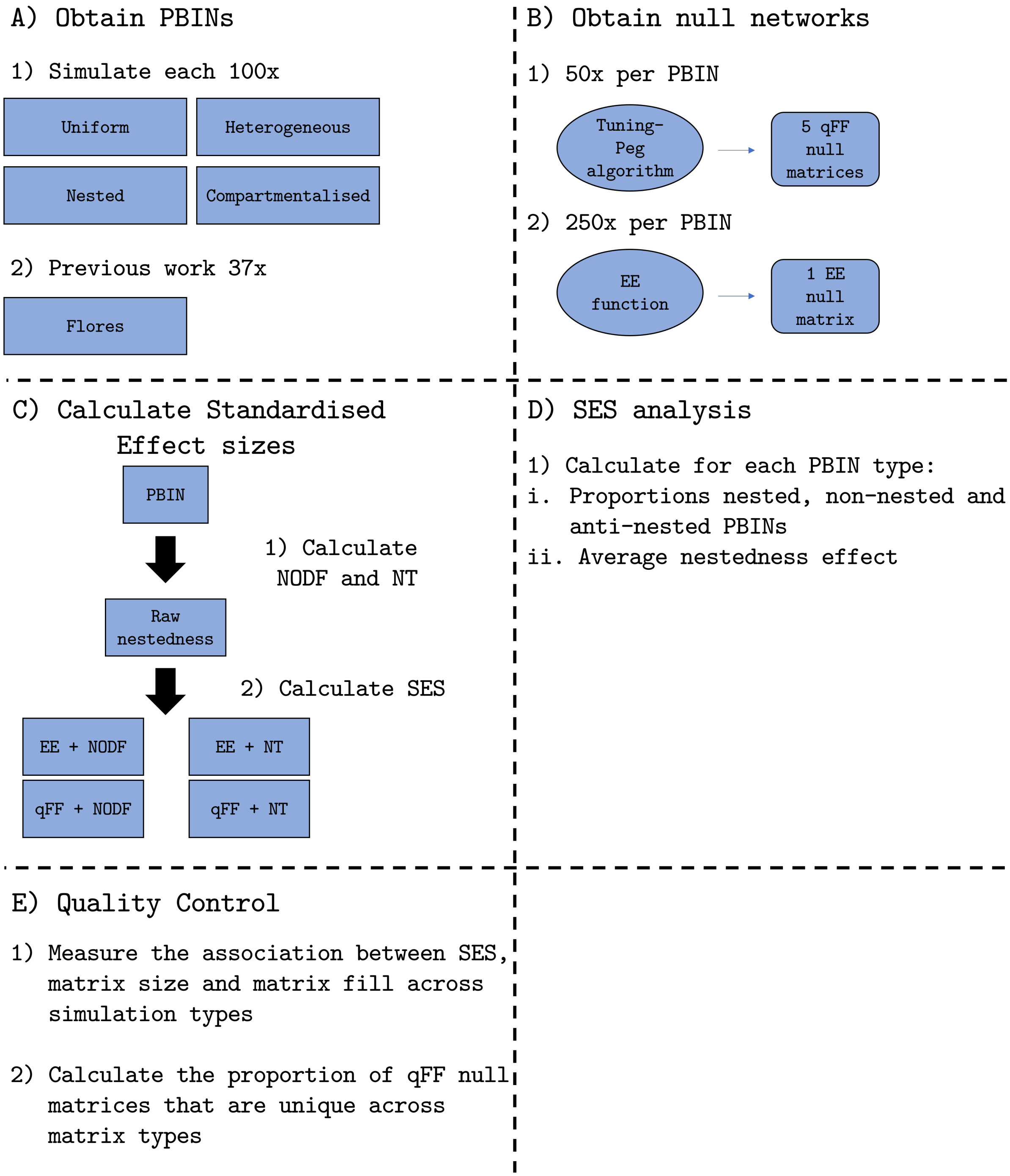

Figure 2: An overview of the methods employed. A) PBINs were either simulated or obtained from a previous meta-analysis. B) Then, qFF and EE null networks were simulated for each PBIN. C) These null networks were used to calculate standardised effect sizes (SESs) for each combination of nestedness metric and null model. D) Next, these SES values were analysed to ask whether PBINs were on average nested, non-nested or anti-nested across each of the four SES combinations. E) Finally, two quality control analyses were performed. The association between SES, matrix fill and matrix size was estimated (E1). In addition, the proportion of unique qFF null matrices was visualised as a function of matrix size (E2).

2.2 Simulation of theoretical infectivity matrices

Theoretical infectivity matrices were generated as follows. The number of rows and columns of a matrix were drawn from a uniform distribution with bounds of 5 and 25. The fill of the matrix was drawn from a uniform distribution with bounds of 0.2 and 0.8 (as in (12)). Given these matrix properties, a set of three matrices was produced: one with uniform marginal totals ("U", i.e. a completely random matrix), one with heterogeneous marginal totals ("H", i.e. a matrix with diverging row and column totals but without imposed overlapping fill) and one with nestedness ("N"). U and H matrices were generated equal (U) and unequal (H) probability sampling with R's `sample` function (31). In contrast, N matrices were generated through a nestedness maximisation procedure.

Specifically, U matrices were generated by sampling positive matrix cells (given the matrix size and fill) with equiprobable cell weights. That is, if the matrix was drawn to have 20 cells and a fill of 0.5, 10 cells would be randomly sampled to be positive. In contrast, H matrices were generated by sampling positive matrix cells (given the matrix size and fill) with variable cell weights. Row weights and column weights were obtained by sampling from an exponential distribution, with a rate drawn from a uniform ditribution with values between 1 and 2 (as in (12)). Cell weights were then obtained by multiplying row and column weights. Finally, positive matrix cells were sampled using these weights. Drawing row and column weights from exponential distributions ensured that the set of H matrices was likely to contain entries with high, intermediate and low row and column fill. R's `sample` function employed inverse transform sampling (31). Therefore positive cells were drawn sequentially, with the probability of a cell being positive being proportional to its weight, given the weights of all other remaining cells.

N matrices were generated by a sequential swap procedure. Starting with an H matrix, positive and zero cells were swapped, with only swaps that increased the NODF metric being retained. Swapping was repeated until 2000 sequential swaps did not increase NODF. 100 sets of matrix properties (i.e. size and fill) and associated theoretical matrices were simulated. Since the downstream TP algorithm involved division by row and column totals to calculate row and column discrepancies, the theoretical matrices were constrained to having non-empty rows and columns (12). Cell weights were sampled using the R functions `sample` and `rexp` and NODF was calculated using the `vegan` package in R (31, 32).

In addition, 100 compartmentalised matrices were simulated ("C"). First, 10^5 compartmentalised matrix specifications were simulated with fill between 0.2 and 0.8. Out of these, 100 specifications were sampled, ensuring a uniform distribution of matrix fill. To simulate the matrix specifications, the number of rows and columns of a matrix were drawn from a uniform distribution with bounds of 5 and 25. Additionally, the number of modules was drawn from a uniform distribution with bounds of 2 and the smallest dimension of the matrix. Thus, a matrix with 6 rows and 10 columns could have 2 to 6 modules. Each module had a minimum of one row and one column, with additional rows and columns separately drawn from a multinomial distribution, with weights separately drawn from an exponential distribution with a rate parameter of 4. The heterogeneous weights ensured that some matrices contained large modules, which were required to include high-fill compartmentalised matrices in the drawn matrix specifications. After sampling 100 matrix specifications with a uniform fill distribution, these matrices were generated by following the specification of modules (i.e. module cells were filled and non-module cells were left empty). Sampling was performed using R's `runif` and `rexp` functions (31).

2.3 Experimental data processing

Experimental data was obtained from the supplementary materials of Flores *et al.* (13). Flores *et al.* provided one Excel file with 38 sheets, each containing one PBIN. These sheets were split into separate csv files using the `pandas` library (33). These csv files were processed using `tidyverse` packages (34), including the removal of empty rows and columns. This is required for the TP algorithm, which involves division by row and column totals to calculate row and column discrepancies (12). Importantly, division by 0 causes the algorithm to fail. Since the network from Kudva *et al.* had a fill of 1 after removing empty rows, it was excluded from the analysis (35). The complete fill of this network prevented it from being randomised, and therefore the significance of its nestedness could not be determined.

2.4 Construction of the null matrices

For each PBIN, the TP algorithm was run 50 times, resulting in 250 qFF null matrices. Note that the five qFF matrices from a single TP run are structurally independent (12), preventing the need to run the TP algorithm 250 times. The TP algorithm was executed using the R script provided by Strona *et al.* (12), with a grid increment of 10. This work was parallelised using the `foreach`, `parallel` and `doParallel` packages (31, 36, 37). The qFF matrices, defined here as cells (0, 0.2), (0.1, 0.2), (0.2, 0.2), (0.2, 0.1) and (0.2, 0), were extracted from the TP objects using the `igraph` package (38). 250 EE null matrices were generated using the `EE` function provided by Strona *et al.* (12).

2.5 Calculation of nestedness effect sizes

To measure nestedness, two metrics were used in this work: nestedness temperature (T) and nestedness metric based on overlap and decreasing fill (NODF) (39, 40). Since T decreases with increased nestedness, $NT = 100 - T$ was used in this work (as in (13, 40)). Both metrics were computed using the `vegan` package (32). For each combination of PBIN and null matrix type (EE or qFF) a standarised effect size was calculated as $SES = \frac{x-\mu}{\sigma}$, where x is the NT or NODF value of a PBIN, μ is the mean NT or NODF value of the 250 null matrices and σ is the standard deviation of the NT or NODF distribution in the null matrices (12). To determine whether a PBIN was nested, non-nested or anti-nested, its SES was compared to the bounds of the 90% confidence interval of a standard normal distribution (analogous to a one-sided hypothesis test). An SES greater than 1.64 indicated nestedness, an SES less than -1.64 indicated anti-nestedness and an SES between these bounds indicated non-nestedness. To obtain p-values for significant average nestedness and anti-nestedness, one-sided t-tests on the standardised effect size were performed with μ = 1.64 and μ = -1.64, respectively. To obtain p-values for significant non-nestedness, equivalence tests with bounds (-1.64, 1.64) were performed using the `TOSTER` package (41).

2.6 Association between SES, size and fill across simulation types

To assess whether SES was associated with size and fill across simulation types, this data was both visualised with simple linear regression models and analysed through linear mixed modelling. Simple linear regression plots were generated using `ggplot2` (42). Two linear mixed models were fit using the `lme4` package (43). The first model was fit to the SES data of non-compartmentalised matrices, with SES as the outcome variable and predictor variables for scaled size, scaled fill, null model, nestedness metric, simulation type and their interactions. Since each set of size and fill generated one uniform, one heterogeneous and one nested matrix, and each matrix was analysed with two null models and two nestedness metrics, the model included a random intercept for each set of size and fill (with 12 data points per group). The second model was fit to the SES data of compartmentalised matrices with the same set of predictor variables, excluding simulation type. This model also included a random intercept for each set of size and fill (with 4 data points per group). Parameter estimates and p-values were obtained with the `lmerTest` package (44) by re-fitting the same model for each combination of nestedness metric and null model (and simulation type, in the case of non-compartmentalised matrices), with the reference levels set to the nestedness metric and null model (and simulation type) of interest.

2.7 Association between the proportion of qFF null matrices that were unique and matrix size

To assess how the proportion of qFF null matrices that were unique varied along matrix size, this data was visualised with scatterplots generated with `ggplot2` (42) for each matrix source. Uniqueness was determined using R's `unique` function (31). Uniqueness was visualised prior to reordering of null matrices, as well as after reordering by the NODF or T functions (`nestednodf` and `nestedtemp` functions from the `vegan` package (32), respectively). The NODF algorithm reorders matrices by arranging their columns and rows by decreasing marginal totals (40). In contrast, the T algorithm reorders matrices through an algorithm that aims to minimise T (i.e. maximise NT) (32, 45).

2.8 Visualisation of PBINs and null nestedness distributions

Each PBIN was visualised using the `ggplot2` package (42). In addition, distributions of nestedness values in the null matrices were visualised for each PBIN across all combinations of null models and nestedness metrics. These distributions were visualised using the `ggplot2` package (42), with the nestedness of the original PBIN and its SES marked on the plots. These visualisations are available online (28).

3 Results

3.1 Across all null model and nestedness metric combinations, uniform matrices were significantly non-nested

To assess the ability of qFF and EE to characterise random matrices as non-nested, nestedness tests were performed on uniform matrices (U). In U matrices interactions were allocated to matrix cells with equal probability. As a consequence, every "phage" and "host" combination had an equal probability of being positive. Since the simulation did not involve creating marginal totals or overlapping fill, U matrices were not expected to show nestedness.

Firstly, the proportion of nested, anti-nested and non-nested matrices was calculated according to each combination of null model and nestedness metric. Across all combinations, the majority of matrices were non-nested, with 7% or less of matrices being classified as significantly nested (Figure 3). In addition, the mean SES values were statistically equivalent to 0 ("p-value non-nested" in Table 1). Overall, these results show that both qFF and EE were able to characterise U matrices as non-nested, irrespective of nestedness metric. Nevertheless, the qFF combinations classified a greater proportion of networks as non-nested compared to EE combinations.

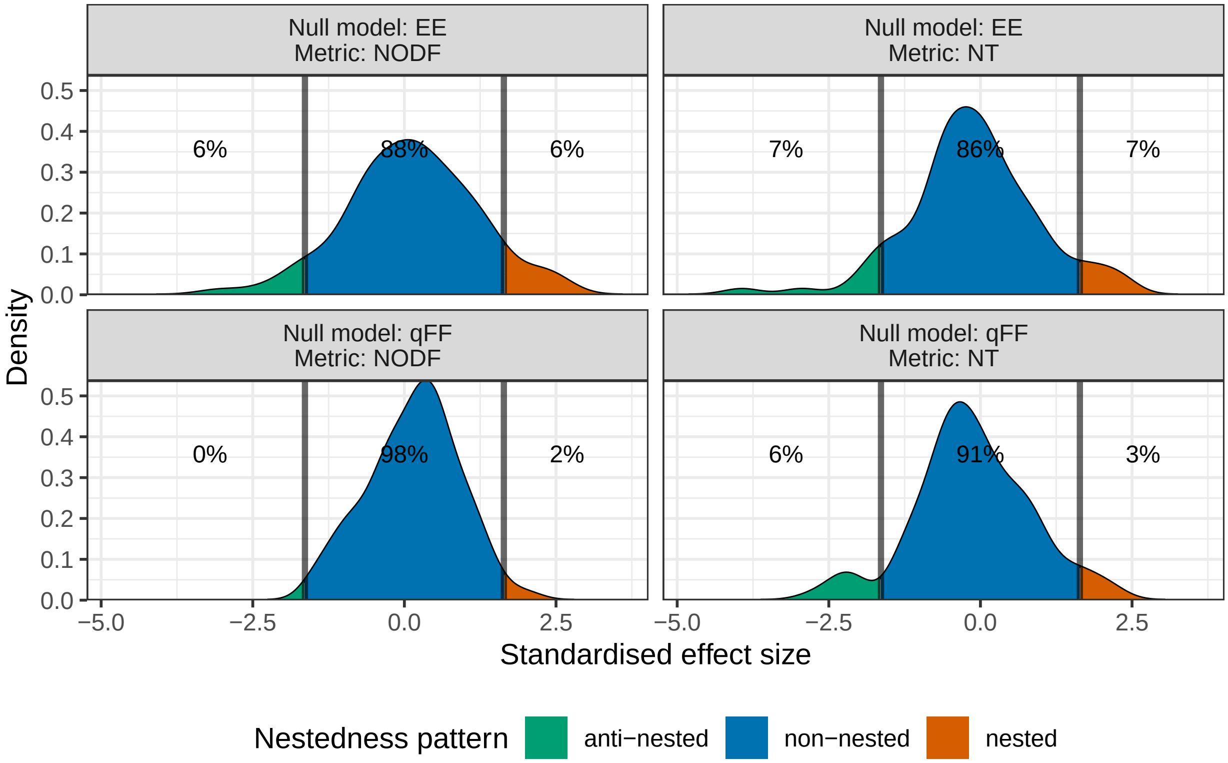

Figure 3: The distributions of standardised effect sizes across nestedness metrics and null models for the simulated uniform matrices. Percentages indicate the percentage of simulated matrices that were classified as anti-nested, non-nested or nested based on their SES. Across all combinations, the majority of matrices were correctly classified as non-nested. Nevertheless, qFF did classify more PBINs as non-nested than EE, across both nestedness metrics.

Table 1: Across both metrics and null models, the mean standardised effect size was statistically equivalent to 0 (p-value non-nested < 0.05).

metric	null model	mean ses	CI low	CI high	t statistic	df	p-value nestedness	p-value non nested	p-value anti nestedness
NT	EE	-0.111	-0.32	0.0987	-16.6	99	1	1.26e-26	1
NT	qFF	-0.131	-0.32	0.0578	-18.7	99	1	2.43e-29	1
NODF	EE	0.128	-0.0832	0.3390	-14.2	99	1	5.15e-26	1
NODF	qFF	0.146	-0.000776	0.2940	-20.2	99	1	3.54e-37	1

3.2 The NODF and qFF combination reliably identified nested matrices as nested

To assess the ability of qFF and EE to characterise nested matrices as nested, nestedness tests were performed on nested matrices (N). These were generated by a sequential swapping procedure which maximised the NODF of a given random matrix. As a consequence, N matrices were expected to be more nested than most alternative “infectivity” patterns given the number of “phages”, “hosts” and “interactions”.

Firstly, the proportion of nested, anti-nested and non-nested matrices was calculated according to each combination of null model and nestedness metric. While most combinations classified at least 95% of matrices as significantly nested, the NT and qFF combination only considered about 65% significantly nested (Figure 4). Nevertheless, all combinations showed significant average nestedness (“p-value nestedness” < 0.05 in Table 2). Overall, all methods could reliably identify nested matrices as nested. However, the reduced proportion of networks that were significantly nested under qFF and NT reflected the differing definitions underlying NT and NODF: the nested matrices, which were generated by maximising NODF, did not always carry a significant NT score with qFF.

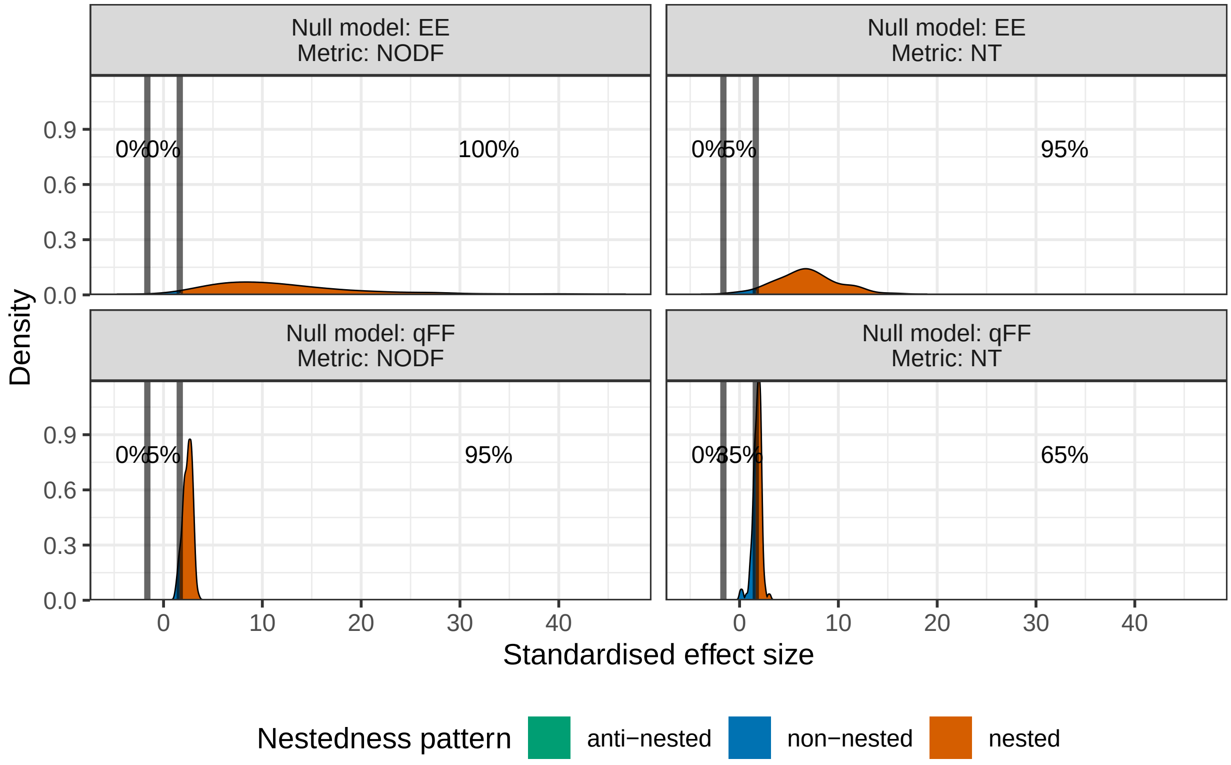

Figure 4: The distributions of standardised effect sizes across nestedness metrics and null models for the simulated nested matrices. Percentages indicate the percentage of simulated matrices that were classified as anti-nested, non-nested or nested based on their SES. Across all combinations, the majority of matrices were correctly classified as nested. However, the combination of NT and qFF was somewhat prone to classifying nested networks that were generated by maximising NODF as non-nested.

Table 2: Across all metric and null model combinations, nested simulated matrices were on average significantly nested (p-value nestedness < 0.05).

metric	null model	mean ses	CI low	CI high	t statistic	df	p-value nestedness	p-value non nested	p-value anti nestedness
NT	EE	6.83	6.22	7.45	16.70	99	8.71e-31	1.000	1
NT	qFF	1.77	1.69	1.85	3.14	99	0.00113	0.999	1
NODF	EE	11.90	10.50	13.40	14.50	99	1.76e-26	1.000	1
NODF	qFF	2.43	2.34	2.51	17.70	99	1.19e-32	1.000	1

3.3 Only with EE was there an erroneous significant nestedness effect in matrices with heterogeneous marginal totals

To assess the ability of qFF and EE to characterise matrices with varying range sizes but without overlapping ranges as non-nested, nestedness tests were performed on matrices with heterogeneous marginal totals (H). In H matrices, the probability of a “phage” and “host” interacting depended on their respective infectivity and host range sizes, simulated with weights drawn from an exponential distribution. Therefore, while the host and infectivity range sizes were simulated to vary as in a nested network, any overlap in fill would arise by chance.

Firstly, the proportion of nested, anti-nested and non-nested matrices was calculated according to each combination of null model and nestedness metric. Note that by chance, a subset of networks from this simulation will have had overlapping fill, and would therefore be nested. While approximately 78.5% of H matrices were classified as non-nested with qFF, the majority of H matrices were wrongly classified as nested with EE (91%, Figure 5). In other words, H matrices were only significantly nested compared to networks without a similar generalist to specialist range in resistance and infectivity. Similarly, while the average effect size with qFF was statistically equivalent to 0 (“p-value non-nested” < 0.05), the average effect size indicated significant nestedness with EE (“p-value nestedness” < 0.05, Table 3). Overall, these results show that only qFF could reliably identify H matrices as non-nested.

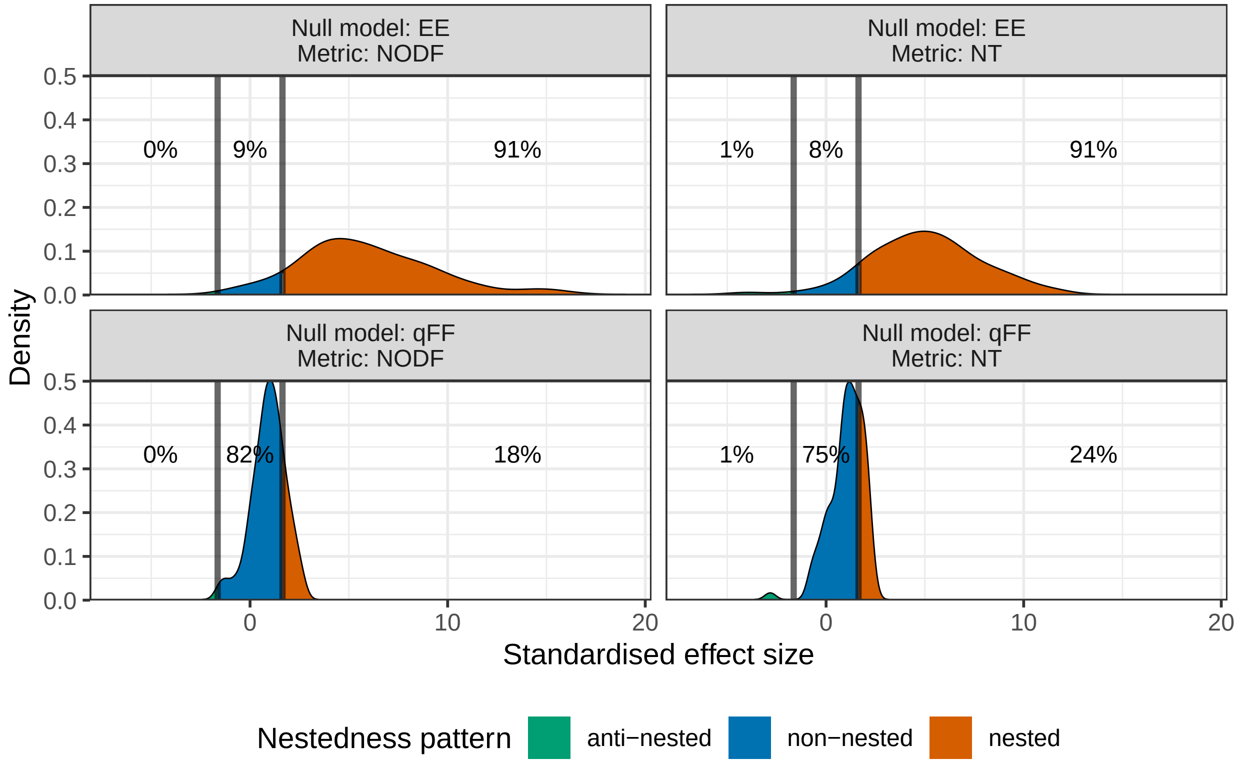

Figure 5: The distributions of standardised effect sizes across nestedness metrics and null models for the simulated matrices with heterogeneous marignal totals. Percentages indicate the percentage of simulated matrices that were classified as anti-nested, non-nested or nested based on their SES. Using qFF, the majority of networks were correctly classified as non-nested. In contrast, using EE, the majority of networks were wrongly classified as nested.

Table 3: The qFF null model correctly suggested that heterogeneous matrices were on average significantly non-nested. In contrast, the EE null model wrongly suggested that the heterogeneous matrices were on average significantly nested.

metric	null model	mean ses	CI low	CI high	t statistic	df	p-value nestedness	p-value non nested	p-value anti nestedness
NT	EE	5.050	4.490	5.60	12.2	99	7.63e-22	1	1
NT	qFF	1.030	0.860	1.20	-7.17	99	1	6.7e-11	1
NODF	EE	5.900	5.240	6.55	12.8	99	5.35e-23	1	1
NODF	qFF	0.935	0.766	1.10	-8.34	99	1	2.27e-13	1

3.4 Only the NODF metric generally considered compartmentalised matrices anti-nested

Since previous work suggests that some compartmentalised networks are anti-nested (24), and compartmentalisation has been observed in empirical networks (3–6, 20), the nestedness of compartmentalised networks, under the different null models and nestedness metrics, was investigated. Firstly, the proportion of nested, anti-nested and non-nested matrices was calculated according to each combination of null model and nestedness metric (Figure 6). Interestingly, only with NODF were the majority of matrices considered anti-nested. In contrast, with NT and qFF the matrices were split between non-nested and anti-nested, and with NT and EE, matrices were split between nested, non-nested and anti-nested. In terms of average effects, the NT and EE combination was the only combination that did not show a significant anti-nestedness effect (Table 4). Overall, these results suggests that compartmentalised networks only had a predictable anti-nested pattern when using the NODF metric, while the nestedness pattern was more variable with the NT metric. Importantly, this highlights that co-evolution between bacteria and phages may not always be expected to result in nestedness. In constrast, if NODF is used, local adaptation or FSD would be expected to result in anti-nestedness regardless of null model used.

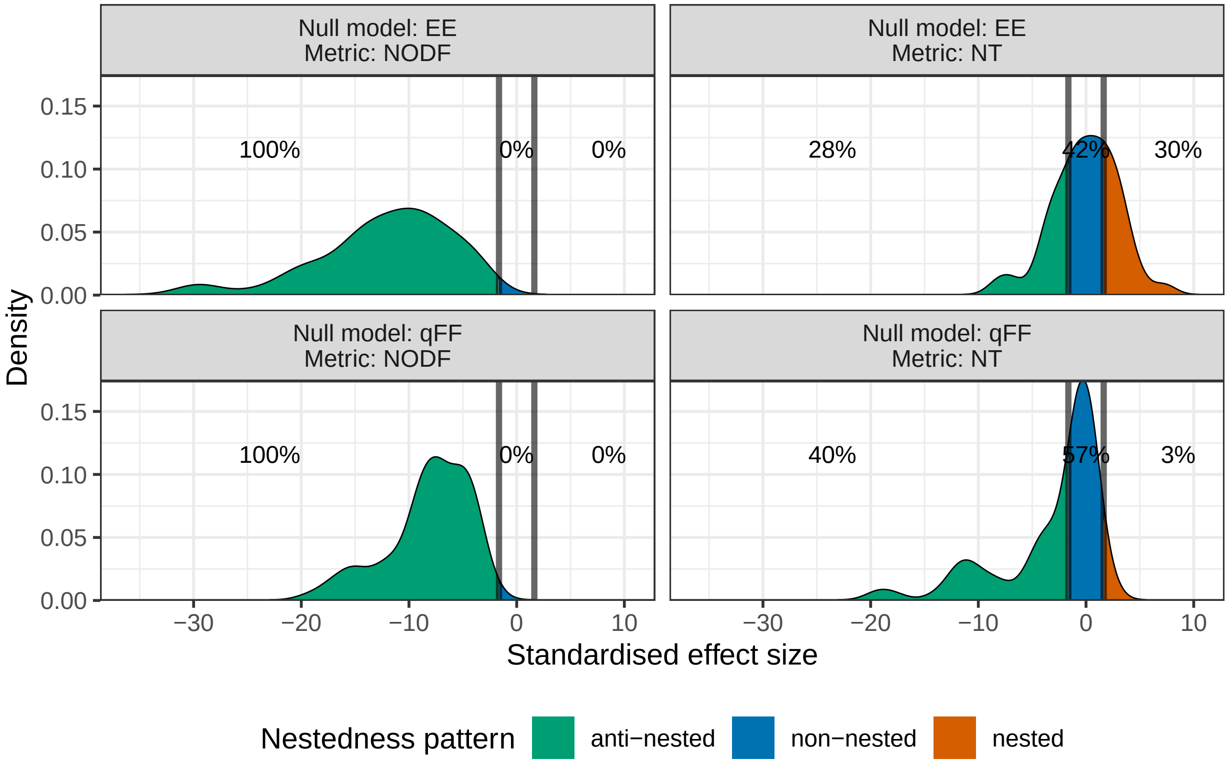

Figure 6: The distributions of standardised effect sizes across nestedness metrics and null models for the simulated compartmentalised matrices. Percentages indicate the percentage of simulated matrices that were classified as anti-nested, non-nested or nested based on their SES. Using NODF, the majority of networks were classified as anti-nested. In contrast, using NT, networks were split between the three categories.

Table 4: With NODF, networks were on average significantly anti-nested. With NT, networks were significantly non-nested.

metric	null model	mean ses	CI low	CI high	t statistic	df	p-value nestedness	p-value non nested	p-value anti nestedness
NT	EE	-0.074	-0.674	0.526	-5.68	99	1	5.48e-07	1
NT	qFF	-3.18	-4.14	-2.22	-9.99	99	1	0.999	0.000984
NODF	EE	-12.1	-13.3	-10.9	-22.5	99	1	1	1.27e-31
NODF	qFF	-8.16	-8.91	-7.4	-25.6	99	1	1	1.76e-31

3.5 Only with qFF was the SES largely invariant to network size and fill

To assess whether matrix size may have affected the SES values reported above, the association between matrix size and SES was investigated across all null model and nestedness metric combinations (Figure 7). The plot suggests that the SES of nested and heterogeneous matrices varied with matrix size with the EE null model. Additionally, the SES of compartmentalised matrices appeared to vary with matrix size across most, if not all, combinations of nestedness metric and null model.

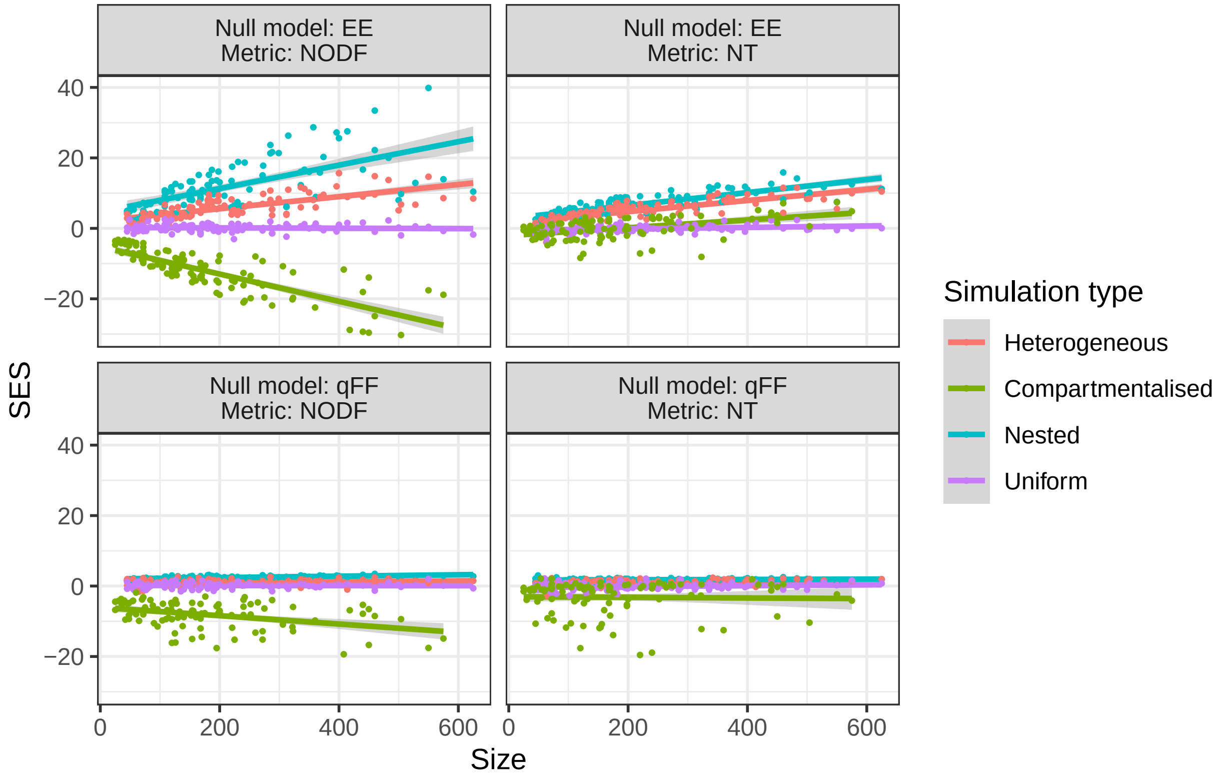

Figure 7: The association between SES and matrix size across simulation type, null model and nestedness metric combinations. SES appears to associate with matrix size for heterogeneous and nested matrices when using the EE null model. Additionally, SES appears to vary along size with compartmentalised matrices.

To assess whether matrix fill may have affected the SES values reported above, the association between matrix fill and SES was also investigated across all null model and nestedness metric combinations (Figure 8). Similar to size, the SES of nested and heterogeneous matrices appeared to vary with matrix size with the EE null model. Additionally, the SES of compartmentalised matrices appeared to vary with matrix fill across most, if not all, combinations of nestedness metric and null model.

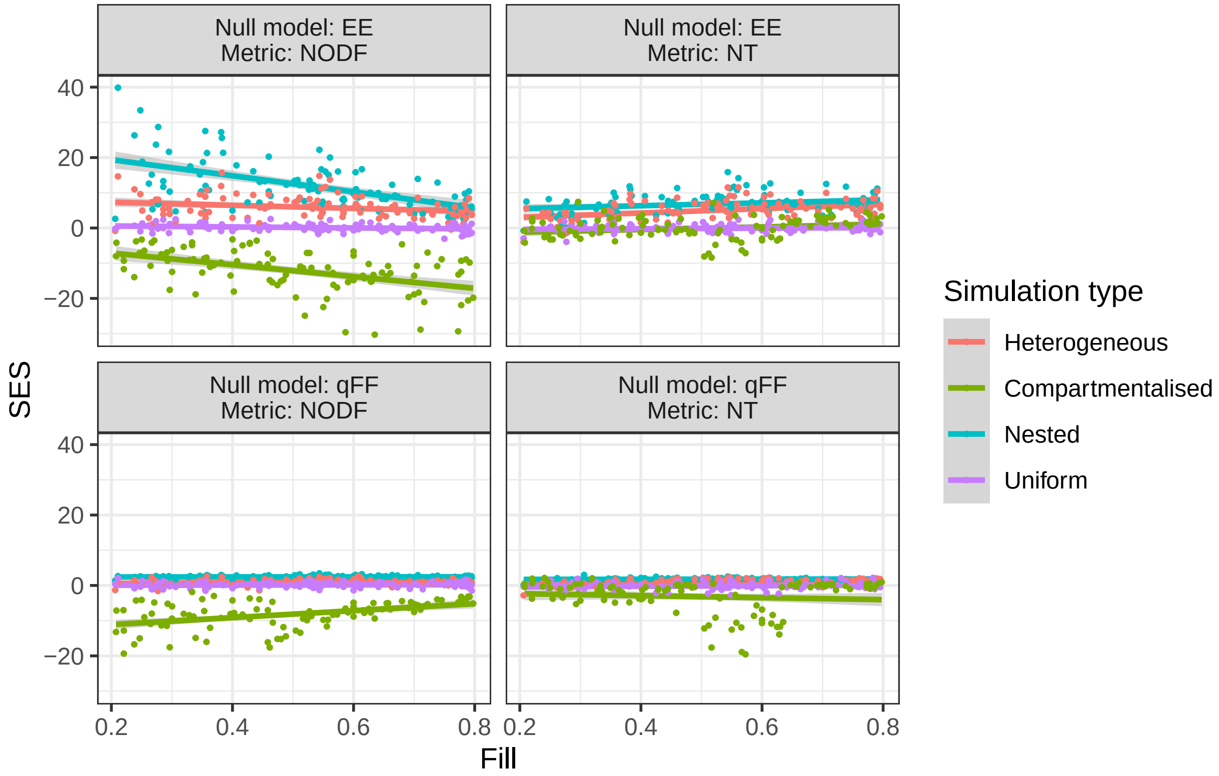

Figure 8: The association between SES and matrix fill across simulation type, null model and nestedness metric combinations. SES appears to associate with matrix fill for heterogeneous and nested matrices when using the EE null model. Additionally, SES appears to vary along fill with compartmentalised matrices.

To assess the strength of the associations with size and fill, parameter estimates were obtained for scaled size, scaled fill and their interaction (Table 5). The largest parameter estimates were observed with the EE null model. Additionally, on nested and heterogeneous matrices, qFF parameter estimates were generally of a smaller scale than EE parameter estimates. On compartmentalised matrices, the NODF + qFF combination showed strong associations with size and fill, despite still being of smaller scale than the NODF + EE combination. Overall, this analysis suggests that when performing nestedness tests on simulated matrices, use of the qFF null model was less susceptible to variation with matrix size and fill, however analysis of compartmentalised matrices may require accounting for matrix size and fill.

Table 5: Parameter estimates for scaled size, scaled fill and their interaction across simulation type, null model and nestedness metric combinations, when fitting linear mixed models to SES. The table suggests that SES was largely invariant to these factors for uniform matrices. For nested and heterogeneous matrices, SES appeared to associate with most factors only when using the EE null model. For compartmentalised matrices, SES appeared to associate with combinations of these factors across all nestedness metrics and null models.

metric	null model	size estimate	size p-value	fill estimate	fill p-value	interaction estimate	interaction p-value
uniform							
NODF	qFF	0.003	0.981	0.072	0.54	-0.151	0.14
NODF	EE	-0.037	0.754	-0.234	0.048	-0.040	0.694
NT	qFF	0.168	0.157	0.129	0.273	-0.075	0.462
NT	EE	0.292	0.014	0.140	0.236	-0.154	0.13
nested							
NODF	qFF	0.279	0.019	-0.016	0.892	-0.020	0.847
NODF	EE	5.414	< 0.001	-4.207	< 0.001	-2.483	< 0.001
NT	qFF	0.059	0.616	0.023	0.848	0.015	0.885
NT	EE	2.487	< 0.001	0.467	< 0.001	-0.254	0.013
heterogeneous							
NODF	qFF	0.187	0.115	0.187	0.114	-0.108	0.29
NODF	EE	2.564	< 0.001	-0.934	< 0.001	-0.749	< 0.001
NT	qFF	0.275	0.021	0.403	0.001	-0.053	0.606
NT	EE	2.082	< 0.001	0.877	< 0.001	-0.210	0.04
compartmentalised							
NODF	qFF	-1.283	< 0.001	1.647	< 0.001	0.768	0.016
NODF	EE	-5.261	< 0.001	-3.073	< 0.001	-1.523	< 0.001
NT	qFF	-0.169	0.606	-0.490	0.129	-0.191	0.547
NT	EE	1.359	< 0.001	0.718	0.026	-0.302	0.341

3.6 Only with the EE null model was there a non-zero nestedness effect in empirical PBINs

The sections above showed that the behaviour of EE and qFF varied across nested, heterogeneous and uniform matrices. In addition, compartmentalised matrices were more likely to be classified as anti-nested when using NODF than when using NT. Importantly, the combination of NT and EE was shown to erroneously classify heterogeneous PBINs as nested, while the NODF and qFF combination correctly classified many as non-nested. Since empirical PBINs were previously concluded to generally show nestedness using the NT and EE combination (13),
the data was re-analysed with all null model and metric combinations. Firstly, the proportion of nested, anti-nested and non-nested matrices was calculated according to each combination of null model and nestedness metric (Figure 9). While many matrices were considered significantly nested with EE, the majority of matrices were considered non-nested with qFF. In other words, the use of qFF greatly reduced the number of PBINs that were considered significantly nested (Figure 9). With NT, the number of significantly nested PBINs was reduced from 25 to 4. With NODF, the number was reduced from 16 to 2. In addition, the NODF and qFF combination considered 10 networks to be anti-nested. Since NODF with qFF did not suffer from NT and EE's high change of erroneously suggesting nestedness in non-nested networks with heterogeneous marginal totals, the true proportion of nested matrices in the Flores data set was estimated to be 0.05. In terms of average effects, only with NT were these significant: with EE there was significant average nestedness, while with qFF there was significant average non-nestedness (Table 6). Overall, this analysis suggested that the PBINs in the Flores data set were generally not significantly nested when compared to null matrices with similar resistance and infectivity range sizes. In other words, the Flores networks generally lacked a significant pattern of nested subsets, a key component of ARD dynamics.

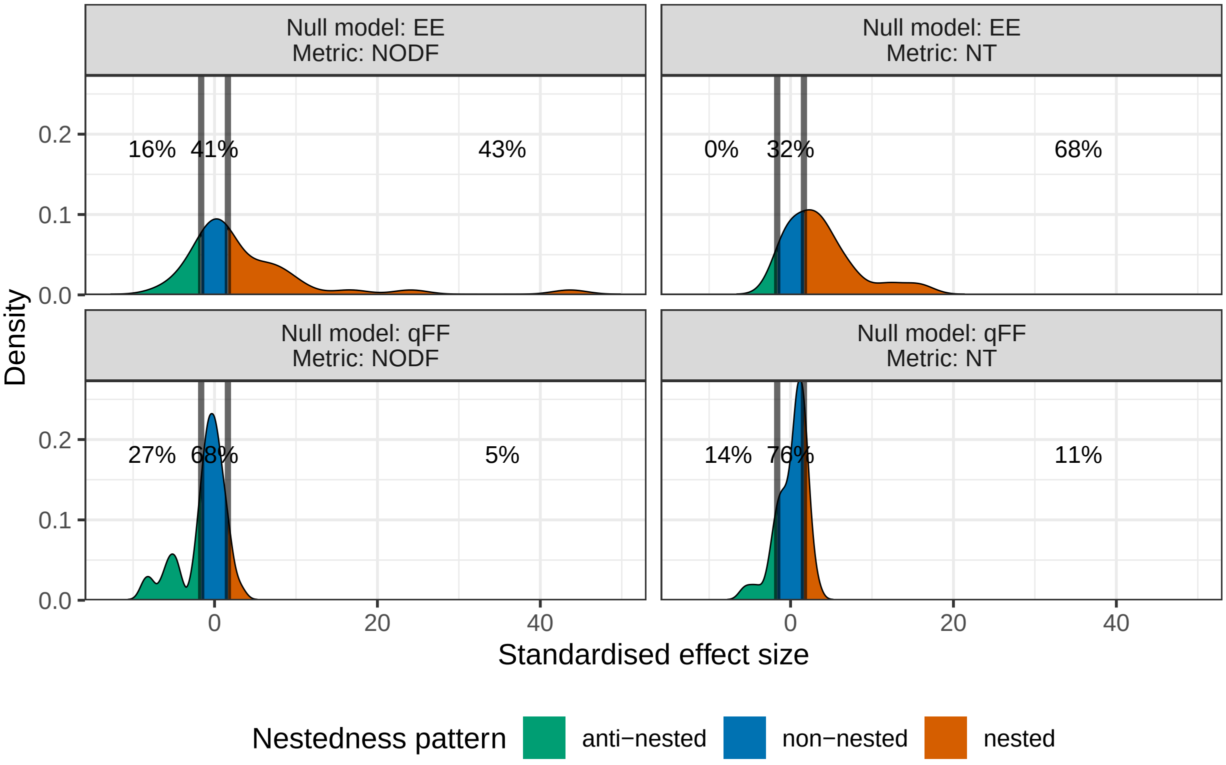

Figure 9: The distributions of standardised effect sizes across nestedness metrics and null models for the empirical Flores networks. Percentages indicate the percentage of simulated matrices that were classified as anti-nested, non-nested or nested based on their SES. The EE and NT combination considered the majority of networks to be significantly nested. In contrast, approximately the same number of networks were considered non-nested or nested using EE and NODF. Using qFF, the majority of networks were considered non-nested.

Table 6: The Flores networks were on average significantly nested according to the EE and NT combination. In contrast, the networks were significantly non-nested according to the qFF and NODF combination. There were no significant effects for the average nestedness pattern using NODF.

metric	null model	mean ses	CI low	CI high	t statistic	df	p-value nestedness	p-value non nested	p-value anti nestedness
NT	EE	3.75	2.28	5.22	2.9	36	0.00317	0.997	1.000
NT	qFF	0.148	-0.452	0.749	-5.06	36	1	6.33e-06	1.000
NODF	EE	3.84	0.868	6.81	1.5	36	0.0715	0.929	1.000
NODF	qFF	-1.25	-2.15	-0.347	-6.51	36	1	0.189	0.811

3.7 With the smallest simulated matrices, qFF could struggle to generate unique null matrices after NODF or NT reordering

Given that qFF generates null matrices by reshuffling the original matrix under near-maximal constraints on row and column discrepancies, it was important to verify that qFF was able to generate unique null matrices. Without a completely unique set of null matrices, the standard deviation of null nestedness may not be accurately estimated, resulting in a biased SES estimate. Crucially, “unique” was nestedness metric-specific, as NODF and NT relied on different reshuffling methods prior to calculating nestedness. Across all combinations of matrix type and nestedness metric, the proportion of unique null matrices was graphed as a function of matrix size (Figure 10).

Of note, when considering small matrices generated by qFF, on nested and heterogeneous networks, and rearranged by NODF or NT, not all rearranged null networks were unique. This was to be expected, given the combination of a wide range of row and column fill values alongside a small number of positive interactions with which to make up those marginals. In addition, qFF struggled to generate unique null matrices on small compartmentalised networks, especially in combination with the NT metric. This analysis suggests that care should be taken when interpreting SES values generated with qFF null matrices derived from small PBINs.

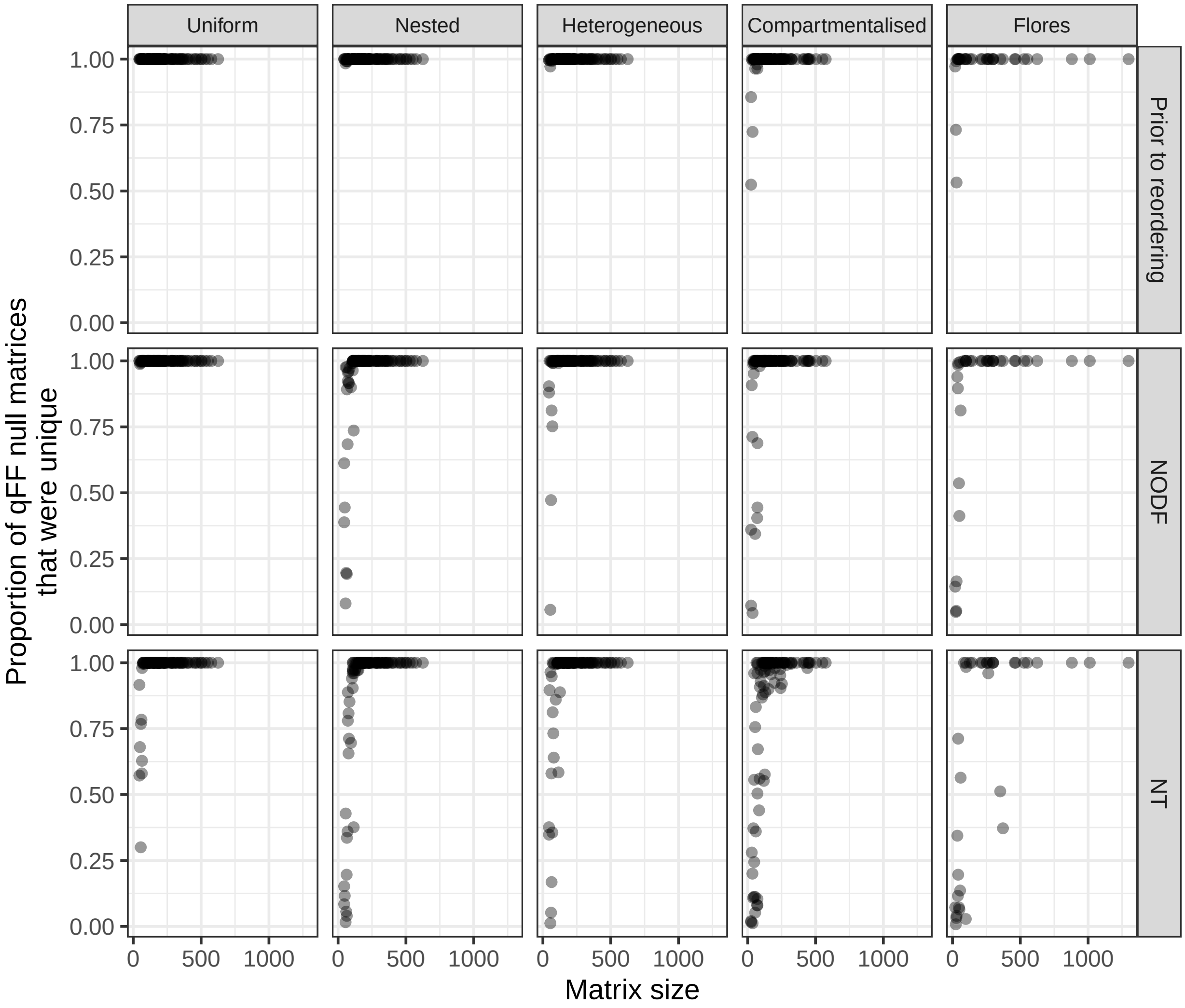

Figure 10: The association between the proportion of qFF null matrices that are unique and matrix size across matrix source and nestedness metric. While most null matrices were unique prior to reordering, some small null matrices lost their uniqueness after reordering by the NODF or NT algorithms.

4 Discussion

Bacteria and phages can co-evolve through arms-race dynamics (ARD), in which resistance and infectivity increase over time (1). Understanding the drivers of this dynamic is important for phage therapy, where ARD could be employed to generate phages with broad infectivity ranges (11). The study of complex Phage-Bacteria Infection Networks (PBINs) has relied on network metrics to distinguish pattern from noise (2). In these analyses, nestedness has been used as a key indicator of ARD, as the nested pattern captures the nested subsets expected to arise from ARD (2). However, previous work has employed a nestedness metric (NT) and a nestedness null model (EE) that did not allow for a meaningful test of ARD. In this work, the performance of the NT and NODF nestedness metrics, in combination with the EE and qFF null models, was compared on simulated and empirical PBINs.

On uniform networks, all combinations of metric and null model correctly identified the matrices as generally non-nested. On nested networks, all combinations of metric and null model correctly identified the matrices as generally nested. However, the NT and qFF combination only characterised 65% of nested matrices as nested, highlighting that NT and NODF differ in their definitions of nestedness. On heterogeneous matrices, EE wrongly identified the matrices as generally nested. In contrast, qFF correctly identified the matrices as generally non-nested. Compartmentalised networks were generally identified as anti-nested by NODF. With NT, there was no significant average effect. Given the superior performance of NODF and qFF, this combination was used to interpret the nestedness of empirical PBINs. This combination suggested that only 2 out of 37 networks were significantly nested. Amongst the remaining networks, 25 were considered non-nested and 10 were considered anti-nested. Overall, this work suggests that the qFF and NODF combination was the most reliable. In addition, this work suggested that a minority of the empirical PBINs were significantly nested.

The conclusion that nestedness may be rare in empirical PBINs is in contrast to the conclusion reached by Flores *et al.* on the same empirical data (13). However, note that the NT and EE combination was used by the authors to conclude widespread nestedness. A recent meta-analysis reiterated the conclusion that that PBINs are generally nested, however this work also used the Temperature metric (46). The authors set out to compare two implementations of the Temperature metric, the original NTC implementation and the BINMANTEST implementation (46). The latter was also the implementation used here, however it does not overcome most of the shortcomings of the original NTC algorithm (40). As the simulations in this work showed, using NT with EE was unreliable due to its propensity to classify heterogeneous matrices as nested. Therefore, using more robust methods, we concluded that the majority of networks in the Flores data were not significantly nested. A future meta-analysis, using the data from Molina *et al.* (46), could test whether this conclusion holds on an independent data set.

Future work lies in analysing the nestedness of experimental PBINs with known co-evolutionary dynamics with qFF and NODF, to ask whether nestedness can be observed to arise from ARD co-evolution. Key to this will be analysing PBINs that span isolates from the entire co-evolutionary experiment (7), rather than separate PBINs for isolates from separate time points (15), as the former is more likely to capture the escalation in resistance and infectivity (2). Additionally, the PBIN row and columns could be ordered by sampling time, rather than by marginal totals as is the default in NODF, to more specifically test for the escalation of resistance and infectivity across time (9, 40). For ecological PBINs, consideration of population scale is required, as PBINs spanning several meta-populations may show nested compartments rather than network-wide nestedness (3, 4). Overall, this work suggests that future analyses would benefit from considering whether a test for ARD is appropriate for the PBIN at hand, prior to performing a nestedness test. Additionally, a future meta-analysis using networks known to have undergone ARD could assess whether nestedness can be detected in such networks, using the robust combination of qFF and NODF. Given previous work that suggested nestedness SES values to associate with network fill when using the EE null model (40), and the importance of having a statistical test that is invariant to network size and fill, the present work assessed the association between SES, network size and network fill across the simulation types and null models. Sizeable associations with size, fill and/or their interaction appeared more common with the EE null model than with the qFF null model. Importantly, this analysis suggests that the SES estimates obtained with qFF on non-compartmentalised matrices were not strongly affected by network size and fill, adding further weight to the conclusion that the combination of qFF and NODF is suitable for nestedness testing. Future work using NODF for nestedness tests on compartmentalised networks may still need to account for network size and fill, due to the strong associations identified in this work.

The present work also highlighted the propensity of the qFF null model to generate null matrices, from small networks, that were not unique after reordering by NODF or NT. Since the null matrices are used to calculate the SES, this could lead to a biased SES estimate on small PBINs. Crucially, the number of matrices with non-unique null matrix sets may have been inflated by the large number of qFF null matrices, 250, used in this work. Future work could assess whether the number of qFF null matrices should be adjusted based on the size of the PBIN. Nevertheless, this work also raised the question of whether it is meaningful to test for nestedness on small PBINs, given the small number of permutations that are possible while retaining similar marginal totals. This, in turn, complicates answering the question raised in the introduction: “does the PBIN have significantly more overlap in resistance and infectivity range sizes than null matrices with a very similar range of generalists and specialists?”

This work suggested that compartmentalised networks, which are expected to arise from local adaptation or FSD co-evolution, are generally anti-nested. This extends previous work that suggested anti-nestedness for two specific instances of compartmentalisation (24). Within PBIN analysis, previous work has used Barber’s modularity metric to infer compartmentalisation (3–5, 13, 16–20): this metric quantifies the extent to which infections occur within pre-defined modules (47), often using a maximisation algorithm to identify the optimal number of modules (48). The present work does not advocate for using anti-nestedness instead of a significant Barber’s modularity for inferring local adaptation or FSD co-evolution. Instead, this analysis highlighted that if anti-nestedness is observed, particularly with NODF, the network may be compartmentalised, and further modularity analysis may prove fruitful. Nevertheless, since the present work used perfectly compartmentalised networks, future work could assess whether compartmentalised networks with some infections outside of modules, as may be expected from natural PBINs (2), are also significantly anti-nested.

This work extends previous attempts to characterise the behaviour of NT and NODF under varying null models (12, 21, 40, 49). While previous work has inferred error rates under a range of null models and/or simulation types, none have done so using random, nested, heterogeneous and compartmentalised simulated networks. Additionally, a novel comparison is made between the widely used EE + NT combination, and the more reliable qFF + NODF combination. Future work lies in simplifying the process of obtaining qFF null matrices. In the present work, qFF matrices were obtained after simulating the entire TP landscape. However, given that the extremes of the landscape are defined by EE and FF (12), future work could specifically obtain qFF matrices through a sequential swap that starts with FF matrices and only swaps up until qFF. This would allow users to obtain qFF null matrices at greater speed and ease. Additionally, future work may lie in the comparison of qFF to the “average” null model used in more recent PBIN experimental work (20), implemented through BIMAT (50), and originally developed by Bascompte *et al.* (51). However, this null model was found to perform very similarly to EE with NODF (“CE” in (49)), suggesting that qFF would be more robust.

It should be noted that qFF was defined as consisting of 5 cells rather than the 8 bottom left cells as originally suggested by Strona *et al.* (12) (i.e. cells (0, 0.1), (0.1, 0) and (0.1, 0.1) were excluded from qFF in this study). When developing this analysis, a trade-off was observed between the two definitions. The 8-cell qFF incorporates lower average host and resistance range discrepancies, thereby being more prone to Type II errors on nested matrices (i.e. the SES from qFF is brought closer to that of FF (21)). Instead, the 5-cell qFF has higher average host range and resistance range size discrepancies, thereby being at greater risk of Type I errors on non-nested matrices. A lower Type II error on nested networks was favoured, as the Type I error of 5-cell qFF with NODF on heterogeneous matrices was deemed sufficiently low, considering that some of these networks would have been significantly nested by chance (Figure 5). Future work lies in a formal comparison of the error rates of 5 and 8-cell qFF, as well as their dependency on network size, as the relative importance of the two error rates is likely to vary between studies.

Overall, this work advocates for the use of qFF and NODF. Additionally, given the collection of empirical networks analysed, this work suggests that nestedness is not a common property of PBINs. Future work could assess the nestedness of empirical PBINs using the robust methodology proposed here, while also taking into account whether ARD are expected to have occured within the network. Finally, improvement in the methods used to infer ARD from PBINs will allow for more accurate identification of the conditions underlying ARD, towards the more effective design of phages for therapeutics (11).

5 References

1. Buckling A, Rainey PB (2002) Antagonistic coevolution between a bacterium and a bacteriophage. *Proceedings of the Royal Society of London Series B: Biological Sciences* 269(1494):931–936.
2. Weitz JS, et al. (2013) Phage–bacteria infection networks. *Trends in Microbiology* 21(2):82–91.
3. Flores CO, Valverde S, Weitz JS (2013) Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *The ISME Journal* 7(3):520–532.
4. Van Cauwenberghe J, et al. (2021) Spatial patterns in phage- rhizobium coevolutionary interactions across regions of common bean domestication. *The ISME Journal*:1–15.
5. Wendling Carolin C., Goehlich Henry, Roth Olivia (2018) The structure of temperate phage–bacteria infection networks changes with the phylogenetic distance of the host bacteria. *Biology Letters* 14(11):20180320.
6. LeGault KN, et al. (2021) Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. *Science*. doi:10.1126/science.abg2166.
7. Fortuna MA, et al. (2019) Coevolutionary dynamics shape the structure of bacteria-phage infection networks. *EVOLUTION* 73(5):1001–1011.
8. Hall AR, Scanlan PD, Morgan AD, Buckling A (2011) Host–parasite coevolutionary arms races give way to fluctuating selection. *Ecology Letters* 14(7):635–642.
9. Ulrich W, Almeida-Neto M, Gotelli NJ (2009) A consumer's guide to nestedness analysis. *Oikos* 118(1):3–17.
10. Mariani MS, Ren Z-M, Bascompte J, Tessone CJ (2019) Nestedness in complex networks: Observation, emergence, and implications. *Physics Reports* 813:1–90.
11. Brockhurst MA, Koskella B, Zhang Q-G (2021) Bacteria-phage antagonistic coevolution and the implications for phage therapy. *Bacteriophages: Biology, Technology, Therapy*, eds Harper DR, Abedon ST, Burrowes BH, McConville ML (Springer International Publishing, Cham), pp 231–251.
12. Strona G, Ulrich W, Gotelli NJ (2018) Bi-dimensional null model analysis of presence-absence binary matrices. *Ecology* 99(1):103–115.
13. Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS (2011) Statistical structure of host–phage interactions. *PNAS* 108(28):E288–E297.
14. Payrató-Borràs C, Hernández L, Moreno Y (2020) Measuring nestedness: A comparative study of the performance of different metrics. *Ecology and Evolution* 10(21):11906–11921.
15. Gurney J, et al. (2017) Network structure and local adaptation in co-evolving bacteria-phage interactions. *MOLECULAR ECOLOGY* 26(7):1764–1777.
16. Larsen ML, Wilhelm SW, Lennon JT (2019) Nutrient stoichiometry shapes microbial coevolution. *ECOLOGY LETTERS* 22(6):1009–1018.
17. Gupta A, et al. (2022) Leapfrog dynamics in phage-bacteria coevolution revealed by joint analysis of cross-infection phenotypes and whole genome sequencing. *Ecology Letters* 25(4):876–888.
18. Beckett SJ, Williams HTP (2013) Coevolutionary diversification creates nested-modular structure in phage-bacteria interaction networks. *Interface Focus* 3(6):20130033.
19. Shaer Tamar E, Kishony R (2022) Multistep diversification in spatiotemporal bacterial-phage coevolution. *Nat Commun* 13(1):7971.
20. Borin JM, et al. (2023) Rapid bacteria-phage coevolution drives the emergence of multi-scale networks. doi:10.1101/2023.04.13.536812.
21. Ulrich W, Gotelli NJ (2007) Null model analysis of species nestedness patterns. *Ecology* 88(7):1824–1831.
22. Moore JE, Swihart RK (2007) Toward ecologically explicit null models of nestedness. *Oecologia* 152(4):763–778.
23. Georjon H, Bernheim A (2023) The highly diverse antiphage defence systems of bacteria. *Nat Rev Microbiol*:1–15.
24. Almeida-Neto M, Jr PRG, Lewinsohn TM (2007) On nestedness analyses: Rethinking matrix temperature and anti-nestedness. *Oikos* 116(4):716–722.

25. Poulin R, Guégan J-F (2000) Nestedness, anti-nestedness, and the relationship between prevalence and intensity in ectoparasite assemblages of marine fish: A spatial model of species coexistence. *International Journal for Parasitology* 30(11):1147–1152.

26. Koskella B, Brockhurst MA (2014) Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol Rev* 38(5):916–931.

27. Gaba S, Ebert D (2009) Time-shift experiments as a tool to study antagonistic coevolution. *Trends in Ecology & Evolution* 24(4):226–232.

28. Herman E (2024) Nestedness of simulated and empirical PBINs. doi:10.5281/zenodo.11422735.

29. R Core Team (2023) *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria) Available at: `https://www.R-project.org/`.

30. Mölder F, et al. (2021) *Sustainable data analysis with snakemake* (F1000Research) doi:10.12688/f1000research.29032.2.

31. R Core Team (2022) *R: A language and environment for statistical computing* (Vienna, Austria) Available at: `https://www.R-project.org/`.

32. Oksanen J, et al. (2020) *Vegan: Community ecology package* Available at: `https://CRAN.R-project.org/package=vegan`.

33. team T pandas development (2023) *Pandas-dev/pandas: pandas* (Zenodo) doi:10.5281/zenodo.7658911.

34. Wickham H, et al. (2019) Welcome to the <span class="nocase">tidyverse</span>. *Journal of Open Source Software* 4(43):1686.

35. Kudva IT, Jelacic S, Tarr PI, Youderian P, Hovde CJ (1999) Biocontrol of escherichia coli O157 with O157-specific bacteriophages. *Appl Environ Microbiol* 65(9):3767–3773.

36. Microsoft, Weston S (2022) *Foreach: Provides foreach looping construct* Available at: `https://CRAN.R-project.org/package=foreach`.

37. Corporation M, Weston S (2020) *doParallel: Foreach parallel adaptor for the 'parallel' package* Available at: `https://CRAN.R-project.org/package=doParallel`.

38. Csardi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal* Complex Systems:1695.

39. Atmar W, Patterson BD (1993) The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia* 96(3):373–382.

40. Almeida-Neto M, Guimarães P, Guimarães Jr PR, Loyola RD, Ulrich W (2008) A consistent metric for nestedness analysis in ecological systems: Reconciling concept and measurement. *Oikos* 117(8):1227–1239.

41. Caldwell AR (2022) Exploring equivalence testing with the updated TOSTER r package. doi:10.31234/osf.io/ty8de.

42. Wickham H (2016) *ggplot2: Elegant graphics for data analysis* (Springer-Verlag New York) Available at: `https://ggplot2.tidyverse.org`.

43. Bates D, et al. (2019) lme4: Linear mixed-effects models using 'eigen' and S4. Available at: `https://CRAN.R-project.org/package=lme4` [Accessed August 23, 2019].

44. Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13):1–26.

45. Rodríguez-Gironés MA, Santamaría L (2006) A new algorithm to calculate the nestedness temperature of presence–absence matrices. *Journal of Biogeography* 33(5):924–935.

46. Molina F, et al. (2021) A new pipeline for designing phage cocktails based on phage-bacteria infection networks. *Front Microbiol* 12. doi:10.3389/fmicb.2021.564532.

47. Barber MJ (2007) Modularity and community detection in bipartite networks. *Phys Rev E* 76(6):066102.

48. Liu X, Murata T (2009) Community detection in large-scale bipartite networks. *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pp 50–57.

49. Strona G, Fattorini S (2014) On the methods to assess significance in nestedness analyses. *Theory Biosci* 133(3):179–186.

50. Flores CO, Poisot T, Valverde S, Weitz JS (2016) BiMat: A MATLAB package to facilitate the analysis of bipartite networks. *Methods in Ecology and Evolution* 7(1):127–132.

51. Bascompte J, Jordano P, Melián CJ, Olesen JM (2003) The nested assembly of plant–animal mutualistic networks. *PNAS* 100(16):9383–9387.

Chapter 5: General discussion

P. aeruginosa lung infections are a major cause of morbidity and mortality in Cystic Fibrosis (CF) patients (1). Given the difficulty of eradicating such infections with antibiotics, phages are being considered as novel CF therapeutics (2). This thesis set out to increase the understanding of phage-bacteria interactions in clinical CF *P. aeruginosa* isolates, the pan-immune system of these bacterial isolates, and the methods with which co-evolutionary patterns are discerned in Phage-Bacteria Infection Networks (PBINs).

In the first data chapter, phage resistance was investigated in light of receptor modifications and defence systems carried by clinical CF *P. aeruginosa* isolates. While it has long been known that bacteria rely on extracellular and intracellular defence mechanisms (3), the range of known defence systems has recently been substantially expanded (4). The aims of the present work were to assess whether similarity in receptor modifications or defence system repertoire was associated with similarity in phage resistance, whether carrying more receptor mutations or defence systems was associated with broader resistance ranges, and whether individual receptor mutations or defence systems associated with resistance to specific phages. Similarity in defence system repertoire was associated with similarity in phage resistance, suggesting that defence systems are associated with resistance specificity. In contrast, only the number of variants in lipopolysaccharide (LPS) receptor genes was associated with resistance breadth. Given that only four T4P phages were included, the study was likely underpowered to detect a similar effect for T4P biosynthesis genes. Additionally, given that the sample consisted of 23 hosts with highly divergent defence system repertoires, the distribution of defence systems may have been too patchy to observe an association between the number of defence systems and resistance breadth. Interestingly, Costa *et al.* did find clinical *P. aeruginosa* isolates with more systems to have broader resistance ranges (5). While this thesis exclusively studied *P. aeruginosa* isolates from CF patients, Costa *et al.* studied strains from a wide variety of clinical settings including sputum, blood cultures and contact lenses (5). Future work could compare the diversity of defence systems in the two isolate collections. Additionally, the phage panels used in the two assays could be swapped, to ask whether the association is specific to the phage panels used. Finally, the genetic diversity in the phage panels could be compared.

In terms of individual associations between phage resistance and the presence of genetic variants, variants in *rmlA,* involved in the production of the LPS O-antigen, and *fimT,* a gene of unknown function thought to be involved in T4P biosynthesis, appeared to associate with resistance. Future work could clone the identified variants into PAO1 to test for phage resistance, analogous to the approach taken by Costa *et al.* with defence systems (5). Since variants in *rmlA* have previously been shown to provide phage resistance (6–8), in addition to a trade-off with ceftazidime resistance (8), the present work suggests that phage therapy could exploit this synergy in clinical CF isolates. Future work could investigate the involvement of *fimT* in phage resistance, which as of yet is elusive, using the studied clinical CF isolates. Finally, the study highlighted the difficulty in associating phage resistance with the presence of individual defence systems in a small panel of hosts with divergent defence system repertoires: while presence of druantia appeared to associate with phage resistance, an adsorption assay suggested that phages were not adsorbing to the druantia-carrying hosts. While more laboursome, the approach taken by Costa *et al.* may be preferable in future work: measure adsorption across the entire infectivity network, and specifically investigate resistance mediated by defence systems within phage-host pairs that show adsorption (5). Finally, given that the present and previous phage panels were obtained from sewage systems (5, 8), and the availability of sputum samples from which the present isolates were obtained, future work could investigate the resistance of the CF isolates to phages from their native clinical environment. Nevertheless, the first data chapter suggested that receptor variation and defence system carriage could play complementary roles in determining phage resistance. This suggests that it will be important to characterise receptor variants alongside defence system repertoires, especially given the potential for phage-antibiotic synergism in phage therapy (9).

In the second data chapter, the pan-immune system of clinical CF *P. aeruginosa* isolates was characterised. Given recent reports of variability in *P. aeruginosa* defence systems (10, 11), and reports of changing defence system repertoires in natural settings (12, 13), this work investigated whether such patterns could be found in the clinical setting. While the isolates carried a broad and open pan-immune system, closely-related isolates carried mostly identical repertoires, and distantly-related isolates never shared identical repertoires. In addition, despite most systems localising to variable genomic regions, systems were rarely found in prophages and signs of horizontal movement of systems across the phylogenetic tree were rare. Nevertheless, given that many variable regions containing defence systems could not be grouped into spots due to the fragmented nature of the assemblies, it is possible that some evidence for horizontal gene transfer was missed. Future work using long-read assemblies could re-assess the evidence for the low frequency of similar defence system spots between distant isolates. Finally, the defence system repertoire appeared to be mostly stable over time within closely-related isolates obtained from the same patient. Importantly, this work suggested that the prevailing narrative of frequent defence system movement (4) does not apply to clinical CF *P. aeruginosa* isolates. Future work using environmental *P. aeruginosa* isolates, obtained over time, could ask whether movement appears more frequently outside of the clinical setting, where bacteria may encounter a more diverse and changing set of phages. In contrast to the stability discovered in the defence system repertoires, previous work with CF *P. aeruginosa* isolates identified many modifications in LPS, which is frequently used by phages for attachment (14). As *P. aeruginosa* isolates are isolated as part of future phage therapy trials, it will be interesting to ask whether therapeutic phage pressure induces defence system movement, and whether it selects for a different set of receptor modifications from those identified in the clinical isolates studied in this thesis. Understanding the dynamics of both modes of resistance may prove fruitful for the development of effective phage therapeutics, whereby phages could be developed to work synergistically with antibiotics and to overcome common modes of resistance within the CF lung.

In the final data chapter, the dependence of nestedness and anti-nestedness patterns on the choice of nestedness metric and null model was investigated. Nestedness has been of interest in PBIN analysis due to its theoretical association with arms-race dynamics (ARD), a form of co-evolution whereby hosts and phages continuously increase their resistance and infectivity capabilities, respectively (15). Additionally, previous work suggested that anti-nestedness may be a property of compartmentalised networks (16), which are thought to arise from fluctuating selection dynamics (FSD) or local adaptation (15). Importantly, previous analyses of PBINs have relied on a null model (EE) and sometimes a metric (T) that are prone to returning significant nestedness with non-nested networks (17). Given the recent development of the Tuning-Peg algorithm, and its quasi-FF (qFF) null model which appeared to have good statistical properties (18), the effects of the choice of nestedness metric and null model were estimated on simulated and empirical networks. Overall, this work revealed that the qFF null model more accurately classified simulated networks as nested or non-nested. In addition, the analysis suggested that compartmentalised networks are generally anti-nested. Crucially, the study suggested that the widely-cited notion of nestedness being frequent in empirical PBINs should be revised, as the robust qFF null model suggested that very few of the PBINs were nested. To ease application of the qFF null model, a simplified function that obtains qFF matrices without evaluating the full Tuning-Peg landscape could be developed for the `vegan` package in R, complementing its current set set of nestedness functions (19). Importantly, the present work lacked an explicit assessment of the nestedness of experimental PBINs with known underlying co-evolutionary dynamics. Given that such experimental work has previously been published (20–26), future work could consider whether robustly estimated nestedness or anti-nestedness are actually good indicators of experimental ARD or FSD dynamics, respectively. If these turn out to be reliable indicators, the NODF metric and the qFF null model investigated here could help in the development of phage therapeutics and in understanding within-patient co-evolutionary dynamics (27).

Overall, this thesis highlighted the complexities of phage resistance and defence system dynamics in clinical CF *P. aeruginosa*. Namely, receptor modifications and defence system carriage may play complementary roles in determining resistance, and the isolate's defence system repertoires may stay relatively stable through time. Additionally, this thesis has suggested that the field of PBIN nestedness analysis requires a shift from the commonly used EE null model, to the more robust qFF null model. Given the need for novel antimicrobial therapeutics for CF patients, which acquire chronic lung infections with antimicrobial resistant pathogens, phages could become invaluable therapeutic tools. Future therapeutic development could be improved by considering the full spectrum of phage resistance mechanisms, as well as the conditions that promote beneficial modes of co-evolution in phage cultivation and within-patient application. Such developments may allow for existing antibiotic therapy for *P. aeruginosa* infections in CF patients to be complemented with antibiotic-synergising phages, offering a novel approach to treating one of the biggest sources of morbidity and mortality within the disorder.

References

1. Thornton CS, Parkins MD (2023) Microbial Epidemiology of the Cystic Fibrosis Airways: Past, Present, and Future. *Seminars in Respiratory and Critical Care Medicine* 44(02):269–286.
2. Chan BK, Stanley G, Modak M, Koff JL, Turner PE (2021) Bacteriophage therapy for infections in CF. *Pediatric Pulmonology* 56(S1):S4–S9.
3. Dy RL, Richter C, Salmond GPC, Fineran PC (2014) Remarkable mechanisms in microbes to resist phage infections. *Annual Review of Virology* 1(1):307–331.
4. Georjon H, Bernheim A (2023) The highly diverse antiphage defence systems of bacteria. *Nature Reviews Microbiology*:1–15.
5. Costa AR, et al. Accumulation of defense systems in phage resistant strains of Pseudomonas aeruginosa. doi:10.1101/2022.08.12.503731.
6. Wright RCT, Friman V-P, Smith MCM, Brockhurst MA (2018) Cross-resistance is modular in bacteria–phage interactions. *PLOS Biology* 16(10):e2006057.
7. Garbe J, Bunk B, Rohde M, Schobert M (2011) Sequencing and characterization of pseudomonas aeruginosa phage JG004. *BMC Microbiology* 11(1):102.
8. Nordstrom HR, et al. (2022) Genomic characterization of lytic bacteriophages targeting genetically diverse pseudomonas aeruginosa clinical isolates. *iScience* 25(6). doi:10.1016/j.isci.2022.104372.
9. Mangalea MR, Duerkop BA (2020) Fitness Trade-Offs Resulting from Bacteriophage Resistance Potentiate Synergistic Antibacterial Strategies. *Infection and Immunity* 88(7). doi:10.1128/IAI.00926-19.
10. Tesson F, et al. (2021) *Systematic and quantitative view of the antiviral arsenal of prokaryotes.*
11. Johnson MC, et al. (2023) Core defense hotspots within pseudomonas aeruginosa are a consistent and rich source of anti-phage defense systems. *Nucleic Acids Research*:gkad317.
12. Hussain FA, et al. (2021) Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science* 374(6566):488–492.
13. LeGault KN, et al. (2021) Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. *Science*. doi:10.1126/science.abg2166.
14. Camus L, Vandenesch F, Moreau K (2021) From genotype to phenotype: Adaptations of pseudomonas aeruginosa to the cystic fibrosis environment. *Microbial Genomics* 7(3):000513.
15. Weitz JS, et al. (2013) Phage–bacteria infection networks. *Trends in Microbiology* 21(2):82–91.
16. Almeida-Neto M, Jr PRG, Lewinsohn TM (2007) On nestedness analyses: rethinking matrix temperature and anti-nestedness. *Oikos* 116(4):716–722.
17. Ulrich W, Gotelli NJ (2007) Null Model Analysis of Species Nestedness Patterns. *Ecology* 88(7):18241831.
18. Strona G, Ulrich W, Gotelli NJ (2018) Bi-dimensional null model analysis of presence-absence binary matrices. *Ecology* 99(1):103–115.
19. Oksanen J, et al. (2020) *Vegan: Community ecology package* Available at: `https://CRAN.R-project.org/package=vegan`.
20. Gurney J, et al. (2017) Network structure and local adaptation in co-evolving bacteria-phage interactions. *MOLECULAR ECOLOGY* 26(7, SI):1764–1777.
21. Fortuna MA, et al. (2019) Coevolutionary dynamics shape the structure of bacteria-phage infection networks. *Evolution* 73(5):1001–1011.
22. Gupta A, et al. (2022) Leapfrog dynamics in phage-bacteria coevolution revealed by joint analysis of cross-infection phenotypes and whole genome sequencing. *Ecology Letters* 25(4):876–888.
23. Borin JM, et al. Rapid bacteria-phage coevolution drives the emergence of multi-scale networks. doi:10.1101/2023.04.13.536812.
24. Poullain V, Gandon S, Brockhurst MA, Buckling A, Hochberg ME (2008) The Evolution of Specificity in Evolving and Coevolving Antagonistic Interactions Between a Bacteria and Its Phage. *Evolution* 62(1):1–11.

25. Scanlan PD, Hall AR, Lopez-Pascua LDC, Buckling A (2011) Genetic basis of infectivity evolution in a bacteriophage. *MOLECULAR ECOLOGY* 20(5):981–989.

26. Hall AR, Scanlan PD, Morgan AD, Buckling A (2011) Host-parasite coevolutionary arms races give way to fluctuating selection. *ECOLOGY LETTERS* 14(7):635–642.

27. Brockhurst MA, Koskella B, Zhang Q-G (2021) Bacteria-Phage Antagonistic Coevolution and the Implications for Phage Therapy. eds Harper DR, Abedon ST, Burrowes BH, McConville ML (Springer International Publishing, Cham), pp 231–251.