Introduction

The major human bacterial pathogen, *Pseudomonas aeruginosa,* causes multidrug-resistant infections, particularly in people with underlying immunodeficiencies or inflammatory lung diseases, such as Cystic Fibrosis (CF). However, it remains unclear how *P. aeruginosa* has evolved into a highly adapted, globally disseminated pathogen causing both CF and non-CF infections.

Rationale

We therefore sought to understand the pathogenic evolution of *P. aeruginosa,* by defining when and how epidemic clones emerged and spread, exploring how they have adapted to specific hosts, and understanding how within-host evolution has influenced ongoing transmission patterns.

Results

We analysed a global collection of 9,829 isolates of *P. aeruginosa,* identifying 21 major clones, which we term 'epidemic'. These epidemic clones were more likely to be detected in infected humans than in the non-built environment, caused most clinical *P. aeruginosa* infections worldwide, were widely distributed across the phylogenetic tree, and had all spread globally. We estimate that these epidemic clones emerged from ancestral locations distributed around the world, and then expanded non-synchronously between the late 17^{th} and late 20^{th} centuries, potentially driven by changes in human population density, migration patterns, and/or air pollution. Through inferring a pan-genome graph, we identified significant differences between epidemic and sporadic isolate genomes in the acquisition of genes involved in specific cellular processes, such as transcriptional control and metabolism (***Figure 0***).

We found that epidemic clones appeared to have intrinsic preferences for CF or non-CF individuals and discovered that clinical isolates clustered in transcriptional space based on the host-preference of epidemic clones, with a clear expression signature of genes positively and negatively associated with CF affinity. We found that high CF affinity clones were better able to survive within CF macrophages, in part mediated by expression of the stringent response modulator DksA1, suggesting that enhanced host innate immune evasion might explain the intrinsic success at infecting CF patients of certain epidemic clones (***Figure 0***).

Examining the recent mutation history of individual clones to understand how epidemic clones of *P. aeruginosa* have adapted to the human host through multiple rounds of within-patient evolution by analysing, we identified 224 out of 5641 genes that had a higher total mutational burden than expected by chance, which we term 'pathoadaptive'.

We found that the products of these pathoadaptive genes were tightly interconnected, indicating their likely coordinated functional roles. Many genes were more frequently mutated

in either CF or non-CF isolates, suggesting that distinct functional programmes were being modified as part of host-specific adaptation (***Figure 0***).

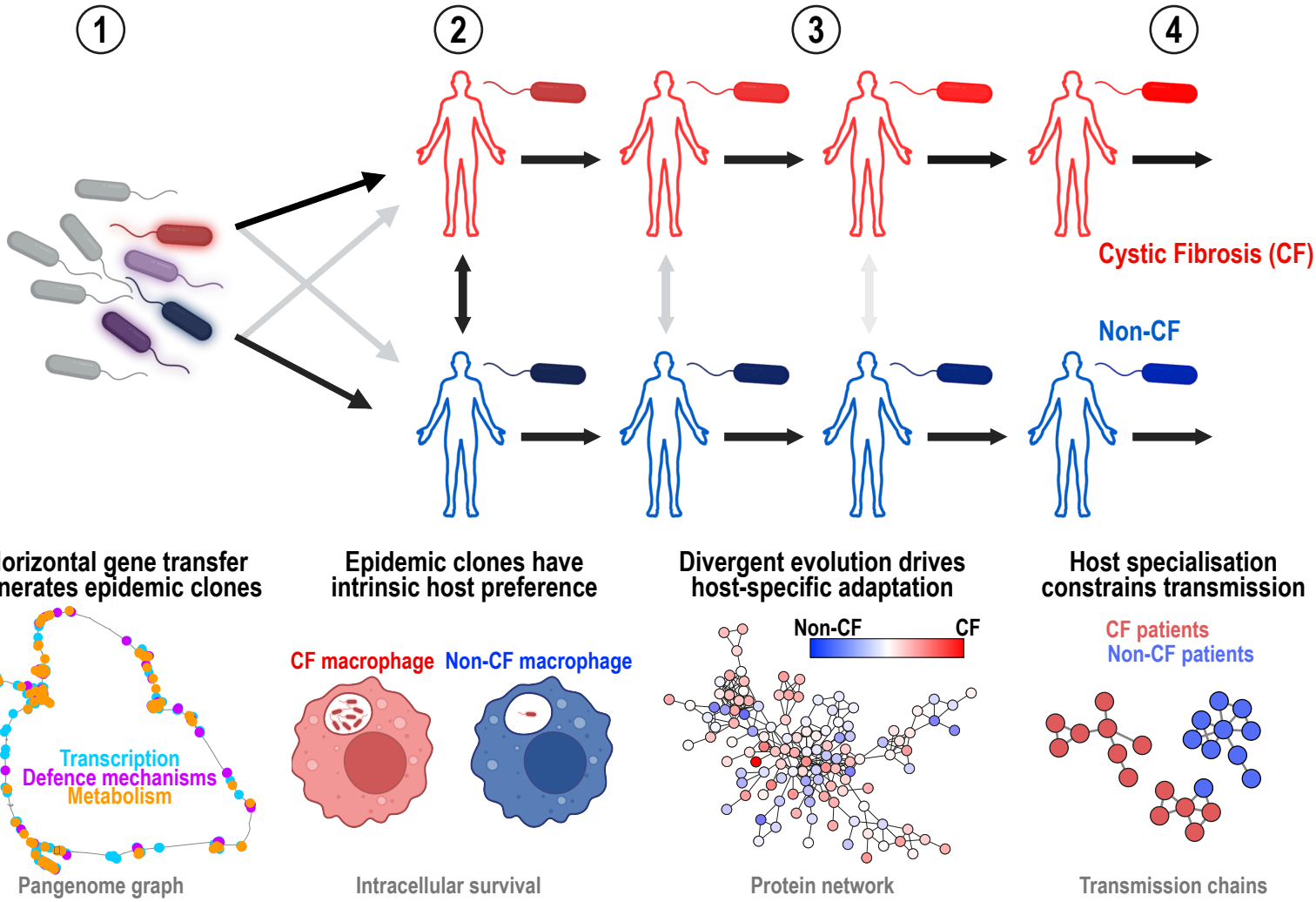
Pathoadaptive genes were frequently associated with changes in transmissibility and/or host-specific adaptation, thereby potentially driving host specialisation. In support of this notion, we found strong evidence of cross-infection either between CF patients or between non-CF patients, but very little CF to non-CF transmission (***Figure 0***).

Conclusion

Our findings describe the key sequential steps involved in the evolution of *P. aeruginosa* from an environmental organism to a major human pathogen: saltatory evolution caused by horizontal gene transfer generating epidemic clones; varying intrinsic host affinities of these clones (linked to specific transcriptional changes enabling survival within macrophages); and multiple rounds of convergent, host-specific adaptation, eventually resulting in the loss of their ability to transmit between different patient groups. Our work thus highlights the importance of global surveillance and cross-infection prevention in averting the emergence of future epidemic clones.

Caption

Figure 0. Host-specific evolution of *Pseudomonas aeruginosa.* We define key steps in the pathogenic evolution of *P. aeruginosa*: (1) Environmental clones with epidemic potential are created through horizontal transfer of genes with specific cellular roles (revealed through comparative pan-genome graph analysis of epidemic and sporadic isolates); (2) Emerging epidemic clones have intrinsic varying preference for Cystic Fibrosis (CF) hosts, which is transcriptionally driven and reflects increased ability to survive within CF macrophages; (3) Multiple rounds of within-host evolution then leads to distinct evolutionary trajectories for isolates infecting CF versus non-CF patients, mediated through different patterns of mutations in 224 ‘pathoadaptive’ genes which influence transmissibility and/or host-specific adaptation; (4) As a result of host specialisation, transmission is constrained between CF and non-CF patients. Top panel created with BioRender.com


1
2
3
4
Cystic Fibrosis (CF)
Non-CF
Horizontal gene transfer generates epidemic clones
Epidemic clones have intrinsic host preference
Divergent evolution drives host-specific adaptation
Host specialisation constrains transmission
Transcription
Defence mechanisms
Metabolism
CF macrophage
Non-CF macrophage
Non-CF
CF
CF patients
Non-CF patients
Pangenome graph
Intracellular survival
Protein network
Transmission chains

Evolution and host-specific adaptation of *Pseudomonas aeruginosa*

Aaron Weimann[1,2,3,4], Adam M Dinan[1,2,3], Christopher Ruis[1,2,3,4], Audrey Bernut[5], Stéphane Pont[5], Karen Brown[1,2,6], Judy Ryan[1,2], Lúcia Santos[7], Louise Ellison[2], Emem Ukor[2,6], Arun P. Pandurangan[1,8,9], Sina Krokowski[1,2], Tom L. Blundell[1,8,9], Martin Welch[8], Beth Blane[9], Kim Judge[11], Rachel Bousfield[9,10], Nicholas Brown[10], Josephine M. Bryant[11], Irena Kukavica-Ibrulj[12], Giordano Rampioni[13,14], Livia Leoni[13], Patrick T. Harrison[7], Sharon J Peacock[9,10], Nicholas R. Thomson[11,15], Jeff Gauthier[12], Jo L Fothergill[16], Roger C Levesque[12], Julian Parkhill[4], R. Andres Floto[1,2,3,6,9,10]

1. Victor Phillip Dahdaleh Heart & Lung Research Institute, University of Cambridge, Cambridge UK
2. University of Cambridge Molecular Immunity Unit, MRC Laboratory of Molecular Biology, Cambridge UK
3. Cambridge Centre for AI in Medicine, University of Cambridge UK
4. Department of Veterinary Medicine, University of Cambridge; Cambridge, UK.
5. Laboratory of Pathogens and Host Immunity (LPHI), UMR5235, CNRS/Université de Montpellier, Montpellier, France.
6. Cambridge Centre for Lung Infection, Royal Papworth Hospital, Cambridge, UK
7. Department of Physiology, Bioscience Institute, University College Cork, Ireland
8. Department of Biochemistry, University of Cambridge, Cambridge, UK
9. Department of Medicine, University of Cambridge, Cambridge UK
10. Cambridge University Hospitals Trust, Cambridge, UK
11. Wellcome Sanger Institute, Hinxton, UK
12. Institut de biologie intégrative et des systèmes (IBIS), Université Laval; Québec City, Québec, Canada.
13. Department of Science, University Roma Tre, Rome, Italy
14. IRCCS Fondazione Santa Lucia, Rome, Italy
15. London School of Hygiene and Tropical Medicine, london, UK
16. Department of Clinical Infection, Microbiology and Immunology, University of Liverpool, Liverpool UK

1

Correspondence to:
Andres Floto (arf27@cam.ac.uk),
Julian Parkhill (jp369@cam.ac.uk), or
Roger Levesque (rclevesq@ibis.ulaval.ca)

Abstract

The major human bacterial pathogen, *Pseudomonas aeruginosa,* causes multidrug-resistant infections in people with underlying immunodeficiencies or structural lung diseases, such as Cystic Fibrosis (CF). We show that a few environmental isolates, driven by horizontal gene acquisition, have become dominant epidemic clones that have sequentially emerged and spread through global transmission networks over the past 200 years. These clones demonstrate varying intrinsic propensities for infecting CF or non-CF individuals (linked to specific transcriptional changes enabling survival within macrophages), have undergone multiple rounds of convergent, host-specific adaptation, and have eventually lost their ability to transmit between different patient groups. Our findings thus explain the pathogenic evolution of *P. aeruginosa* and highlight the importance of global surveillance and cross-infection prevention in averting the emergence of future epidemic clones.

P. aeruginosa is found widely in natural and man-made environments (*1–5*) and has become an increasingly important opportunistic human pathogen, causing acute nosocomial lung, soft tissue, and systemic infections (*6*), as well as chronic pulmonary infections in individuals with underlying inflammatory lung diseases, such as Chronic Obstructive Pulmonary Disease (COPD) (*7*), CF (*8*), and non-CF bronchiectasis (*9*), where it causes increased morbidity and mortality (*8*, *10*). Antimicrobial resistance (AMR) in *P. aeruginosa* is increasing globally (recognised by its inclusion in the WHO ESKAPE pathogen list (*11*)) and is responsible for over 300,000 deaths annually (*12*).

Although individuals frequently acquire *P. aeruginosa* independently from the environment, hospital-based person-to-person transmission is well recognised in people with CF, leading to strict cohort and individual segregation within clinics (*8*), but has not been considered a risk in other patient cohorts (*13*). Nevertheless, epidemic clones of *P. aeruginosa* have been identified in both CF and non-CF infection studies (*14–17*), suggesting the presence of wide-spread transmission networks. In CF, these epidemic clones (such as the Liverpool Epidemic Strain; LES) are associated with accelerated inflammatory lung damage and worse clinical outcomes, demonstrating that epidemic clones can emerge with increased virulence in particular patient groups.

We therefore sought to understand the pathogenic evolution of *P. aeruginosa,* by defining when and how epidemic clones emerged and spread, exploring how they have adapted to specific hosts, and understanding how within-host evolution has influenced ongoing transmission patterns.

Results

Phylogenetic analysis reveals global epidemic clones.

We first analysed a globally-distributed collection of 9,829 human, animal, and environmental isolates of *P. aeruginosa,* including 9,573 human clinical samples from 2,765 patients (*14*, *18–27*) (***Figure 1A; Table S1***), which we grouped into 596 genetically-related clones (based on pairwise single nucleotide polymorphism (SNP) distances) and then stratified by the number of patients infected by each clone (***Figure 1B; Figure S1; Supplementary Methods***). We identified 21 major clones each containing isolates from at least 30 individuals, which we term ‘epidemic’ and refer to by their majority multi-locus sequence type (*28*).

We found that these epidemic clones were more likely to be detected in infected humans than in the non-built environment (Fisher exact test $p = 7.80 \times 10^{-12}$), caused 51 % of all clinical *P. aeruginosa* infections worldwide, were widely distributed across the phylogenetic tree, and had all spread globally (***Figure 1C; Figure S1; Table S1***).

Inferring population expansion and geography of epidemic clones over time.

Since our sample collections spanned over 100 years (from 1900 to 2018), we wondered whether we could explore the historical origins of epidemic clones using Bayesian temporal reconstruction (*29*). We estimate that epidemic clones emerged non-synchronously between the late 17th and late 20th centuries (***Figure S2***) and (through Skyline demographic modelling (*30*); ***Figure S3***) have each undergone at least one major population expansion between 1850 and 2000 (***Figure 1D***), suggesting (considering only extant clones) that *P. aeruginosa* has undergone relatively recent changes in host-pathogen dynamics, potentially driven by changes in human population density, migration patterns (*31*, *32*), and/or increased susceptibility to infections (caused, for example, by increased air pollution during industrialisation (*33*–*35*)).

For epidemic clones with geographically clustered clades (association test p value <0.05), we implemented Bayesian phylogeographic methods (*36*) to infer the ancestral location of clones (accepting the limitations of our opportunistic sample collection). In some instances, for example ST235, we were able to find statistical support for the direction of intercontinental spread (from South America to North America and Europe, and then subsequently Asia and Africa; ***Figure S4***), whereas, for other clones such as ST17 and ST27, we could identify only that transmission was restricted to between Europe and North America and peaked in the second half of the 20th century (***Figure S4***). We conclude that epidemic clones have likely arisen from ancestral locations distributed around the world.

Horizontal gene transfer may drive emergence of epidemic clones.

We next asked why some *P. aeruginosa* clones had become epidemic and investigated whether gene acquisition, through horizontal transfer, might have driven large jumps in human infectivity through saltatory evolution (*37*) (meaning abrupt changes in evolutionary fitness caused by sudden large genetic changes), as previously observed in *Mycobacterium abscessus* (*38*). To accurately analyse the *P. aeruginosa* accessory genome, we first inferred a pan-genome graph (using *Panaroo* (*39*)), with nodes as clusters of orthologous genes and two nodes linked by an edge if they were found adjacent in any contig (***Figure 1E***). We then compared the accessory genomes of representatives of ancestral epidemic clones with those of sporadic isolates and found that epidemic clones had significant enrichment of genes involved in transcriptional regulation, inorganic ion transport, lipid metabolism, and protein turnover, with significant depletion of genes involved in bacterial defence and secretion (Fisher exact test, FDR = 0.1; ***Figure 1E; Figure S5, Table S2***); suggesting that fundamental changes in bacterial physiology might have driven the development of epidemic clones.

Epidemic clones vary in their intrinsic host preference.

We next examined the types of human infection caused by the epidemic clones and found that ST146, also known as the Liverpool Epidemic Strain, caused infection almost exclusively in

people with CF while other clones (e.g., ST175 and ST309) caused infection only in non-CF individuals, with a range of CF and non-CF host distributions for other clones (***Figure 2A***). While our collection was opportunistically sampled, our findings were replicated when re-analysing existing systematic UK surveillance data (*40*) for the frequency of CF and non-CF lung infections caused by each epidemic clone (***Figure S6***). We hypothesised that this apparent clone-specific host preference might relate to intrinsic differences in bacterial behaviour between clones. We found no evidence, on pangenome analysis, for an association of host affinity with acquisition of genes with a particular function (***Figure S7***) and therefore we reasoned that changes in gene expression could explain differences in host preference. We therefore analysed a previous transcriptomic study of clinical *P. aeruginosa* isolates (nearly all from non-CF patients; (*25*)) that were represented in our sample collection, and found that clinical isolates clustered in transcriptional space based on the host-preference of epidemic clones ($p < 0.001$; ***Figure 2B; Supplementary Methods****)*.

We next explored whether there were any patterns in gene expression associated with clone predisposition for specific human hosts and identified, using a negative binomial generalised linear model, a clear expression signature of 624 genes positively associated, and 514 genes inversely associated, with affinity for causing CF infection (Wald test, FDR = 0.05***; Figure 2C; Table S3***).

Increased survival within macrophages of high CF affinity clones

To identify a potential mechanism by which differential gene expression could alter the host-preference of epidemic clones, we conducted a multi-dimensional phenotypic characterisation of 49 representative isolates (almost all from non-CF patients to minimise the contribution of secondary host-specific adaptation) from epidemic clones with low (ST235, ST111), intermediate (ST253), and high levels of CF affinity (ST17, ST27, ST146) obtained from the International Pseudomonas Consortium Database (*21*). We initially examined isolate behaviour in established assays of *Pseudomonas* virulence (biofilm formation, siderophore production, swim and twitch motility, and production of caseinase and gelatinase) but could find no correlation with clonal host affinity (***Figure S8***).

We therefore decided to test the ability of representative isolates of different epidemic clones to withstand intracellular killing by macrophages that, together with neutrophils, are thought to be the first line of defence against bacterial lung infection (*41*, *42*). We found significantly increased intracellular survival and replication of isolates from the high CF affinity clone (ST27) compared to isolates from the low/intermediate clones ST111 and ST235 in both wildtype and CF (F508del homozygous) isogenic macrophage cell lines (***Figure 2D; Table S4***), suggesting that enhanced host innate immune evasion might explain the intrinsic success of particular epidemic clones in infecting CF patients.

Host preference of epidemic strains is mediated by DksA1 expression.

To further explore the bacterial mechanisms contributing to enhanced intracellular survival of bacterial isolates from high CF affinity clones, we interrogated the differentially-expressed gene set (*Figure 2C*) and discovered that both the expression of the stringent response modulator DksA1 and the activation of its regulon were associated with CF affinity (***Figure 3A***). Since DksA1 had previously been implicated in enhancing *P. aeruginosa* survival within mouse macrophages and increasing tolerance to H_2O_2 (*43*), we examined the impact of deleting DksA1 on bacterial survival by using *in vitro* and *in vivo* models of non-CF and CF infection employing *P. aeruginosa* PAO1 wildtype and isogenic DksA1 and DksA2 double knockout (ΔDksA1,2), and complemented (ΔDksA1,2::DksA1) strains (ΔDksA1,2 double knockout mutants were used due to gene redundancy (*44*).

We first tested bacterial survival in macrophages and found that, while all three strains were effectively killed in wildtype THP1 macrophages, only the ΔDksA1,2 strain could be killed by isogenic CF (F508del knock-in) cells while wildtype and complemented bacteria were able to resist macrophage killing and replicate intracellularly (***Figure 3B***). Our findings reveal a number of important features of host-*Pseudomonas* interactions: firstly, that there are intrinsic defects in CF macrophages that facilitate intracellular survival of *P. aeruginosa* (observations that are supported by previous *in vitro* reports (*45*) and by our *in vivo* experiments in zebrafish, where deletion (*46*) or morpholino knockdown of the Cystic Fibrosis Transmembrane Regulator (*cftr*) compromises survival after intravenous infection (***Figure S9***)); secondly, that intracellular survival in CF macrophages is mediated by DskA1, raising intriguing mechanistic questions about the role of the stringent response in surviving the phagosomal environment; and finally, that differences in DksA1 expression across epidemic clones may explain their observed different abilities to survive within macrophages and, potentially as a consequence, their varying intrinsic host preferences.

To explore the potential role of DksA1 further, we examined the behaviour of fluorescently labeled wildtype, ΔDksA1,2 mutant and complemented *P. aeruginosa* PAO1 during *in vivo* infection in zebrafish larvae (***Figure 3C-G***). We observed increased survival of both control and *cftr* morpholino-treated fish after infection with ΔDksA1,2 bacteria compared to wildtype and complemented strains (***Figure 3D***); findings which correlated with an observed decreased survival of ΔDksA1,2 bacteria *in vivo* (***Figure 3E***).

We then exploited the optical transparency of zebrafish larvae to track the interaction of macrophages with *P. aeruginosa* following intramuscular infection using intravital confocal microscopy. Utilizing a fluorescent macrophage reporter fish line (*Tg(mpeg1:mcherry-F)ump2*; (*47*)), we could clearly identify macrophages and distinguish extracellular from intracellular

fluorescent bacteria (***Figure 3F***). We found no difference in the mobilization of macrophage to the site of infection, or the proportion of infected macrophages, in control and *cftr* morpholino fish infected with wildtype, ΔDksA1,2 mutant or complemented *P. aeruginosa* PAO1 bacteria (***Figure S9***; ***Figure 3G***). We did however observe a clear reduction in macrophage bacterial burden (suggesting reduced intracellular replication) following ΔDksA1,2 infection of both control and *cftr* morpholino fish lines (***Figure 3G***), confirming the critical role of DksA1 for intracellular survival during *in vivo P. aeruginosa* infection.

Taken together, our data indicate that intrinsic elevations in DksA1 expression levels in some epidemic clones may have enabled them to exploit potential innate immune defects in CF and adopt the specific evolutionary strategy of replicating within macrophages.

Convergent host-specific adaptation of *P. aeruginosa*

We next examined how, once selected from the environment, epidemic clones of *P. aeruginosa* have adapted to the human host through multiple rounds of within-patient evolution by analysing the recent mutation history of individual clones.

By reconstructing mutations that had likely occurred since the emergence of each clone, we found strong evidence for convergent molecular evolution, identifying 224 out of 5641 genes that had a higher total mutational burden than expected by chance (Poisson test, FDR = 0.05; ***Figure 4A***; ***Table S5***), which we term 'pathoadaptive'. Mutations in these pathoadaptive genes were more likely to be non-synonymous and deleterious (by variant effect annotation (*48*)) than those found in other genes (Fisher exact test $p < 1.0 \times 10^{-16}$), and were predicted to be more likely to cause protein dysfunction, as estimated by both sequence conservation methods (*SIFT* (*49*); Wilcoxon rank-sum test, $p = 9.04 \times 10^{-15}$) and structural modelling approaches (*FoldX* (*50*); Wilcoxon rank-sum test, $p = 1.34 \times 10^{-6}$) (***Figure S10***), suggesting that pathoadaptation is largely driven by loss-of-function mutations. We explored the functional impact of pathoadaptive mutations experimentally by using existing RNAseq datasets (*25*) to examine the effect of transcription factor variants on expression of their previously characterised regulons (*51–54*) and found that clinical isolates with pathoadaptive variants in several transcription factor had statistically significant shifts in regulon expression levels compared to controls (two tailed t-test with adjusted p-value < 0.00014; ***Figure S10***), supporting the concept of a general loss-of-function evolutionary process driving *P. aeruginosa* pathoadaptation.

We were able to functionally annotate the majority of these pathoadaptive genes using prior published information (*55*), identifying many of them as having established roles in recognised pathogenic processes including biofilm formation, antibiotic resistance and LPS modification (***Figure S11***). The number of genes with an established function was much higher among

pathoadaptive genes than in other genes (Fisher exact test $p < 1.0 \times 10^{-16}$; **Figure S11**), potentially reflecting their central role in *P. aeruginosa* pathobiology. We also characterised the function of 41 pathoadaptive genes experimentally by *de novo* screening relevant transposon mutants in a series of functional assays to quantify virulence traits (***Supplementary Methods; Figure S11***).

We next examined the nature of host adaptation achieved by individual *P. aeruginosa* isolates by using the profile of their pathoadaptive gene mutations to map them in evolutionary space (defined by the presence or absence of mutations in the 224 pathoadaptive genes). We found that CF isolates clustered separately from others and had accumulated more mutations, suggesting that the CF lung is a distinct niche with different selective pressures compared to other lung or non-lung environments (***Figure 4B***).

We found that the products of these pathoadaptive genes were tightly interconnected, with more protein-protein interactions than expected by chance (*STRING* database (*56*); $p < 1 \times 10^{-16}$; ***Figure 4C; Figure S12***), indicating their likely coordinated functional roles. We observed that 70 genes were more frequently mutated in CF isolates while 55 genes were more commonly mutated in non-CF isolates (Fisher exact test, FDR = 0.1). Among genes that were more commonly mutated in CF or non-CF, we found several overrepresented pathways (using Gene Ontology biological pathway enrichment analysis with TopGO (*57*)), suggesting that distinct functional programmes were being modified as part of host-specific adaptation (***Figure 4C***). For example, CF isolates were more likely to have mutated AlgU, a key regulator of mucoidy (*58*) (with mutations occurring predominantly at the interface between this sigma-H factor and its negative regulator protein, MucA), and PcnA, a putative nicotinamidase (with mutations found within the protein core or at sites of protein-protein interaction; ***Figure S13***). In contrast, non-CF isolates were more likely to have mutated LadS, a calcium-responsive histidine kinase (*59*) (with mutations concentrated in the N-terminal (sensor) and transmembrane domains), and in the putative choline transporter BetT2 (*60*), with helix-breaking mutations found within the transmembrane domain (***Figure S13***).

Distinct evolutionary trajectories lead to host specialisation

We then used ancestral state reconstruction to determine the order of acquisition of each pathoadaptive mutation and thereby recreate the evolutionary trajectory of each isolate. We found that, on average, CF isolates had longer trajectories than non-CF isolates (with 20.5 compared to 11.2 steps, Wilcoxon signed-rank test $p < 1.0 \times 10^{-16}$, ***Figure S14***). By looking at the frequency of mutations in pathoadaptive genes at each evolutionary time-step, we were able to cluster genes into 5 groups with distinct temporal signatures (***Figure 5A; Figure S15***), suggesting that mutations in specific genes may be important at different stages of evolution (as noted previously, specifically for AMR evolution (*61*)).

We wondered therefore whether specific sets of pathoadaptive genes might be driving distinct evolutionary processes such as host-specific adaptation, person-to-person transmission, or both. To examine this, we inferred the impact of each pathoadaptive gene on host-specific pathoadaptation (by examining the relative frequency of gene mutations occurring in CF compared to non-CF lung isolates), and on bacterial transmissibility (based on the frequency of specific gene mutations being found in isolates from at least two patients) to create a map of the contribution of each pathoadaptive gene to each of these evolutionary processes (***Figure 5B***), annotating each gene by previously known or experimentally derived function, or the type of temporal mutation signature observed (***Figure 5C; Figure S16***).

We found that, while some pathoadaptive genes were associated with changes in either transmissibility or host-specific pathoadaptation, many were implicated in both processes (***Figure S16***). For example, mutations in several genes (such as *mvfR* and *morA*) occurred early in the evolutionary trajectories of isolates and were associated with both adaptation to the non-CF host and increased transmissibility, while mutations in other genes (such as in *fusA1* and *algU*) occurred late in evolutionary journeys and were associated with adaptation to the CF host and decreased transmissibility (***Figure S16***).

Since we observed a likely deleterious impact of several pathoadaptive mutations on transmissibility, we examined whether pathoadaptation might lead to host specialisation and result in reduced transmission of isolates between CF and non-CF individuals. To explore this possibility, we used the genomic relatedness of isolates to plot the number (and proportion) of transmission links over a range of SNP pairwise thresholds (representing transmission chains of various lengths) and found strong evidence for CF-to-CF patient transmission and non-CF to non-CF patient cross-infection but very little CF to non-CF transmission (***Figure 5D***). Additionally, we reconstructed transmission clusters at a specific SNP threshold (26 SNPs), based on the measured genetic diversity within individual patients (***Supplementary Methods***), and found transmission clusters of variable sizes but very few containing both CF and non-CF patients (***Figure 5E***). It seems likely therefore that host-specific pathoadaptation of epidemic *P. aeruginosa* clades limits transmission between different hosts.

Discussion

Our findings describe the key sequential steps involved in the evolution of *P. aeruginosa* from an environmental organism to a major human pathogen. We identify horizontal gene acquisition as a likely driver for the emergence of epidemic clones from the environment through saltational evolution (as previously described for *M. abscessus* (*38*)) and infer their spatio-temporal spread which suggests an increasing rate of new epidemic clone expansions over time (accepting that only extant clones are considered). We identify an intrinsic and

variable host-specific affinity across epidemic clones with CF preference potentially causally associated with improved intracellular survival in macrophages. We then describe how deleterious mutations in a discrete set of functionally interrelated genes likely mediate further host specialisation (through multiple rounds of within-patient adaptation) and onward transmission, thereby plausibly explain the observed lack of person-to-person transmission between CF and non-CF patients.

Our work highlights the importance of preventing pathogenic evolution by minimising cross-infection, not just within CF cohorts (where infection control measures are well established) but also between non-CF patients, and emphasises how global surveillance and targeted monitoring of high-risk patient groups will be needed to detect expansion, pathoadaptation, and transmission of new and extant epidemic *P. aeruginosa* clones.

Methods summary

Genomic datasets and clone assignment

We collated Pseudomonas aeruginosa genome datasets from studies of antibiotic resistance (*18*, *23*–*25*), from individuals with cystic fibrosis (*14*, *26*) and non-CF bronchiectasis (*22*); from the International *Pseudomonas* Consortium (*21*); and from studies targeting high-risk clones (*19*, *20*, *27*). Newly sequenced genomes from the TeleCF study, which involved adults with CF (n = 15) who underwent home monitoring for six months and were chronically infected with *Pseudomonas*, and from bacteaemia infections (n = 365) as part of the UK BSAC bacteraemia resistance programme (*62*) and from patients attending hospitals in Cambridgeshire, UK were included. DNA was extracted using QIAxtractor (QIAgen), and samples were sequenced on the Illumina HiSeq 2000 and 2500 and X10 platforms.

Variants were called by mapping reads against the *P. aeruginosa* PAO1 reference genome (accession number AE004091.2) using the multiple_mappings_to_bam 1.6 pipeline with default parameters (https://github.com/sanger-pathogens/bact-gen-scripts) employing BWA (*63*) for mapping followed by stringent QC filtering and removing samples with an excess number of minority variants. Ariba 2.14.6 (*64*) was used for multi-locus sequencing typing (*28*). FastTree (2.1.10) was used to infer a global phylogenetic tree (*65*).

Clone were assigned by first grouping samples based on pairwise SNP distances using the ultra-metric pairwise group method with arithmetic means (UPGMA) and then applying a cut-off of 7000 SNPs. SNP-sites was used to infer a clone-specific alignment of variable sites (*66*). Gubbins version 2.4.1 (*67*) was used to remove recombination for individual clones with at least four available genomes.

Dating and phylogeography

Molecular dating was performed for all 21 epidemic clone separately. Potential hypermutators (distorting the temporal signal) were removed by identifying samples with an unusual ratio of

transition and transversion mutations. The temporal signal was assessed with TempEst (*68*) by comparing collection dates with root-to-tip distances using non-dated phylogenetic tress inferred with RAxML 8.2.12 (*69*). The significance of the signal was assessed using a permutation test using a custom script (https://github.com/chrisruis/tree_scripts/blob/main/bootstrap_TempEst_rttd_date.R). Clones with a significant temporal signal in this test ($P < 0.05$) were taken forward for molecular dating with BEAST 2.6.6 (*29*). We modelled the population history using the coalescent Bayesian skyline population prior. Convergence was assessed with Tracer 1.7.1 (*70*) with 10% burn-in. For clones that didn't pass the bootstrap randomisation test (N = 9), a uniform prior for the substitution rate was set informed by the above clones.

For clones that passed the initial test, we ran a more thorough date randomisation test as described previously (*71*). The estimated median substitution rates and most recent common ancestor dates for randomised BEAST runs (n=10) did not overlap with those of the runs using real collection dates, indicating a significant temporal signal. To test whether each epidemic clone has undergone a historical population expansion, we analysed Bayesian skyline plot estimates of relative genetic diversity across the posterior distribution.

The association index was computed to find evidence of geographic clustering within clone phylogenies (*72*). We identified clones for further spatiotemporal analysis where less than 5% of randomisations had a higher association index than the non-permuted dataset. Asymmetric phylogeographic discrete trait reconstructions of the isolate continents were then performed using the BEAST classic 1.9.0 package of BEAST 2.6.6 (*29*). Subsampling to account for overrepresentation of certain continents was repeated five times and results compared between subsamples. Spread 0.9.7.1 (*73*) was used to identify candidate migration routes between continents (Bayes factor >= 3).

Pan-genome analysis

Genomes were assembled from short-read data and Panaroo 1.2.8 (*39*) was used to cluster the gene sequences from all samples into gene families and to infer a graphical pan-genome, which was reduced, ordered against the *P. aeruginosa* PAO1 genome, pruned of long-range connections, and then visualised (see Supplementary Methods for details). Parsimony ancestral character state reconstruction was then used to infer gene gains and losses on the branches of the rooted tree leading to the ancestral epidemic and sporadic clones. Gene functions were then annotated using EggNog-mapper 2.1.6 (*74*), with the number of genes gained and annotated within a specific COG functional category compared using a Fisher exact test (adjusted p-value < 0.05).

Macrophage infection experiments

Isogenic F508del homozygous THP1 cells were created from wild type THP1 cells (obtained from ATCC) using CRISPR-Cas9 editing and confirmed by Sanger sequencing (see Supplementary Methods). Wild type (WT) and F508del THP-1 monocytes were cultured, seeded at 200,000 cells/mL, and differentiated into macrophages (as previously described (*75*)) before being exposed to pooled clinical isolates of *P. aeruginosa* at a multiplicity of infection (MOI) of 1:1 and then incubated at 37 ^{o}C for 1 hour before the supernatant was removed and cells were lysed at 1h time point or incubated in fresh media for further time points (2h or 4h) before supernatant removal, cell lysis, and DNA extraction and sequencing. Strain abundance was quantified using the mSWEEP 1.4.0 sequence-based deconvolution method (*76*). Strains with less than 1% abundance at the 1h time point were excluded from the analysis.

Transcriptomic analysis

Gene expression data for clinical *P. aeruginosa* strains (and the UCBPP- PA14 wildtype control strain) was obtained (*25*), and pseudoaligned to strain-specific gene indices to produce abundance estimates using Kallisto (*77*). Length-scaled abundance estimates were size-factor normalised by the median ratio method and modelled as a response to CF proportion per genomic cluster (as defined by the number of CF vs non-CF patients and environmental samples) using a negative binomial generalised linear model (GLM) with DESeq2 (*78*). The coefficients for gene models were assessed using the Wald test (adjusted p-value < 0.05). To assess the distribution and clustering of transcriptional diversity of strains with respect to CF proportion, we used k-means clustering (k=20) on the principal components (PCs) of the gene expression data, and then then calculated the mean standard deviation (σ) of the CF proportion by cluster (mean σ = 0.135). A permutation test was used to assess significance (see Supplementary Methods for details)

Zebrafish infection models

The following zebrafish lines were used (see Supplementary Methods for details): wild type AB line; the knockout cftr sh540 mutant (*46*); Tg(mpeg1:mcherry-F)ump2 line (*47*). The morpholino for cftr knockdown (5'-GACACATTTTGGACACTCACACCAA-3') were prepared and injected into one-cell-stage as previously described (*79*). Systemic infections were achieved by microinjection of GFP-expressing *P. aeruginosa* strains into the caudal vein of 30 hours post-fertilization (hpf) zebrafish embryos as previously described (*80*), with survival post infection assessed daily and viable *in vivo P. aeruginosa* quantified by colony forming units (CFU) at 1 day post infection (dpi). Macrophage responses were examined by intramuscular injection of anesthetized Tg(mpeg1:mcherry-F)ump2 larvae at 3 days post fertilisation (dpf) with GFP-expressing fluorescent *P. aeruginosa* as previously described (*79*, *81*). Macrophage

chemotaxis, phagocytosis, and intracellular *P. aeruginosa* burden were quantified by confocal microscopy (see Supplementary Methods for details).

Mutational burden analysis

Treetime 0.8.1 (*82*) was used to reconstruct ancestral character states of every nucleotide position in every clone. We then implemented a pipeline (*83*) to identify single nucleotide changes and annotate variant effect in their phylogenetic context using the gene annotation from Pseudomonas.com (PAO1 107) and the ancestral character state reconstructions (*55*). Parsimony ancestral character state reconstruction was used to infer ancestral insertions and deletions, using SNPeff (*48*) for variant effect annotation.

We assessed the mutational burden of every gene based on the number of non-synonymous variants across all clones (using a Poisson test, adjusted p-value < 0.05). The 224 genes passing the adjusted p-value threshold were used to query the STRING 11.5 database (*56*) of protein-protein interaction. Pathoadaptive genes were assigned to 17 functional categories based on the gene products description on Pseudomonas.com (*55*) (Figure S1). A Fisher exact test was used to compare the number of assigned with the number of unassigned genes among pathoadaptive genes and non-hits.

Impact of amino acid changes on protein stability and structural analysis

All amino acid changes were analysed with SIFT 4G 6.2.1 (*49*) and FoldX 5 (*50*) (see Supplementary Methods for details). A two-tailed t-test was used to compare the averaged scores per gene/protein scores between mutational burden test hits and non-hits. Mutational frequencies were mapped on the structural models of the identified hotspot genes in *P. aeruginosa* using the Chimera molecular modelling package (*84*). Models were downloaded from the Protein Data Bank and UniProt (*85*).

Phenotyping of pathoadaptive gene mutations

PAO1 mutants with transposon insertions in 154 pathoadaptive genes (selected from the Manoil library (*86*) were arrayed in 96 well plates and imaged using the Phenobooth Imager (Singer Instruments) to quantify the following phenotypic traits: swimming motility, twitching motility, siderophore production, caseinase activity, gelatinase activity, and rhamnolipid production (see Supplementary Methods for details).

To assess the association between genetic variants and the expression of transcription factor (TF) regulons, gene expression data from (*25*) were pseudoaligned to strain-specific gene sets and the normalised expression levels of TF regulons were compared between strains with and without genetic variants using Welch's two-sample *t*-tests (adjusted p-value < 0.05) (see Supplementary Methods for details).

Transmission and host selectivity of pathoadaptive mutations

To assess the transmissibility of pathoadaptive changes, the number of mutations that had been observed in at least two isolates (from different patients) was compared with hitherto untransmitted mutations using a Fisher exact test (adjusted p-value < 0.1). TopGO 2.4.6 was used for functional enrichment analysis of the host-specific Gene Ontology biological pathway annotation compared to background (*57*), using annotations from *Pseudomonas.com* (*55*) (p-value < 0.05).

Mutations in pathoadaptative genes were stratified by the (ancestral) infection type (CF or non-CF) of every branch based on outgroup-rooted rooted clone trees. To assess host-specific pathoadapdation, the number of CF vs non-CF mutations were compared using a Fisher exact test (adjusted p-value < 0.1). Mutations on branches with non-concordant ancestral infection types were discarded.

Trajectories were inferred as the sequence of mutations in pathoadaptive genes since the emergence of the clone ancestor as implied by the PAO1-rooted tree stratified by cystic fibrosis (CF) and non-CF infection types. Mutation frequencies were position normalised and the frequency plots of the 40 genes with the lowest p-value from the mutational-burden test were manually assigned into five groups of genes with similar frequency curve shapes. Trendlines were generated by locally-weighted smoothing.

We established a relatedness cut-off to define potential transmission links using pairwise SNP differences between pairs of isolate genomes from the same patient (n = 81 patients). We then identified potential transmission events as isolates from the same clone sampled from different patients that differed by 26 SNPs or fewer, visualised using Cytoscape.

References

1. J. L. Whitby, A. Rampling, PSEUDOMONAS ÆRUGINOSA CONTAMINATION IN DOMESTIC AND HOSPITAL ENVIRONMENTS. *Lancet* **299**, 15–17 (1972).

2. S. K. Green, M. N. Schroth, J. J. Cho, S. K. Kominos, V. B. Vitanza-jack, Agricultural plants and soil as a reservoir for Pseudomonas aeruginosa. *Appl. Microbiol.* **28**, 987–991 (1974).

3. J.-P. Pirnay, S. Matthijs, H. Colak, P. Chablain, F. Bilocq, J. Van Eldere, D. De Vos, M. Zizi, L. Triest, P. Cornelis, Global Pseudomonas aeruginosa biodiversity as reflected in a Belgian river. *Environ. Microbiol.* **7**, 969–980 (2005).

4. L. M. Ringen, C. H. Drake, A study of the incidence of Pseudomonas aeruginosa from various natural sources. *J. Bacteriol.* **64**, 841–845 (1952).

5. S. Crone, M. Vives-Flórez, L. Kvich, A. M. Saunders, M. Malone, M. H. Nicolaisen, E. Martínez-García, C. Rojas-Acosta, M. Catalina Gomez-Puerto, H. Calum, M. Whiteley, R. Kolter, T. Bjarnsholt, The environmental occurrence of Pseudomonas aeruginosa. *APMIS* **128**, 220–231 (2020).

6. ECDC, “Healthcare-associated infections acquired in intensive care units. Annual epidemiological report for 2017” (European Centre for Disease Prevention and Control., 2019).

7. T. F. Murphy, A. L. Brauer, K. Eschberger, P. Lobbins, L. Grove, X. Cai, S. Sethi, Pseudomonas aeruginosa in chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **177**, 853–860 (2008).

8. S. Rajan, L. Saiman, Pulmonary infections in patients with cystic fibrosis. *Semin. Respir. Infect.* **17**, 47–56 (2002).

9. S. Finch, M. J. McDonnell, H. Abo-Leyah, S. Aliberti, J. D. Chalmers, A Comprehensive Analysis of the Impact of Pseudomonas aeruginosa Colonization on Prognosis in Adult Bronchiectasis. *Ann. Am. Thorac. Soc.* **12**, 1602–1611 (2015).

10. M. R. Kosorok, L. Zeng, S. E. West, M. J. Rock, M. L. Splaingard, A. Laxova, C. G. Green, J. Collins, P. M. Farrell, Acceleration of lung disease in children with cystic fibrosis after Pseudomonas aeruginosa acquisition. *Pediatr. Pulmonol.* **32**, 277–287 (2001).

11. E. Tacconelli, E. Carrara, A. Savoldi, S. Harbarth, M. Mendelson, D. L. Monnet, C. Pulcini, G. Kahlmeter, J. Kluytmans, Y. Carmeli, M. Ouellette, K. Outterson, J. Patel, M. Cavaleri, E. M. Cox, C. R. Houchens, M. L. Grayson, P. Hansen, N. Singh, U. Theuretzbacher, N. Magrini, WHO Pathogens Priority List Working Group, Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect. Dis.* **18**, 318–327 (2018).

12. K. S. Ikuta, *et al* Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **400**, 2221–2248 (2022).

13. P. J. Mitchelmore, J. Randall, M. J. Bull, K. A. Moore, P. A. O’Neill, K. Paszkiewicz, E. Mahenthiralingam, C. J. Scotton, C. D. Sheldon, N. J. Withers, A. R. Brown, Molecular epidemiology of Pseudomonas aeruginosa in an unsegregated bronchiectasis cohort sharing hospital facilities with a cystic fibrosis cohort. *Thorax*, doi: 10.1136/thoraxjnl-2016-209889 (2017).

14. P. J. Stapleton, C. Izydorcyzk, S. Clark, A. Blanchard, P. W. Wang, Y. Yau, V. Waters, D. S. Guttman, Pseudomonas aeruginosa strain sharing in early infection among children with cystic fibrosis. *Clin. Infect. Dis.*, doi: 10.1093/cid/ciaa788 (2020).

15. G. A. Tramper-Stranders, C. K. van der Ent, T. F. W. Wolfs, J. L. L. Kimpen, A. Fleer, U. Johansen, H. K. Johansen, N. Høiby, Pseudomonas aeruginosa diversity in distinct paediatric patient groups. *Clin. Microbiol. Infect.* **14**, 935–941 (2008).

16. T. E. Woo, R. Lim, M. G. Surette, B. Waddell, J. C. Bowron, R. Somayaji, J. Duong, C. H. Mody, H. R. Rabin, D. G. Storey, M. D. Parkins, Epidemiology and natural history of Pseudomonas aeruginosa airway infections in non-cystic fibrosis bronchiectasis. *ERJ Open Res* **4** (2018).

17. A. AbdulWahab, S. J. Taj-Aldeen, E. Ibrahim, S. H. Abdulla, R. Muhammed, I. Ahmed, Y. Abdeen, O. Sadek, M. Abu-Madi, Genetic relatedness and host specificity of Pseudomonas aeruginosa isolates from cystic fibrosis and non-cystic fibrosis patients. *Infect. Drug Resist.* **7**, 309–316 (2014).

18. V. N. Kos, M. Déraspe, R. E. McLaughlin, J. D. Whiteaker, P. H. Roy, R. A. Alm, J. Corbeil, H. Gardner, The resistome of Pseudomonas aeruginosa in relationship to phenotypic susceptibility. *Antimicrob. Agents Chemother.* **59**, 427–436 (2015).

19. M. P. Moore, I. L. Lamont, D. Williams, S. Paterson, I. Kukavica-Ibrulj, N. P. Tucker, D. T. D. Kenna, J. F. Turton, J. Jeukens, L. Freschi, B. A. Wee, N. J. Loman, S. Holden, S. Manzoor, P. Hawkey, K. W. Southern, M. J. Walshaw, R. C. Levesque, J. L. Fothergill, C. Winstanley, Transmission, adaptation and geographical spread of the Pseudomonas aeruginosa Liverpool epidemic strain. *Microb Genom* **7** (2021).

20. C. López-Causapé, L. M. Sommer, G. Cabot, R. Rubio, A. A. Ocampo-Sosa, H. K. Johansen, J. Figuerola, R. Cantón, T. J. Kidd, S. Molin, A. Oliver, Evolution of the Pseudomonas aeruginosa mutational resistome in an international Cystic Fibrosis clone. *Sci. Rep.* **7**, 5555 (2017).

21. L. Freschi, A. T. Vincent, J. Jeukens, J.-G. Emond-Rheault, I. Kukavica-Ibrulj, M.-J. Dupont, S. J. Charette, B. Boyle, R. C. Levesque, The Pseudomonas aeruginosa pan-genome provides new insights on its population structure, horizontal gene transfer and pathogenicity. *Genome Biol. Evol.*, doi: 10.1093/gbe/evy259 (2018).

22. Y. Hilliam, M. P. Moore, I. L. Lamont, D. Bilton, C. S. Haworth, J. Foweraker, M. J. Walshaw, D. Williams, J. L. Fothergill, A. De Soyza, C. Winstanley, Pseudomonas aeruginosa adaptation and diversification in the non-cystic fibrosis bronchiectasis lung. *Eur. Respir. J.* **49** (2017).

23. J. Chilam, S. Argimón, M. T. Limas, M. L. Masim, J. M. Gayeta, M. L. Lagrada, A. M. Olorosa, V. Cohen, L. T. Hernandez, B. Jeffrey, K. Abudahab, C. M. Hufano, S. B. Sia, M. T. G. Holden, J. Stelling, D. M. Aanensen, C. C. Carlos, Philippines Antimicrobial Resistance Surveillance Program, Genomic surveillance of Pseudomonas aeruginosa in the Philippines, 2013-2014. *Western Pac Surveill Response J* **12**, 4–18 (2021).

24. E. Del Barrio-Tofiño, C. López-Causapé, G. Cabot, A. Rivera, N. Benito, C. Segura, M. M. Montero, L. Sorlí, F. Tubau, S. Gómez-Zorrilla, N. Tormo, R. Durá-Navarro, E. Viedma, E. Resino-Foz, M. Fernández-Martínez, C. González-Rico, I. Alejo-Cancho, J. A. Martínez, C. Labayru-Echverria, C. Dueñas, I. Ayestarán, L. Zamorano, L. Martinez-Martinez, J. P. Horcajada, A. Oliver, Genomics and Susceptibility Profiles of Extensively Drug-Resistant Pseudomonas aeruginosa Isolates from Spain. *Antimicrob. Agents Chemother.* **61**, AAC.01589-17 (2017).

25. A. Khaledi, A. Weimann, M. Schniederjans, E. Asgari, T.-H. Kuo, A. Oliver, G. Cabot, A. Kola, P. Gastmeier, M. Hogardt, D. Jonas, M. R. Mofrad, A. Bremges, A. C. McHardy, S. Häussler, Predicting antimicrobial resistance in Pseudomonas aeruginosa with machine learning-enabled molecular diagnostics. *EMBO Mol. Med.* **12**, e10264 (2020).

26. R. L. Marvig, L. M. Sommer, S. Molin, H. K. Johansen, Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. *Nat. Genet.* **47**, 57–64 (2015).

27. J. F. Turton, L. Wright, A. Underwood, A. A. Witney, Y.-T. Chan, A. Al-Shahib, C. Arnold, M. Doumith, B. Patel, T. D. Planche, J. Green, R. Holliman, N. Woodford, High-Resolution Analysis by Whole-Genome Sequencing of an International Lineage (Sequence Type 111) of Pseudomonas aeruginosa Associated with Metallo-Carbapenemases in the United Kingdom. *J. Clin. Microbiol.* **53**, 2622–2631 (2015).

28. B. Curran, D. Jonas, H. Grundmann, T. Pitt, C. G. Dowson, Development of a multilocus sequence typing scheme for the opportunistic pathogen Pseudomonas aeruginosa. *J. Clin. Microbiol.* **42**, 5644–5649 (2004).

29. R. Bouckaert, J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, A. J. Drummond, BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).

30. A. J. Drummond, A. Rambaut, B. Shapiro, O. G. Pybus, Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192 (2005).

31. W. H. McNeill, Human Migration in Historical Perspective. *Popul. Dev. Rev.* **10**, 1–18 (1984).

32. A. P. Dobson, E. R. Carper, Infectious Diseases and Human Population History. *Bioscience* **46**, 115–126 (1996).

33. F. J. Kelly, J. C. Fussell, Air pollution and airway disease. *Clin. Exp. Allergy* **41**, 1059–1071 (2011).

34. A. J. Chauhan, S. L. Johnston, Air pollution and infection in respiratory illness. *Br. Med. Bull.* **68**, 95–112 (2003).

35. M. Kampa, E. Castanas, Human health effects of air pollution. *Environ. Pollut.* **151**, 362–367 (2008).

36. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* **5**, e1000520 (2009).

37. M. I. Katsnelson, Y. I. Wolf, E. V. Koonin, On the feasibility of saltational evolution. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 21068–21075 (2019).

38. J. M. Bryant, K. P. Brown, S. Burbaud, I. Everall, J. M. Belardinelli, D. Rodriguez-Rincon, D. M. Grogono, C. M. Peterson, D. Verma, I. E. Evans, C. Ruis, A. Weimann, D. Arora, S. Malhotra, B. Bannerman, C. Passemar, K. Templeton, G. MacGregor, K. Jiwa, A. J. Fisher, T. L. Blundell, D. J. Ordway, M. Jackson, J. Parkhill, R. A. Floto, Stepwise pathogenic evolution of *Mycobacterium abscessus*. *Science* **372** (2021).

39. G. Tonkin-Hill, N. MacAlasdair, C. Ruis, A. Weimann, G. Horesh, J. A. Lees, R. A. Gladstone, S. Lo, C. Beaudoin, R. A. Floto, S. D. W. Frost, J. Corander, S. D. Bentley, J. Parkhill, Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).

40. K. Martin, B. Baddal, N. Mustafa, C. Perry, A. Underwood, C. Constantidou, N. Loman, D. T. Kenna, J. F. Turton, Clusters of genetically similar isolates of Pseudomonas aeruginosa from multiple hospitals in the UK. *J. Med. Microbiol.* **62**, 988–1000 (2013).

41. J. Day, A. Friedman, L. S. Schlesinger, Modeling the immune rheostat of macrophages in the lung in response to infection. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11246–11251 (2009).

42. A. Craig, J. Mai, S. Cai, S. Jeyaseelan, Neutrophil recruitment to the lungs during bacterial pneumonia. *Infect. Immun.* **77**, 568–575 (2009).

569 43. A. Fortuna, D. Collalto, V. Schiaffi, V. Pastore, P. Visca, F. Ascenzioni, G. Rampioni, L.
570 Leoni, The Pseudomonas aeruginosa DksA1 protein is involved in H2O2 tolerance and
571 within-macrophages survival and can be replaced by DksA2. *Sci. Rep.* **12**, 10404 (2022).

572 44. A. Fortuna, H. Bähre, P. Visca, G. Rampioni, L. Leoni, The two Pseudomonas aeruginosa
573 DksA stringent response proteins are largely interchangeable at the whole transcriptome
574 level and in the control of virulence-related traits. *Environ. Microbiol.* **23**, 5487–5504
575 (2021).

576 45. K. B. Turton, R. J. Ingram, M. A. Valvano, Macrophage dysfunction in cystic fibrosis:
577 Nature or nurture? *J. Leukoc. Biol.* **109**, 573–582 (2021).

578 46. A. Bernut, C. A. Loynes, R. A. Floto, S. A. Renshaw, Deletion of cftr Leads to an
579 Excessive Neutrophilic Response and Defective Tissue Repair in a Zebrafish Model of
580 Sterile Inflammation. *Front. Immunol.* **11** (2020).

581 47. A. Bernut, J.-L. Herrmann, K. Kissa, J.-F. Dubremetz, J.-L. Gaillard, G. Lutfalla, L.
582 Kremer, Mycobacterium abscessus cording prevents phagocytosis and promotes
583 abscess formation. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E943-52 (2014).

584 48. P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D.
585 M. Ruden, A program for annotating and predicting the effects of single nucleotide
586 polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118;
587 iso-2; iso-3. *Fly* **6**, 80–92 (2012).

588 49. R. Vaser, S. Adusumalli, S. N. Leng, M. Sikic, P. C. Ng, SIFT missense predictions for
589 genomes. *Nat. Protoc.* **11**, 1–9 (2016).

590 50. J. Delgado, L. G. Radusky, D. Cianferoni, L. Serrano, FoldX 5.0: working with RNA, small
591 molecules and a new graphical interface. *Bioinformatics* **35**, 4168–4169 (2019).

592 51. S. Westbrock-Wadman, D. R. Sherman, M. J. Hickey, S. N. Coulter, Y. Q. Zhu, P.
593 Warrener, L. Y. Nguyen, R. M. Shawar, K. R. Folger, C. K. Stover, Characterization of a
594 Pseudomonas aeruginosa efflux pump contributing to aminoglycoside impermeability.
595 *Antimicrob. Agents Chemother.* **43**, 2975–2983 (1999).

596 52. K. Poole, K. Tetro, Q. Zhao, S. Neshat, D. E. Heinrichs, N. Bianco, Expression of the
597 multidrug resistance operon mexA-mexB-oprM in Pseudomonas aeruginosa: mexR
598 encodes a regulator of operon expression. *Antimicrob. Agents Chemother.* **40**, 2021–
599 2028 (1996).

600 53. L. Rust, E. C. Pesci, B. H. Iglewski, Analysis of the Pseudomonas aeruginosa elastase
601 (lasB) regulatory region. *J. Bacteriol.* **178**, 1134–1140 (1996).

602 54. T. L. Yahr, D. W. Frank, Transcriptional organization of the trans-regulatory locus which
603 controls exoenzyme S synthesis in Pseudomonas aeruginosa. *J. Bacteriol.* **176**, 3832–
604 3838 (1994).

605 55. G. L. Winsor, E. J. Griffiths, R. Lo, B. K. Dhillon, J. A. Shay, F. S. L. Brinkman, Enhanced
606 annotations and features for comparing thousands of Pseudomonas genomes in the
607 Pseudomonas genome database. *Nucleic Acids Res.* **44**, D646-53 (2016).

608 56. D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic,
609 N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, C. von Mering, STRING v11: protein–

protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2018).

57. A. Alexa, J. Rahnenführer, T. Lengauer, Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).

58. D. W. Martin, M. J. Schurr, H. Yu, V. Deretic, Analysis of promoters controlled by the putative sigma factor AlgU regulating conversion to mucoidy in Pseudomonas aeruginosa: relationship to sigma E and stress response. *J. Bacteriol.* **176**, 6688–6696 (1994).

59. U. N. Broder, T. Jaeger, U. Jenal, LadS is a calcium-responsive kinase that induces acute-to-chronic virulence switch in Pseudomonas aeruginosa. *Nat Microbiol* **2**, 16184 (2016).

60. C. Chen, G. A. Beattie, Pseudomonas syringae BetT is a low-affinity choline transporter that is responsible for superior osmoprotection by choline over glycine betaine. *J. Bacteriol.* **190**, 2717–2725 (2008).

61. I. Santi, P. Manfredi, E. Maffei, A. Egli, U. Jenal, Evolution of Antibiotic Tolerance Shapes Resistance Development in Chronic Pseudomonas aeruginosa Infections. *MBio* **12** (2021).

62. R. Reynolds, R. Hope, L. Williams, BSAC Working Parties on Resistance Surveillance, Survey, laboratory and statistical methods for the BSAC Resistance Surveillance Programmes. *J. Antimicrob. Chemother.* **62 Suppl 2**, ii15-28 (2008).

63. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *arXiv [q-bio.GN]* (2013). http://arxiv.org/abs/1303.3997.

64. M. Hunt, A. E. Mather, L. Sánchez-Busó, A. J. Page, J. Parkhill, J. A. Keane, S. R. Harris, ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom* **3**, e000131 (2017).

65. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

66. A. J. Page, B. Taylor, A. J. Delaney, J. Soares, T. Seemann, J. A. Keane, S. R. Harris, SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**, e000056 (2016).

67. N. J. Croucher, A. J. Page, T. R. Connor, A. J. Delaney, J. A. Keane, S. D. Bentley, J. Parkhill, S. R. Harris, Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).

68. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, vew007 (2016).

69. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

70. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

71. F. Menardo, S. Duchêne, D. Brites, S. Gagneux, The molecular clock of Mycobacterium tuberculosis. *PLoS Pathog.* **15**, e1008067 (2019).

72. J. Parker, A. Rambaut, O. G. Pybus, Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect. Genet. Evol.* **8**, 239–246 (2008).

73. F. Bielejec, A. Rambaut, M. A. Suchard, P. Lemey, SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics* **27**, 2910–2912 (2011).

74. J. Huerta-Cepas, K. Forslund, L. P. Coelho, D. Szklarczyk, L. J. Jensen, C. von Mering, P. Bork, Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).

75. J. M. Bryant, D. M. Grogono, D. Rodriguez-Rincon, I. Everall, K. P. Brown, P. Moreno, D. Verma, E. Hill, J. Drijkoningen, P. Gilligan, C. R. Esther, P. G. Noone, O. Giddings, S. C. Bell, R. Thomson, C. E. Wainwright, C. Coulter, S. Pandey, M. E. Wood, R. E. Stockwell, K. A. Ramsay, L. J. Sherrard, T. J. Kidd, N. Jabbour, G. R. Johnson, L. D. Knibbs, L. Morawska, P. D. Sly, A. Jones, D. Bilton, I. Laurenson, M. Ruddy, S. Bourke, I. C. Bowler, S. J. Chapman, A. Clayton, M. Cullen, T. Daniels, O. Dempsey, M. Denton, M. Desai, R. J. Drew, F. Edenborough, J. Evans, J. Folb, H. Humphrey, B. Isalska, S. Jensen-Fangel, B. Jönsson, A. M. Jones, T. L. Katzenstein, T. Lillebaek, G. MacGregor, S. Mayell, M. Millar, D. Modha, E. F. Nash, C. O'Brien, D. O'Brien, C. Ohri, C. S. Pao, D. Peckham, F. Perrin, A. Perry, T. Pressler, L. Prtak, T. Qvist, A. Robb, H. Rodgers, K. Schaffer, N. Shafi, J. van Ingen, M. Walshaw, D. Watson, N. West, J. Whitehouse, C. S. Haworth, S. R. Harris, D. Ordway, J. Parkhill, R. A. Floto, Emergence and spread of a human-transmissible multidrug-resistant nontuberculous mycobacterium. *Science* **354**, 751–757 (2016).

76. T. Mäklin, T. Kallonen, S. David, C. J. Boinett, B. Pascoe, G. Méric, D. M. Aanensen, E. J. Feil, S. Baker, J. Parkhill, S. K. Sheppard, J. Corander, A. Honkela, High-resolution sweep metagenomics using fast probabilistic inference. *Wellcome Open Res* **5**, 14 (2020).

77. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

78. M. Love, S. Anders, W. Huber, Differential analysis of count data--the DESeq2 package. *Genome Biol.* **15**, 550 (2014).

79. A. Bernut, C. Dupont, N. V. Ogryzko, A. Neyret, J.-L. Herrmann, R. A. Floto, S. A. Renshaw, L. Kremer, CFTR Protects against Mycobacterium abscessus Infection by Fine-Tuning Host Oxidative Defenses. *Cell Rep.* **26**, 1828-1840.e4 (2019).

80. C. Belon, C. Soscia, A. Bernut, A. Laubier, S. Bleves, A.-B. Blanc-Potard, A Macrophage Subversion Factor Is Shared by Intracellular and Extracellular Pathogens. *PLoS Pathog.* **11**, e1004969 (2015).

81. V. Le Moigne, A.-L. Roux, H. Mahoudo, G. Christien, A. Ferroni, O. Dumitrescu, G. Lina, J.-P. Bouchara, P. Plésiat, J.-L. Gaillard, S. Canaan, G. Héry-Arnaud, J.-L. Herrmann, Serological biomarkers for the diagnosis of Mycobacterium abscessus infections in cystic fibrosis patients. *J. Cyst. Fibros.* **21**, 353–360 (2022).

82. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* **4**, vex042 (2018).

694 83. A. Weimann, C. Ruis, PhyloEffects, Zenodo (2024);
695 https://doi.org/10.5281/zenodo.10606175.

696 84. E. F. Pettersen, T. D. Goddard, C. C. Huang, UCSF Chimera—a visualization system for
697 exploratory research and analysis. *Journal of* (2004).

698 85. B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, C. H. Wu, UniRef: comprehensive
699 and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).

700 86. K. Held, E. Ramage, M. Jacobs, L. Gallagher, C. Manoil, Sequence-verified two-allele
701 transposon mutant library for Pseudomonas aeruginosa PAO1. *J. Bacteriol.* **194**, 6387–
702 6389 (2012).

703 87. A. Weimann, A. Dinan, Computer codes for Weimann et al - Evolution and host-specific
704 pathoadaptation of Pseudomonas aeruginosa, Zenodo (2024);
705 https://doi.org/10.5281/ZENODO.10625500.

706 88. A. Weimann, Data for Weimann et al - Evolution and host-specific pathoadaptation of
707 Pseudomonas aeruginosa, Zenodo (2024); https://doi.org/10.5281/ZENODO.10600286.

708 89. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K.
709 Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis,
710 S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene
711 Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

712 90. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K.
713 Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis
714 Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.
715 *Genome Res.* **20**, 1297–1303 (2010).

716 91. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping
717 and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**,
718 2987–2993 (2011).

719 92. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham,
720 T. Keane, S. A. McCarthy, R. M. Davies, H. Li, Twelve years of SAMtools and BCFtools.
721 *Gigascience* **10** (2021).

722 93. K. A. Jolley, J. E. Bray, M. C. J. Maiden, Open-access bacterial population genomics:
723 BIGSdb software, the Pubmlst.org website and their applications. *Wellcome Open Res*
724 **3**, 124 (2018).

725 94. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, Ggtree : An r package for visualization
726 and annotation of phylogenetic trees with their covariates and other associated data.
727 *Methods Ecol. Evol.* **8**, 28–36 (2017).

728 95. A. J. Page, N. De Silva, M. Hunt, M. A. Quail, J. Parkhill, S. R. Harris, T. D. Otto, J. A.
729 Keane, Robust high-throughput prokaryote de novo assembly and improvement pipeline
730 for Illumina data. *Microb Genom* **2**, e000083 (2016).

731 96. D. R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de
732 Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).

733 97. A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin,
734 S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G.

Tesler, M. A. Alekseyev, P. A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

98. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

99. C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, J. Huerta-Cepas, eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.*, doi: 10.1093/molbev/msab293 (2021).

100. H. C. Valley, K. M. Bukis, A. Bell, Y. Cheng, E. Wong, N. J. Jordan, N. E. Allaire, A. Sivachenko, F. Liang, H. Bihler, P. J. Thomas, J. Mahiou, M. Mense, Isogenic cell models of cystic fibrosis-causing variants in natively expressing pulmonary epithelial cells. *J. Cyst. Fibros.* **18**, 476–483 (2019).

101. D. H. Figurski, D. R. Helinski, Replication of an origin-containing derivative of plasmid RK2 dependent on a plasmid function provided in trans. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 1648–1652 (1979).

102. D. E. Nivens, D. E. Ohman, J. Williams, M. J. Franklin, Role of alginate and its O acetylation in formation of Pseudomonas aeruginosa microcolonies and biofilms. *J. Bacteriol.* **183**, 1047–1057 (2001).

103. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

104. B. C. Louden, D. Haarmann, A. M. Lynne, Use of Blue Agar CAS Assay for Siderophore Detection. *J. Microbiol. Biol. Educ.* **12**, 51–53 (2011).

105. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

Acknowledgements

We would like to thank Sanger Pathogen Informatics for providing informatics support, Humphrey Gardner and Robert E McLaughlin for sharing their unpublished short-read sequencing data of their previously published genomes (*18*), and David Aanensen and Susanne Häußler for sharing their genomic and transcriptomic data in the public domain (*23*, *25*). We would like to thank BSAC, and the laboratory staff at the clinical microbiology lab at Cambridge Unversity Hospitals for their support in isolating the hospital strains. We are grateful to Dr Anne-Béatrice Blanc-Potard for VLM support and the gift of pMF230. We thank Dr Mai Nguyen-Chi for support with access to animal experimentation, and Pr Stephen Renshaw for support and helpful discussions. We acknowledge the ZEFIX-LPHI (ZEbraFIsh and Xenopus platform, Lphi, University of Montpellier) and the Bateson Center (University of Sheffield, Sheffield) Aquarium teams for zebrafish maintenance and care. We wish to thank the imaging facility MRI (Montpellier Ressources Imagerie), and Elodie Jublanc and Vicki Diakou for their assistance. We also thank the University of Montpellier and the University of Sheffield for support.

Funding: This work was supported by: The Wellcome Trust (107032AIA (RAF, AW, CR, AD), 226602/Z/22/Z (RAF, AW); 10224/Z/15/Z JMB, 098051(JP), 220472/Z/20/Z SK); The Botnar Foundation (6063; RAF, AW, TLB, CR, JP); The UK Cystic Fibrosis Trust (Innovation Hub grant 001 (RAF, TLB, JP, SB, AW, CR, LE), NIHR Cambridge Biomedical Research Centre (RAF, KPB), Health Enterprise East (SBRI; KPB, JR, RAF), Royal Papworth Hospital Research Innovation Fund (RAF), the Canadian Institute of Health Research (168006), Cystic Fibrosis Canada (2610), Genome Quebec and Genome Canada (RCL), the Italian Ministry of University and Research (art. 1, commi 314-337 Legge 232/2016; GR, LL), the Italian Cystic Fibrosis Research Foundation (FFC 10/2023; GR), Horizon 2020 Research (H2020-MSCA-IF-2016; AB), the Marie Skłodowska-Curie Innovation Framework Program (CFZEBRA (751977); AB), the FC3R through the ZFishforCFCare project (22FC3R-017; AB) and the French cystic fibrosis foundation Vaincre La Mucoviscidose (VLM, RF20220503060; SP).

Author contributions: AW, JP and RAF conceived the project. AW, RCL, JP and RAF designed the experiments and wrote the manuscript. AW performed the bioinformatic analyses. AD and AW performed the differential expression analysis. AW and CR performed the Bayesian phylodynamic analysis. APP and TLB performed the computational structural modelling. LE and MW performed the virulence phenotyping assays. LS and PTH made the THP1 F508 knockin cell line. KB, JMB, SK, and AW performed the THP1 *in vitro* screen, sequencing and analysis. EU, KB, JR, JMB and RAF undertook the TeleCF study, samples collection, DNA extraction, and analysis. BB, KJ, RB and SJP conducted the bacteraemia sample collection and processing. GR and LL generated knockout bacterial lines and contributed to experimental design. AB and SP conceived and designed experiments and analysed data for the *in vivo* testing in zebrafish models. SP generated fluorescent *P. aeruginosa* strains. NRT provided feedback on the project outline and access to the Sanger computing facilities. JG, JF and RL provided clinical isolates and contributed to analysis. RAF and JP provided supervisory support**. Competing interests:** none. **Ethical approvals:** National Ethical approval for the TeleCF study (ClinicalTrials.gov number NCT01877707). was granted from the National Research Ethics Service (NRES) Committee of Hertfordshire, UK (REC 12/EE/0462). The study design was peer reviewed by the Cambridge Centre for Lung Infection (CCLI) and the Cambridge Institute for Medical Research (CIMR). All zebrafish (Danio rerio) procedures described in the present study were performed by authorized staff and conducted by following the 3Rs -Replacement, Reduction and Refinement- principles in compliance with the European Union guidelines for handling of laboratory animals to local standards set approved by the UK Home Office under Animal Welfare and Ethical Review Body, and by the Direction Sanitaire et Vétérinaire de l'Hérault et Comité d'Ethique pour l'Expérimentation Animale de la région Languedoc Roussillon and the French Ministry of Agriculture (authorization number: APAFIS #36309-2022040114222432). Breeding and maintenance of adult zebrafish were performed at the ZEFIX (Lphi, UMR5294, University of Montpellier, France) and the Bateson Centre (University of Sheffield, UK) Home Office-approved aquaria, according to the local animal welfare standards (license numbers CEEA-LR-B3417237 and P1A4A7A5E). The number of animals used for each procedure was guided by pilot experiments.

Data and Materials availability: Short-read DNA and RNA sequence data for the clinical isolates was downloaded from the European Nucleotide Archive (ENA). THP1 pooled infection assay DNA sequencing data was uploaded to the ENA under PRJEB20836 (Supplementary Table S4). Short-read DNA sequencing data for newly sequenced isolates genomes (or where only assemblies were previously available) was uploaded to the ENA (TeleCF: ERP022089, IPC: PRJNA325248, AZ: PRJEB66158, LES: PRJEB69223, Bronchiectasis UK PRJEB69219). ENA run accessions can be found in Supplementary Table S1. The analysis codes and the PhyloEffects software were made available on GitHub and snapshots provided on Zenodo (*83*, *87*). Intermediate and additional data were made available in a Zenodo repository (*88*).

Figure legends

Figure 1. The emergence of epidemic clones of *Pseudomonas aeruginosa*. (**A**) Geographical location of the whole genome sequenced *P. aeruginosa* isolates obtained from patients, animals, and environment analysed in this study (*n* = 9,573). Number of samples from each location indicated by the size of blue dot. (**B**) Cumulative number of isolates across *P. aeruginosa* clones (defined by clustering genomes using the unweighted pair group method with arithmetic means; see *Supplementary Methods*), arranged by ascending number of genomes per clone and stratified into epidemic (*n* ≥30 isolates/clone; *red*), non-unique (1 < *n* < 30 isolates/clone; *light brown*), and unique (*n* = 1 isolate/clone; *blue*) groups. (**C**) *Left:* Maximum likelihood phylogenetic tree generated from genomes of all study isolates (major epidemic clones labelled in red). *Right* Bar plot representing the number of cities where each epidemic clone was found, coloured by continent. (**D**) Estimated date of first population expansion of 21 epidemic clones (predicted by Bayesian inference using *BEAST* (*29*)) with graph showing median and interquartile range (IQR; *boxplots*), 1.5 times IQR range (*whiskers*), and data points outside this range (*black points*). (**E**) Pangenome graph analysis of ancestral representatives of epidemic clones (*n* = 21) and sporadic clones (*n* = 80), constructed using *Panaroo* (*39*), where nodes represent clusters of orthologous genes and two nodes are connected by an edge if they are adjacent on a contig in any sample from the population, define gene gain events associated with the emergence of epidemic clones (described in detail in *Figure S5*) with genes highlighted that are involved in transcription (*blue*), defense mechanisms (*purple*), and inorganic ion transport and metabolism (*yellow*).For illustration purposes, the graph has been ordered against the genome of *P. aeruginosa* PAO1. *Inset:* magnified section of the pangenome graph is shown to illustrate node and edge structure.

Figure 2 Variable intrinsic host preference of epidemic *P. aeruginosa* clones. (**A**) Proportion of infections caused by epidemic clones (labelled by their majority multi-locus sequence type, ST) in cystic fibrosis (CF; *red*) and non-CF (*blue*) patients. (**B**) UMAP projection of transcriptomes from representative isolates of epidemic clones (*25*), colour-coded by the CF affinity of each clone. Expression data were pseudo-aligned to strain-specific gene indices to produce estimates of gene transcript abundance. (**C**) Transcriptome-wide association of gene expression with CF affinity. Transcript abundances were modelled as a response to the proportion of CF infections caused by each epidemic clone using a negative binomial generalised linear model. Volcano plot visualization of the Log2-fold expression change with CF proportion for every gene in the 99% core genome of *Pseudomonas aeruginosa* (center). Genes with an adjusted p-value of less than 0.05 and a log2 fold change less than -0.5 were coloured in green, genes with a log2 fold change greater than 0.5 were coloured in red. The coefficients for gene models were assessed using the Wald test (FDR = 0.05). Normalized expression counts vs CF proportions per epidemic clone with a trendline for the two genes with the lowest and highest log2 fold change, respectively, are shown above (top left/top right). Bulk RNA seq data was analysed from 241 clinical isolates of epidemic clones (*25*) included in our strain collection. (**D**) Survival of epidemic clones within wildtype (WT) or isogenic F508del knock-in THP1 macrophages at 2 and 4h post infection, expressed as fold change from 1 hour post infection showing median and interquartile range (IQR; *boxplots*), 1.5 times IQR range (*whiskers*. Experiments (carried out at least in duplicate) were performed by exposing THP1 macrophages to pooled isolates of 51 clinical isolates at a multiplicity of infection (MOI) of less than 1. Viable bacteria were isolated from macrophages at time points indicated and grown on solid media. Isolate abundance was quantified using sequence-based deconvolution. Strains with less than 1% abundance at the 1h time point were excluded from the analysis. A difference in the abundance of ST27 strains vs ST111 and ST235 strains at the 4h timepoint was assessed using a two-tailed t-test. * p-value < 0.05, ** p-value < 0.01.

Figure 3 Activation of the DksA1 regulon contributes to Cystic Fibrosis host preference of *P. aeruginosa* clones. (A) Volcano plot visualisation of the Log2-fold expression change with CF proportion for genes positively controlled (*red*) and negatively controlled (*green*) within

the DskA1 regulon as defined by Fortuna et al. (*44*). Bulk RNA seq data was analysed from 241 clinical isolates of epidemic clones *(24)* included in our strain collection. (**B**) DksA1 promotes survival of *P. aeruginosa* within CF macrophages. Viable intracellular *P. aeruginosa* (quantified through enumeration of cell-associated colony forming units; CFU) were measured at 1h and 4h post infection of differentiated wildtype (WT) and isogenic F508del homozygous knockin (CF-F508del) THP1 cells with wildtype (*blue*), isogenic DskA1-DskA2 double knockout (ΔDksA1,2; *pink*), and knockout complemented with DksA1 ((ΔDksA1,2::DksA1; *yellow*) *P. aeruginosa* PAO1. Data (mean ± SEM) are representative of at least three independent experiments performed in at least triplicate. *** p < 0.001; *ns* not significant (two-tailed Student's t-test). (**C**) (B) *Top*: Cartoon of zebrafish (created with *BioRender.com*) illustrating injection site for GFP-labelled fluorescent *P. aeruginosa*. *Bottom*: Representative fluorescence and DIC images of whole infected zebrafish larvae at 1 day post-infection (Scale bar: 150 μm; the labelled yolk sac is autofluorescent). (**D**) Survival analysis of control (*top*) and *cftr* morphant (cftr MO; *bottom*) zebrafish larvae infected intravenously (250-350 CFU) with *P. aeruginosa* PAO1 wildtype (*blue*), ΔDksA1,2 knockout (*pink*), and ΔDksA1,2::DksA1 complemented (*yellow*) fluorescent strains plotted as the percentage of surviving animals over 6 days (average of 2 independent experiments; n = 30-38 fish for each condition); *** $p < 0.001$ (Mantel-Cox Log-rank test). (**E**) Viable *P. aeruginosa* in zebrafish larvae at Day 1 post infection with *P. aeruginosa* PAO1 wildtype (*blue*), ΔDksA1,2 knockout (*pink*), and ΔDksA1,2::DksA1 complemented (*yellow*) fluorescent strains (plotted as mean ± IQR colony forming units (CFU) per fish of at least 3 independent experiments; n = 15-20 larvae per condition. *** $p < 0.001$; *ns* not significant (two-way ANOVA with Tukey's post-test). (**F,G**) Control and *cftr* morphant zebrafish larvae with mCherry-labelled macrophages (*Tg(mpeg1:mcherry-F)ump2 (45)*) were intramuscularly infected with 250-350 GFP-labelled *P. aeruginosa* PAO1 wildtype, ΔDksA1,2 or ΔDksA1,2::DksA1 strains) and the infection tracked using real-time intravital confocal microscopy. (**F**) Representative 3D reconstruction of confocal imaging showing macrophages (*red*) and automatic classification of extracellular (*grey*) and intracellular (*green*) *P. aeruginosa* (Scale bar 10 μm). (**G**) Quantification of the number of infected macrophages at the site of injection (*left*) and the level of intracellular bacterial load (calculated by the volume of bacteria-associated fluorescence observed within each macrophage) at 6 hours post infection with *P. aeruginosa* PAO1 wildtype (*blue*), ΔDksA1,2 knockout (*pink*), and ΔDksA1,2::DksA1 complemented (*yellow*) fluorescent strains. Mean ± IQR of at least 54 cells per condition (from n = 4-6 larvae) recorded from 2 independent experiments. ** $p < 0.01$; *** $p < 0.001$; *ns* not significant (two-way ANOVA with Tukey's post-test).

Figure 4 Host-specific pathoadaptation of *P. aeruginosa* (**A**) Manhattan plot showing nominal p values (plotted as -Log10) from genome-wide mutational burden test across all genes in *P. aeruginosa* PAO1. Significance was assessed using a Poisson test comparing the expected and observed number of mutations in each gene accounting for the proportion of genomes that gene was found in the pan-genome (FDR = 0.1; genes with a significant mutational burden, termed pathoadaptive, shown in *black,* others in *grey*). (**B**) UMAP projections of host adaptation of isolates (based on acquired mutations in pathoadaptive genes) colour-coded by (*left*) number of pathoadaptive mutations and (*right*) type of infection (centroids denoted by larger dots). Isolates without any pathoadaptive mutations were removed from the analysis. (**C**) Protein-protein interaction network for the pathoadaptive genes extracted from the STRING database (only main connected component shown, full graphs shown in *Fig. S12;* (*56*)). Genes are shown as nodes which are connected by an edge if they had an interaction reported in STRING (confidence > 0.7). *Top:* To estimate host-specific pathoadaption, the number of cystic fibrosis (CF) vs non-CF mutations (determined by stratifying mutations in pathoadaptive genes on terminal branches by the infection type of isolates) were compared using a Fisher exact test (FDR = 0.1) and expressed as an odds ratio for each gene. *Bottom:* Gene nodes were colour-coded by class of functional annotation (based on overrepresented pathways using Gene Ontology (*89*) biological process enrichment

analysis with TopGO (*57*) among CF: transmembrane transport and fatty acid biosynthesis, and non-CF: transcriptional regulation and chemotaxis).

Figure 5. Evolutionary trajectories of *P. aeruginosa* during pathoadaptation. (**A**) Normalised frequency of mutations over evolutionary time in specific pathoadaptive genes. The trajectories of the 50 most commonly mutated genes were manually assigned to one of 5 classes (*Figure S15*), based on the shape of their mutation frequency curves (relative size of each class and representative examples (with trendlines from locally-weighted smoothing) shown). (**B**) The relative transmissibility and host-specific adaptation of pathoadaptive genes was calculated. To estimate host-specific pathoadaptation, the number of cystic fibrosis (CF) vs non-CF mutations (determined by stratifying mutations in pathoadaptive genes on terminal branches by the infection type of isolates) were compared using a Fisher exact test (FDR = 0.1) and expressed as an odds ratio. To assess the transmissibility of pathoadaptive changes, the number of mutations that had been observed in at least two isolates were compared with mutations that had only been observed once using a Fisher exact test (FDR = 0.1). Genes were colour-coded if showing significant host-specific adaptation (*blue*), changes in transmissibility (*purple*), or both (*pink*). Genes with zero or infinite odds ratio not shown. (**C**) Functional annotation of pathoadaptive genes associated in (*top*) host-specific adaptation and (*bottom*) changes in transmissibility. (**D**) The number (*top*) and proportion (*bottom*) of transmission links across a range of pairwise SNP thresholds occurring between CF to CF (*red*), CF to non-CF (*yellow*), and non-CF to non-CF (*blue*) individuals (data were down-sized to contain equal numbers of CF and non-CF infections). (**E**) Transmission clusters involving patients with CF (*red*), non-CF (*blue*), or unknown status (*white*). Nodes representing isolates were connected by edges if pairwise SNP distances were 26 SNPs or less. This cut-off represents the 95th percentile of the within-host genetic diversity analysed in 81 patients.

Supplementary Materials

Materials and Methods
Figs. S1 to S16
Captions for Tables S1-7
References (90-105)

Figure 1.

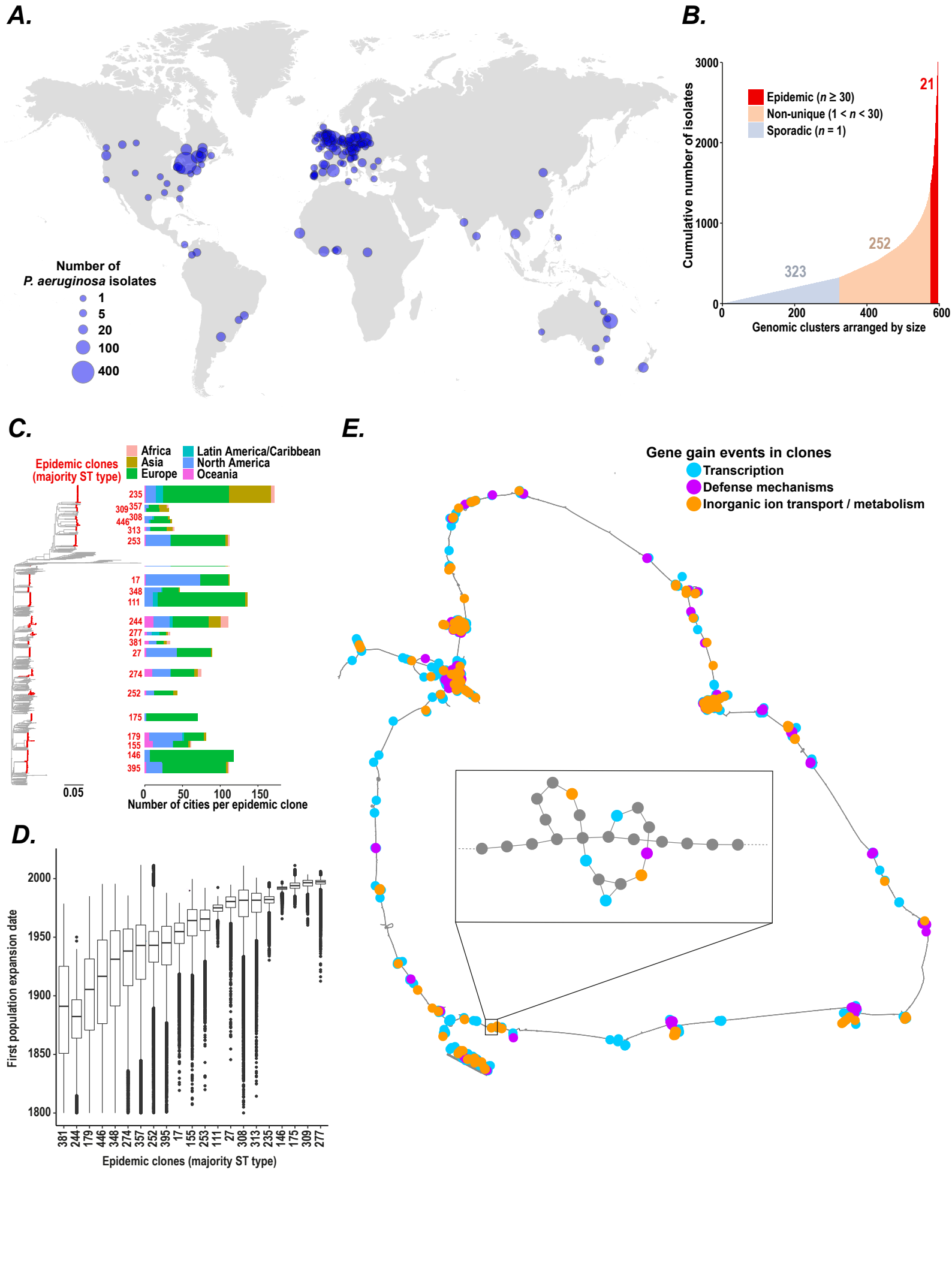

A.
Number of
P. aeruginosa isolates
1
5
20
100
400
B.
Cumulative number of isolates
3000
2000
1000
0
Epidemic (n ≥ 30)
Non-unique (1 < n < 30)
Sporadic (n = 1)
21
252
323
0
200
400
600
Genomic clusters arranged by size
C.
Epidemic clones
(majority ST type)
Africa
Asia
Europe
Latin America/Caribbean
North America
Oceania
235
309
357
446
308
313
253
17
348
111
244
277
381
27
274
252
175
179
155
146
395
0.05
0
50
100
150
Number of cities per epidemic clone
D.
First population expansion date
2000
1950
1900
1850
1800
381
244
179
446
348
274
357
252
395
17
155
253
111
27
308
313
235
146
175
309
277
Epidemic clones (majority ST type)
E.
Gene gain events in clones
Transcription
Defense mechanisms
Inorganic ion transport / metabolism

Figure 2.

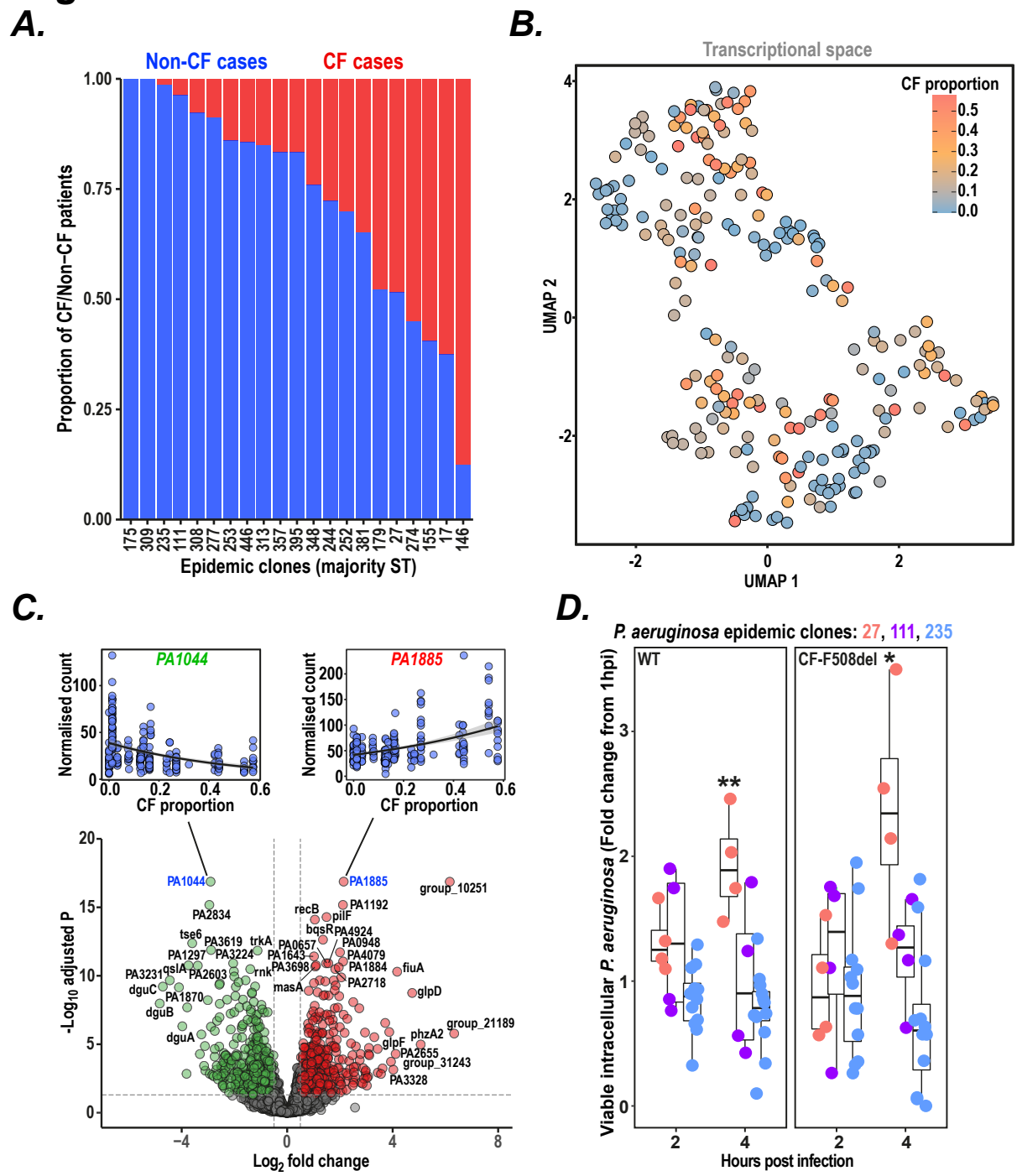

A.
Non-CF cases
CF cases
Proportion of CF/Non−CF patients
1.00
0.75
0.50
0.25
0.00
175 309 235 111 308 277 253 446 313 357 395 348 244 252 381 179 27 274 155 17 146
Epidemic clones (majority ST)
B.
Transcriptional space
CF proportion
0.5
0.4
0.3
0.2
0.1
0.0
UMAP 2
UMAP 1
C.
PA1044
PA1885
Normalised count
CF proportion
-Log10 adjusted P
Log2 fold change
PA1044
PA2834
tse6
PA3619
trkA
PA1297
PA3224
PA3231
qslA
PA2603
dguC
PA1870
dguB
dguA
PA0657
PA1643
PA3698
rnk
masA
recB
pilF
bqsR
PA4924
PA1192
PA0948
PA4079
PA1884
PA2718
fiuA
glpD
group_10251
group_21189
phzA2
glpF
PA2655
group_31243
PA3328
D.
P. aeruginosa epidemic clones: 27, 111, 235
WT
CF-F508del
**
*
Viable intracellular P. aeruginosa (Fold change from 1hpi)
Hours post infection

Figure 3

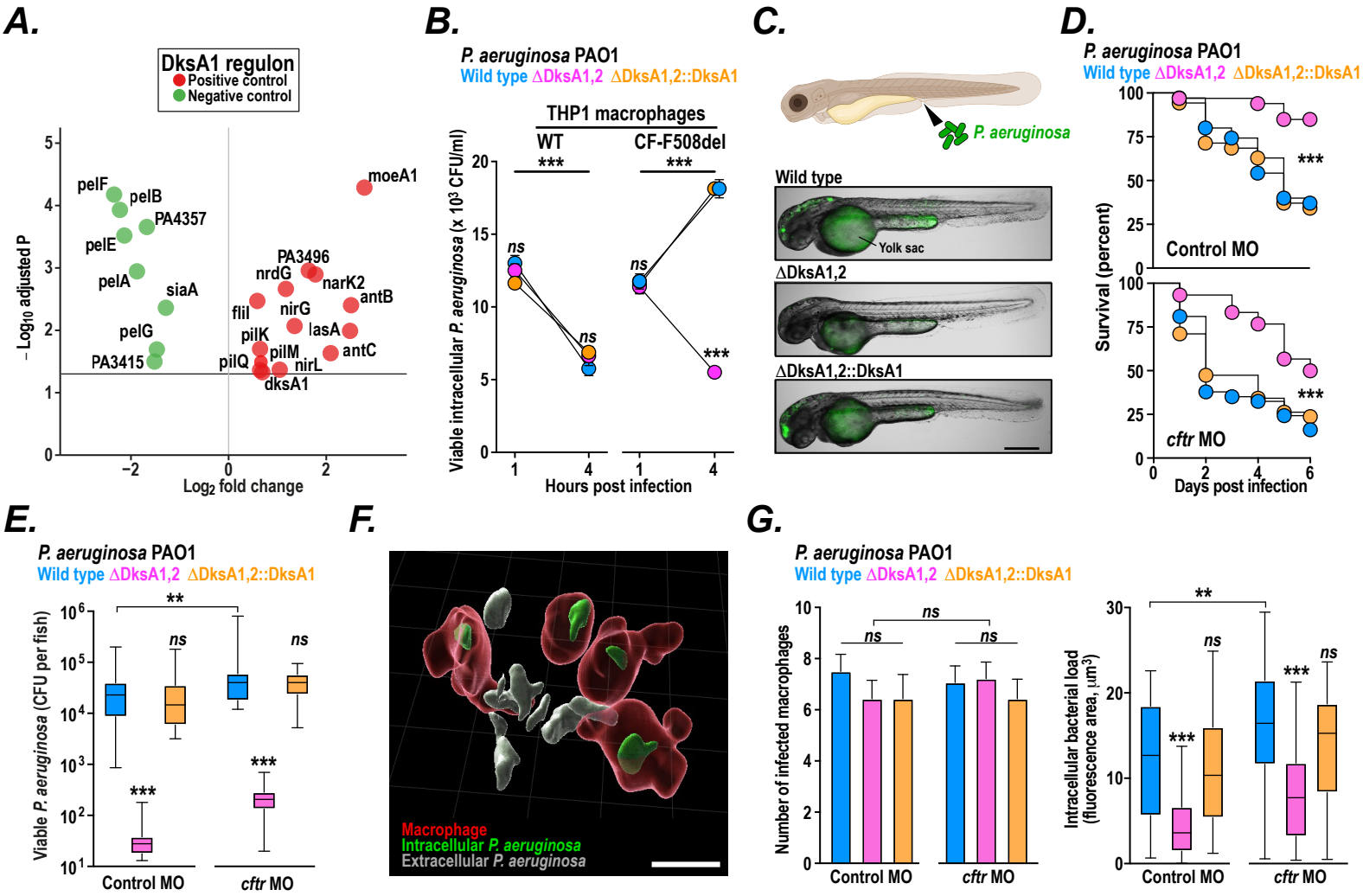

A.
DksA1 regulon
Positive control
Negative control
−Log10 adjusted P
Log2 fold change
pelF
pelB
PA4357
pelE
pelA
siaA
pelG
PA3415
moeA1
PA3496
nrdG
narK2
antB
fliI
nirG
pilK
lasA
antC
pilQ
pilM
nirL
dksA1
B.
P. aeruginosa PAO1
Wild type ΔDksA1,2 ΔDksA1,2::DksA1
THP1 macrophages
WT
CF-F508del
Viable intracellular P. aeruginosa (x 10³ CFU/ml)
Hours post infection
ns
***
C.
P. aeruginosa
Wild type
Yolk sac
ΔDksA1,2
ΔDksA1,2::DksA1
D.
P. aeruginosa PAO1
Wild type ΔDksA1,2 ΔDksA1,2::DksA1
Survival (percent)
Control MO
cftr MO
Days post infection
E.
P. aeruginosa PAO1
Wild type ΔDksA1,2 ΔDksA1,2::DksA1
Viable P. aeruginosa (CFU per fish)
Control MO
cftr MO
F.
Macrophage
Intracellular P. aeruginosa
Extracellular P. aeruginosa
G.
P. aeruginosa PAO1
Wild type ΔDksA1,2 ΔDksA1,2::DksA1
Number of infected macrophages
Intracellular bacterial load (fluorescence area, μm³)
Control MO
cftr MO

Figure 4.

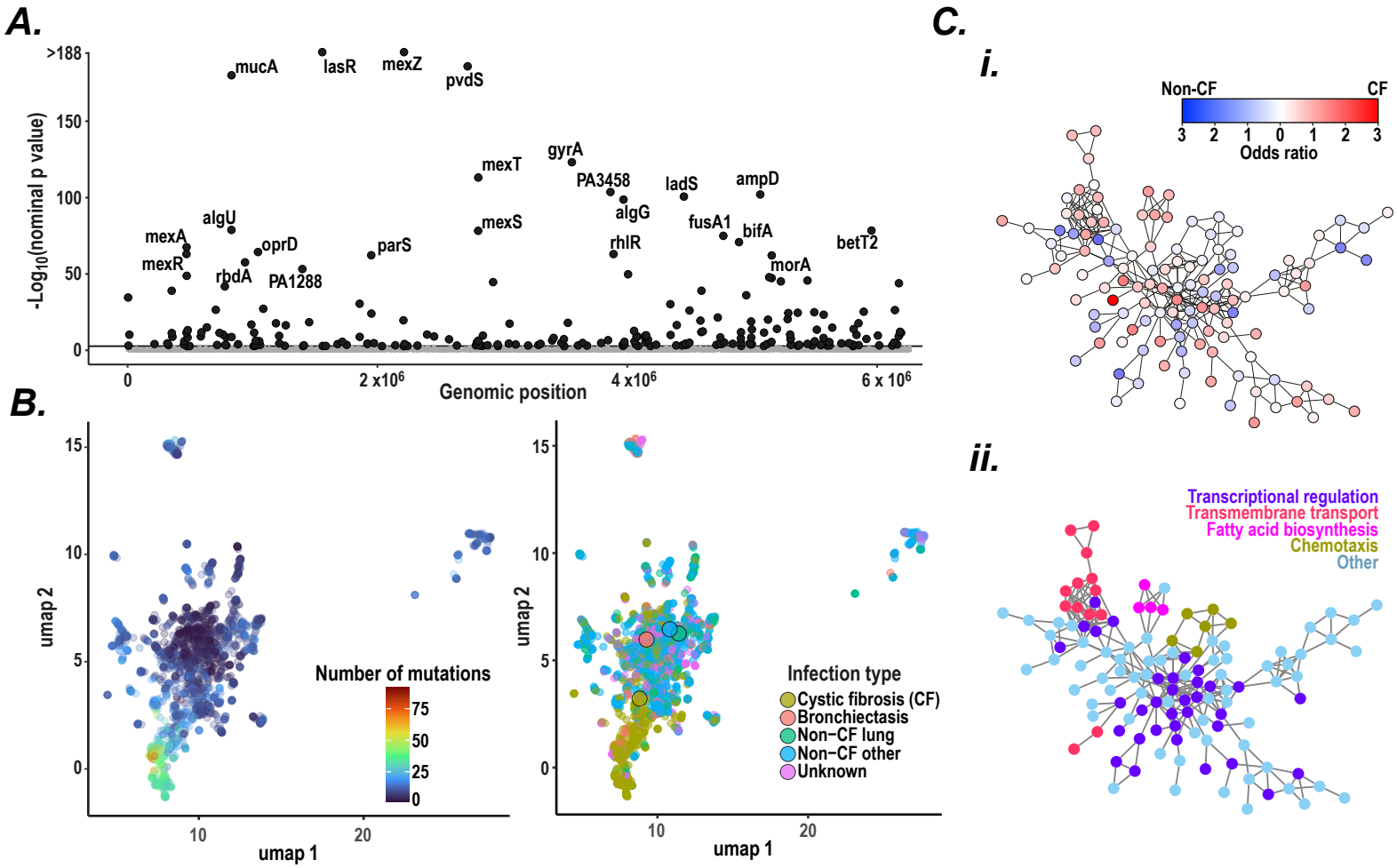
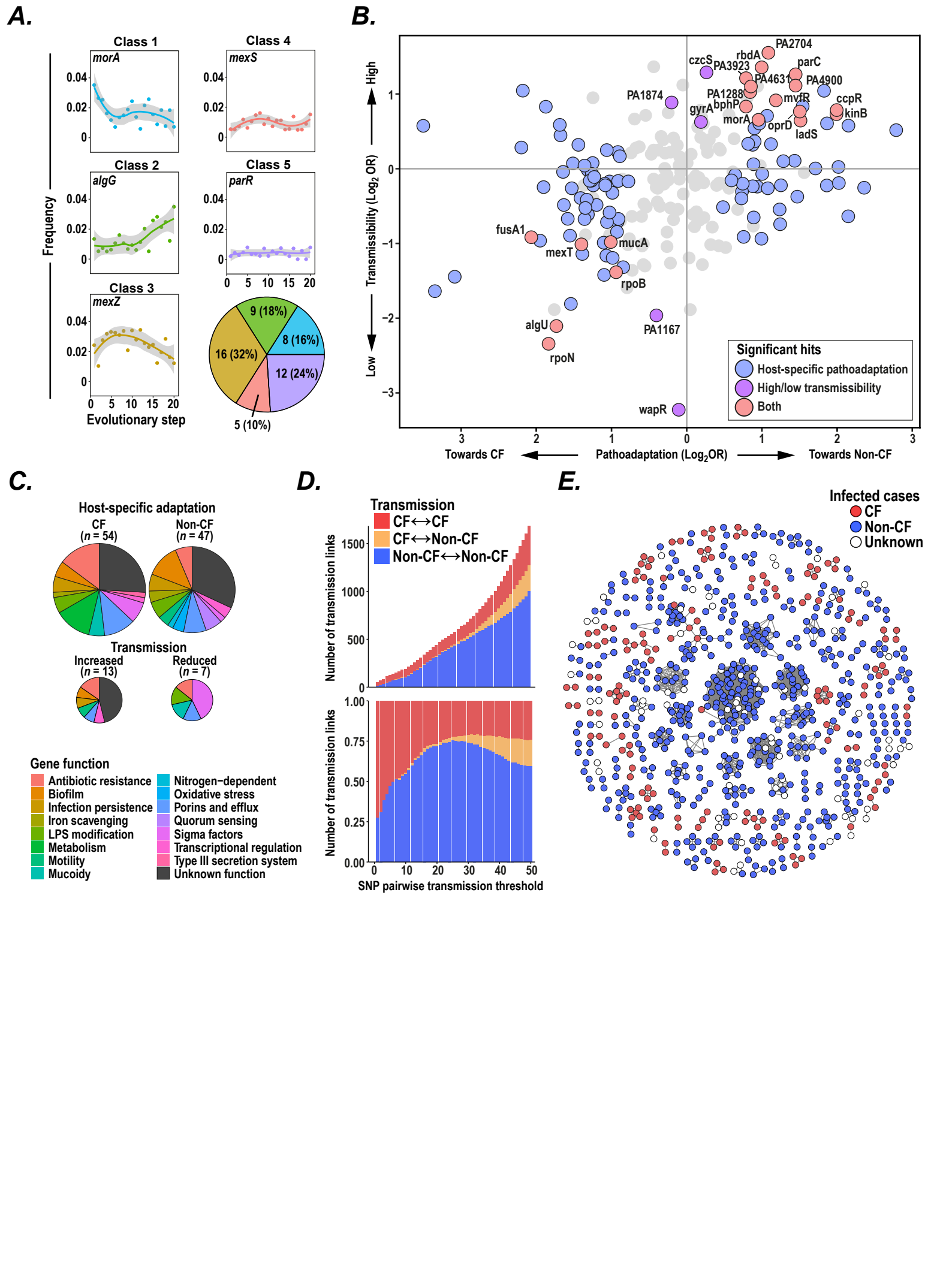

A.
>188
150
100
50
0
-Log10(nominal p value)
mucA
lasR
mexZ
pvdS
mexT
gyrA
PA3458
ladS
ampD
algU
algG
fusA1
bifA
mexA
oprD
parS
mexS
rhlR
betT2
mexR
rbdA
PA1288
morA
0
2 x10^6
4 x10^6
6 x 10^6
Genomic position
B.
15
10
5
0
umap 2
10
20
umap 1
Number of mutations
75
50
25
0
Infection type
Cystic fibrosis (CF)
Bronchiectasis
Non−CF lung
Non−CF other
Unknown
C.
i.
Non-CF
CF
3 2 1 0 1 2 3
Odds ratio
ii.
Transcriptional regulation
Transmembrane transport
Fatty acid biosynthesis
Chemotaxis
Other

Figure 5.


A.
Class 1
morA
Class 4
mexS
Class 2
algG
Class 5
parR
Class 3
mexZ
Frequency
Evolutionary step
9 (18%)
8 (16%)
16 (32%)
12 (24%)
5 (10%)
B.
Transmissibility (Log2 OR)
High
Low
Towards CF
Pathoadaptation (Log2OR)
Towards Non-CF
PA2704
rbdA
czcS
PA3923
parC
PA4631
PA4900
PA1288
mvfR
ccpR
PA1874
gyrA
bphP
kinB
morA
oprD
ladS
fusA1
mexT
mucA
rpoB
algU
PA1167
rpoN
wapR
Significant hits
Host-specific pathoadaptation
High/low transmissibility
Both
C.
Host-specific adaptation
CF
(n = 54)
Non-CF
(n = 47)
Transmission
Increased
(n = 13)
Reduced
(n = 7)
Gene function
Antibiotic resistance
Biofilm
Infection persistence
Iron scavenging
LPS modification
Metabolism
Motility
Mucoidy
Nitrogen−dependent
Oxidative stress
Porins and efflux
Quorum sensing
Sigma factors
Transcriptional regulation
Type III secretion system
Unknown function
D.
Transmission
CF↔CF
CF↔Non-CF
Non-CF↔Non-CF
Number of transmission links
SNP pairwise transmission threshold
E.
Infected cases
CF
Non-CF
Unknown

Science
AAAS

Supplementary Materials for

Evolution and host-specific adaptation of *Pseudomonas aeruginosa*

Aaron Weimann[1,2,3,4], Adam M Dinan[1,2,3], Christopher Ruis[1,2,3,4], Audrey Bernut[5], Stéphane Pont[5], Karen Brown[1,2,6], Judy Ryan[1,2], Lúcia Santos[7], Louise Ellison[2], Emem Ukor[2,6], Arun P. Pandurangan[1,8,9], Sina Krokowski[1,2], Tom L. Blundell[1,8,9], Martin Welch[8], Beth Blane[9], Kim Judge[11], Rachel Bousfield[9,10], Nicholas Brown[10], Josephine M. Bryant[11], Irena Kukavica-Ibrulj[12], Giordano Rampioni[13,14], Livia Leoni[13], Patrick T. Harrison[7], Sharon J Peacock[9,10], Nicholas R. Thomson[11], Jeff Gauthier[12], Jo L Fothergill[15], Roger C Levesque[12], Julian Parkhill[4], R. Andres Floto[1,2,3,6,9,10]

Corresponding author:
Andres Floto (arf27@cam.ac.uk),
Julian Parkhill (jp369@cam.ac.uk), or
Roger Levesque (rclevesq@ibis.ulaval.ca)

The PDF file includes:

Materials and Methods
Figs. S1 to S15
Captions for Tables S1 to S7
References (90-105)

Other Supplementary Materials for this manuscript include the following:

Supplementary Tables 1-7 (tsv)

MATERIAL AND METHODS

Genomic datasets

Published datasets: We utilized the following *Pseudomonas aeruginosa* whole genome sequencing datasets: from studies of antibiotic resistance in healthcare setting in Germany (*1*), Spain (*2*), the Philippines (*3*), and world-wide (*4*); from studies of infection in individuals with Cystic Fibrosis (CF) from Denmark (*5*) and Canada (*6*) and with non-CF bronchiectasis from the UK(*7*); from the International *Pseudomonas* Consortium (*8*); and from studies targeting high-risk clones ST274(*9*), ST111(*62*), and the Liverpool Epidemic Strain (LES)(*63*).

Unpublished datasets: P. aeruginosa isolates from the TeleCF study: TeleCF (NCT01877707) was a single-centre, observational pilot study of home monitoring in adults with CF (*n* = 9) who were chronically infected with *P. aeruginosa* and had experienced at least two acute pulmonary exacerbations in the prior 12 months. Participants collected spontaneously expectorated sputum daily for 6 months in dedicated study freezers provided in their homes. Samples were collected and then thawed to room temperature, treated with an equal volume of 0.1% dithiothreitol, vortexed, incubated at room temperature for 15 minutes, and then vortexed again to ensure homogeneity before being streaked to purity and grown on *Pseudomonas* selective agar base prepared plates with cetrimide and sodium nalidixate supplement (PCN agar; Oxoid). Plates were incubated at 37°C for 48-72 h, to allow for the growth of slow-replicating and small colony variants. Representative colonies were placed into a 96-deep well microplate containing 1 mL of cetrimide broth, and grown for 6 hours at 37°C. These strains were then re-streaked for purity and grown on PCN plates at 37°C for 48-72 hours. Single colonies were chosen and re-arrayed into individual 96-deep well microplates, one microplate per initial sputum sample, and left to undergo static incubation at 37°C for a further 6 hours. To ensure minimal adaptation to the laboratory environment, the strains were stored at -80°C in 25% (v/v) glycerol solution for long-term preservation. DNA was extracted from the strains using a QIAxtractor (QIAgen) instrument, following the manufacturer's instructions. The Illumina protocol was used for library preparation, and sequencing was conducted on both the Illumina HiSeq 2500 and X10 platforms.

P. aeruginosa isolates from bacteraemia infections (n = 365): 224 isolates were cultured from bloodstream infections from patients attending Addenbrookes, Hinchingbrooke, and Papworth Hospitals (UK) between 2006 and 2013 with a further 60 isolates collected between 2017 and 2018. 81 isolates were collected as part of the BSAC bacteraemia resistance programme (*89*) from 25 contributing laboratories distributed across the UK and Ireland between 2001 and 2011 focusing on multi-drug resistant strains. DNA was extracted using QIAxtractor (QIAgen), according to the manufacturer's instructions. Library preparation was done according to the Illumina protocol, and samples were sequenced on the Illumina HiSeq 2000 and 2500 platforms.

Variant calling

Sequencing reads from all samples analysed were mapped against the *P. aeruginosa* PAO1 reference genome (accession number AE004091.2) using the multiple_mappings_to_bam 1.6 pipeline with default parameters (https://github.com/sanger-pathogens/bact-gen-scripts) with BWA (*90*) as the short-read aligner including an indel realignment step using GATK (*91*). Samples where less than 70% of the reads mapped to the reference genome were removed.

SNPs and indels were identified from the aligned reads using the same pipeline employing samtools mpileup (*92*) for generating read pile-ups and bcftools call (*92*) for variant identification. Variants were filtered by base call quality (>=50), mapping quality (>=20) and number of supporting reads (>=8) on the reverse strand (>=3) and the forward strand (>=3). Indels across strains were aggregated into a catalogue and the base at the start of every indel in the catalogue was quality checked as above.

We removed samples with an excess number of minority variants as they likely indicate strain-level contamination. Variants were filtered to identify minority variants using a previously described approach (*93*). Subsequently every sample with more than 8 minority variants was removed. Ariba 2.14.6 (*64*) was used to query the genomes using the multi-locus sequencing type scheme for *P. aeruginosa* (*28*) for the allele-sequence type combinations as available on pubMLST (*94*) (downloaded on Nov 11 2020). Ariba also identifies new allelic combinations. Samples where the sequence type could not be determined with Ariba were removed (n = 7). The oldest sample was chosen to represent every patient sequence type combination to minimise bias from within-host evolution. SNP-sites (*66*) was used to infer an alignment of the variable sites. FastTree (2.1.10) was used to infer a global phylogenetic tree (*67*). Ggtree was used to visualise the trees and produce figures (*69*).

Clone assignment

The sample with the earliest collection date in each sequence type was chosen to represent every patient-sequence type combination to minimise bias from within-host evolution. PairSNP (Version 0.2.0) (https://github.com/gtonkinhill/pairsnp) was used to infer all pairwise SNP distances between sample pairs. The ultra-metric pairwise group method with arithmetic means (UPGMA) was used to cluster samples based on pairwise SNP distances. UPGMA infers a sample dendrogram which is then separated into clusters by applying a discrete SNP distance threshold. A global threshold of 7000 SNPs was chosen to separate the dendrogram into genomic clusters (which we refer to as clones). This threshold was selected to assign all samples from any recently emerged clones into the same cluster while at the same time being liberal enough to allow recently emerged recombinants and hypermutators to be clustered within their ancestral clones (*Supplementary Figure 1*). Samples with the same patient-sequence type combinations that were not included in the clustering were then assigned to the matching clone type.

SNP-sites was used to infer a clone-specific alignment of variable sites (*69*). Gubbins version 2.4.1 (*68*) was used to remove recombination for individual clones with at least four available genomes. Gubbins infers a phylogenetic tree with RaxML (*69*) based on the final alignment.

Dating and phylogeography

Molecular dating was performed for all 21 epidemic clones separately including genome sequences with known collection dates and including only the earliest sample from each patient. To identify potential hypermutator branches that would violate the molecular clock assumption, the ratio of transition and transversion mutations was compared using a Fisher exact test. P-values were corrected using the Benjamini-Hochberg procedure at an FDR threshold of 5%. We initially assessed the temporal signal in each epidemic clone. We reconstructed a non-dated phylogenetic tree using RAxML 8.2.12 (*70*) with the general time reversible (GTR) model of nucleotide substitution and gamma rate heterogeneity with four gamma classes and compared sequence collection dates with root-to-tip distances using TempEst (*71*). Sequences that had accumulated far more or fewer mutations than expected given their collection date were removed from further analyses and a new phylogenetic tree reconstructed as above. To assess the significance of the temporal signal in each epidemic clone, we compared the correlation coefficient between collection date and root-to-tip distance with the real sequence dates with the distribution of correlation coefficients from 1000 date randomisations using a custom script (https://github.com/chrisruis/tree_scripts/blob/main/bootstrap_TempEst_rttd_date.R). Clones with a significant temporal signal in this test (P < 0.05) were taken forward for molecular dating with BEAST 2.6.6 (*72*). We employed the HKY model of nucleotide substitution. We used an uncorrelated relaxed lognormal clock model with a lognormal prior on the mean substitution rate with mean set to the slope of the root-to-tip correlation calculated above and standard deviation set to 0.5. The variation in substitution rate across branches was modelled using a gamma prior

with alpha set to 0.54 and beta set to 0.38. We modelled the population history using the coalescent Bayesian skyline population prior. Three independent runs were conducted for each dataset with 100M steps; convergence was assessed with Tracer 1.7.1 (*72*) with 10% burn-in. For clones that didn't pass the bootstrap randomisation test (N = 9), a uniform prior for the substitution rate was set encompassing values across the 95% HPD estimates for the substitution rates inferred for the above clones. All other priors and parameters were kept the same.

To further establish a temporal signal in epidemic clone that passed the root-to-tip randomization test above, we ran a more thorough date randomisation test (*73*). Here, BEAST was run using a uniform prior with upper and lower bounds set to encompass the full range of substitution rates inferred previously for the clones above and all other priors set as before. The estimates from those BEAST runs were highly similar to the estimates obtained with the informed substitution rate prior. We then performed ten randomization runs where collection dates were randomized across sequences and BEAST run using the same uniform prior on substitution rate. The estimated median substitution rates and most recent common ancestor dates for the randomization runs did not overlap with those of the runs using real collection dates, indicating a significant temporal signal.

To test whether each epidemic clone has undergone a historical population expansion, we analysed Bayesian skyline plot estimates of relative genetic diversity across the posterior distribution. We inferred whether each sampled step in the MCMC chain exhibits an increase in relative genetic diversity of at least twofold relative to the root of the tree and examined the distribution of dates of this increase; median and 95% HPD estimates were calculated for each clone.

Prior to conducting Bayesian phylogeographic analyses, the association index was computed to find evidence of geographic clustering within clone phylogenies (*95*–*97*). Based on randomly permuting the locations 1,000 times, we identified clones for further spatiotemporal analysis where less than 5% of the randomisations had a higher association index than the non-permuted dataset. Samples from the same clone were subsampled to keep only one sequence for every cluster of genetically related samples from the same city to account for regional outbreaks. In addition, locations only containing one sample within a clone were removed from the analysis of that clone.

Asymmetric phylogeographic discrete trait reconstructions were performed using the BEAST classic 1.9.0 package of BEAST 2.6.6 (*38*). The continent was used as the label for each sample. The same priors were used as above including an informed log-normal substitution rate prior and Bayesian Skyline prior on the population size. Additionally, an exponential prior was employed for the rate of lineage movements with mean 1. The relative rates of migration between continent pairs were modelled with a gamma distribution with alpha and beta set to 1. To assess the robustness of our approach in the light of overrepresentation of certain continents, we subsampled the sequences of the most frequent continent so that the two most frequent continents were equally abundant. Subsampling was repeated five times and results compared between subsamples. Spread 0.9.7.1 (*97*) was used to identify migration routes between continents. We only reported routes that had a Bayes factor of at least 3 in both the full sample and in at least four out of five subsamples.

Pan-genome analysis

Genomes were assembled from short-read data using the Velvet or SPAdes assembler using the Assembly Improvement pipeline (*98*). Where several assemblies were available, the assembly with the fewest contigs was selected. Where the number of contigs exceeded 500 or less than 3000 genes were predicted (as determined by panaroo-qc), Shovill 1.1.0 (https://github.com/tseemann/shovill) was used to re-assemble the genome. Panaroo 1.2.8 with -

--clean-mode moderate (*99*) was used to cluster the gene sequences from all samples into gene families and infer a graphical pan-genome (for every patient-clone combinations the earliest sample was used).

Parsimony reconstruction was used to infer the presence or absence of each gene in the common ancestor at the root of each clone tree. We then picked the genome that was most similar to the ancestral genome in terms of gene presence and absence as the ancestral epidemic clone representative. The genome graph was then reduced to a random subset of the sporadic clones (N = 50) and epidemic clone representatives (N = 21). The graph was then ordered against the genome of *Pseudomonas aeruginosa PAO1* and any long-range connections (>100 genes distance) were cut as described previously (*100*). yFiles 1.1.1 (https://www.yworks.com/products/yfiles) was then used to infer a layout in Cytoscape 3.8.2 (*75*) for visualisation. An outgroup-rooted tree was inferred based on the clone representative genomes using FastTree (2.1.10) employing the PA7-type isolate strain AZPAE14941 as the outgroup. Using the reduced pan-genome, parsimony ancestral character state reconstruction was used to infer gene gains and losses on the branches of the rooted tree leading to the ancestral epidemic and sporadic clones. Co-gained genes were then aggregated into events based on the phylogenetic context. EggNog-mapper 2.1.6 (*76*) was used to annotate the gene family of each gene in each event using the representative sequence for the corresponding gene families from Panaroo.

A Fisher exact test was conducted to compare the number of genes gained and annotated with a specific COG category. Multiple-testing correction was applied to account for the number of tests (the number of COG categories) using the Benjamini-Hochberg method to control the false discovery rate at 10%.

THP1 F508del cell line

Isogenic F508del homozygous THP1 cells were created as follows: Alt-R® CRISPR-Cas9 (S.p. HiFi Cas9 Nuclease V3; crRNA; and tracrRNA), and ssODN HDR donor template (Ultramer® DNA Oligos) were purchased from Integrated DNA Technologies (IDT). Genotyping PCR primers were obtained from Eurofins Genomics.

Ribonucleoparticles (RNP) assembly and electroporation of the gene editing reagents were performed as previously described (*76*). Briefly, the 100uM stocks of crRNA and tracrRNA were combined at equimolar concentrations for a final duplex concentration of 44 µM, incubated at 95 ^{o}C for 5 min, and then allowed to cool down to room temperature. The crRNA:tracrRNA complex was combined at a 1:1.2 molar ratio with Cas9 nuclease protein and incubated for 20 min to form the RNP complex. THP-1 cells were electroporated using the Neon™ Transfection System (Thermo Fisher Scientific) and the Neon™ Transfection System 10 µl kit. For each electroporation reaction, 100,000 cells were resuspended in 5 µl of Buffer R, mixed with 7 µl of RNP complex, and electroporated with 300 ng ssODN HDR donor template. For the co-transfection of RNP and ssODN HDR donor template, ssODN HDR donor was added to the RNP:cells mixture before the electroporation step. The cells were seeded in a 24-well plate and 72 h after electroporation editing efficiency was evaluated.

After electroporation, cells were harvested for genomic DNA (gDNA) extraction using the DNeasy Blood and Tissue Kit (Qiagen). The target site was amplified using Q5 Hot Start High-Fidelity 2x Master Mix (New England Biolabs), PCR products were purified, and Sanger sequenced (Eurofins Genomics). The gene editing events were analysed using the Sanger sequencing files from unedited and edited cells as input into the Inference of CRISPR Edits (ICE) web tool (2019, v2.0. Synthego).

THP1 infection assay

Wild type (WT) and F508del THP-1 cells were cultured in RPMI-1640 (Invitrogen) medium supplemented with 10% heat-inactivated fetal bovine serum (Invitrogen), 100U/mL penicillin and 100ug/mL streptomycin (Sigma), in 5% CO_2 humidified atmosphere at 37 ^oC. THP-1 monocytes were seeded at 200,000 cells/mL in 24 well tissue culture plates (Corning), and differentiated into macrophages in the presence of 20 ng/mL phorbol 12-myristate 13-acetate (PMA, Sigma–Aldrich) for 48 hours followed by a recovery period of 24 hours in serum-supplemented RPMI-1640 medium without PMA (as previously described (*77*)). Cell differentiation was verified by detection of morphology changes using light microscopy.

Pooled clinical isolates of *P. aeruginosa* were cultured in low-salt LB (Thermo Fisher) at 37 ^oC with aeration at 200rpm overnight, resuspended in RPMI-1640 supplemented with serum and then added to differentiated WT or F508del THP-1 cells at a multiplicity of infection (MOI) of 1:1, centrifuged at 1800rpm for 3 minutes, and then incubated at 37 ^oC for 1 hour before the supernatant was removed and cells were lysed (using 2% Saponin) at 1h time point or incubated in fresh media for further time points (2h or 4h) before supernatant removal and cell lysis. DNA was extracted using the QiaAmp DNA mini kit (Qiagen).

Strain abundance was quantified using mSWEEP 1.4.0 sequence-based deconvolution method (*78*). First, we built an index using the 51 input genomes using the *build_index* command. The reads for every sample were then pseudoaligned to the index using the *pseudoalign* command. Finally, the abundances of the strains in our samples were estimated by running the *mSWEEP* command. Strains with less than 1% abundance at the 1h time point were excluded from the analysis. A difference in the abundance of ST27 strains vs ST111 and ST235 strains at the 4h relative to the 1h timepoint was assessed using a two-tailed t-test.

Mutant *ΔdksA1-2* and complemented *ΔdksA1-2::DksA1 P. aeruginosa* were cultured in Luria-Bertani medium (LB, Thermo Fisher) supplemented with 100ug/ml tetracycline, and PAO1 was incubated in LB only at 37 ^oC with aeration at 200rpm overnight.

Bacterial suspensions in RPMI-1640 supplemented with serum were added to the differentiated WT and F508del THP-1 cells at a multiplicity of infection (MOI) of 1:1, centrifuged at 1800rpm for 3 minutes and then incubated at 37 °C for 1 hour. THP-1 cells were lysed at timepoints 1 hour and 4 hours using 2% Saponin (Sigma) and colony forming units were calculated by plating the lysates on LB agar.

Transcriptomic analysis

Gene expression data for clinical *P. aeruginosa* strains (and the UCBPP- PA14 wildtype control strain) was obtained as described previously (*47*). Briefly, strains were grown in LB broth at 37 °C and harvested at late log phase (OD600 = 2). Sequencing was done on an Illumina HiSeq 2500. Expression data were pseudoaligned to strain-specific gene indices to produce abundance estimates using Kallisto (*78*) and orthologous genes present in 99% of strains, as determined by *Panaroo* as above, were retained for downstream analysis. Expression levels for missing genes were imputed with 0 values. Abundance estimates were scaled to the median length of each ortholog across strains. Length-scaled abundance estimates `were size-factor normalised by the median ratio method and modelled as a response to CF proportion per sequence type using a negative binomial generalised linear model (GLM) with DESeq2 (*79*). The coefficients for gene models were assessed using the Wald test, and the resulting p-values were corrected using the Benjamini-Hochberg method. Genes with adjusted p-values < 0.05 were considered to be significantly associated with CF proportion.

To assess the distribution and clustering of transcriptional diversity of strains with respect to CF proportion, we used *k*-means clustering on the principal components (PCs) of the gene

expression data. We performed clustering using a range of *k* values and chose *k* = 20 because it was associated with a relatively low within-cluster sum of squares (WCSS) and reasonable cluster sizes (mean cluster size = 12). We ran the algorithm 100 times with different initial centroids and used the initial centroid placement which minimised the WCSS. We then calculated the mean standard deviation (σ) of the CF proportion by cluster (mean σ = 0.135). Next, we randomly sampled the CF proportions per strain and re-calculated the mean σ each time, repeating this process 10,000 times in total. An empirical *P* value was calculated as the fraction of random samples with mean σ at least as low as that observed with our actual data. Since no random samples had mean σ as low as our actual observed value (smallest mean σ = 0.14), the empirical *P* is $< 1 \times 10^{-4}$.

Zebrafish infection experiments

Zebrafish lines: Experimental procedures were performed using the wild type AB line and the knockout cftrsh540 mutant (*101*). Macrophage activity was evaluated using the Tg(mpeg1:mcherry-F)ump2 line harboring red macrophages (*102*). For zebrafish anesthesia procedures, larvae are immersed in a 0.168 mg/mL Tricaine (Sigma-Aldrich) solution in fish water. When required, larvae were cryo-anesthetized by incubation on ice for 10 min then euthanized using an overdose of Tricaine (0.500 mg/mL).

Morpholino injection: Morpholino were purchased from Gene Tools. The morpholino for cftr knockdown (5'-GACACATTTTGGACACTCACACCAA-3') were prepared and injected into one-cell-stage as previously described (*80*). A standard morpholino control (5'-CCTCTTACCTCAGTTACAATTTATA-3') was used as a negative control.

Generation of fluorescent Pseudomonas aeruginosa strains: Plasmids were obtained from Addgene. *P. aeruginosa* GFP+ strains were obtained by triparental mating, using an E. coli strain containing the plasmid pRK2013 (*79*, *81*) as helper strains and an E. coli containing pMF230 (*82*) as the donor strain. GFP+ colonies were selected on LB agar plates complemented with carbenicillin 300 µg/ml and 5 µg/ml tetracycline (LBCarb/Tet) (Sigma-Aldrich).

Pseudomonas aeruginosa inoculates and microinjection into zebrafish embryos and larvae: P. aeruginosa expressing GFP were grown using LBCarb/Tet medium. To prepare *P. aeruginosa* inoculates, 1 ml of LBCarb/Tet medium was inoculated with a single colony of bacteria and incubated at 37°C overnight with shaking. 50 µl of this overnight culture was then added to 5 ml of LBCarb medium and incubated at 37°C with shaking to mid-log phase (OD600 0.6-0.8). *P. aeruginosa* were next harvested by centrifugation and resuspended in a volume of PBS (GibcoTM, Thermo Fisher ScientificTM). Then bacterial suspensions were homogenized with a 26-gauge needle and resuspended at an OD_{600} of 1 in PBS, then kept on ice until zebrafish infection challenges. Systemic infections were carried out by the microinjection of P. aeruginosa into the caudal vein of 30 hours post-fertilization (hpf) zebrafish embryos as described earlier (*48*) with some modifications. Briefly, tricaine-anaesthetized embryos were infected individually with 1-2 nl *P. aeruginosa* inoculate (≈250-350 colony forming units (CFU)). Inocula were checked a posteriori by microinjection onto LB agar plates. Survival post *P. aeruginosa* infection was assessed daily by counting dead embryos (no heartbeat) up to 6 days.

Zebrafish whole bacterial burden analysis: Growth of *P. aeruginosa in vivo* was assessed by CFU analysis at 1 day post-infection (dpi). To determine CFU, groups of five larvae were anesthetized, collected, euthanized and transferred individually into microfuge tubes with 1% Triton X-100 (Merck Millipore) in PBS. Larvae were then mechanically homogenized using a micro pestle (Eppendorf), washed to removed Triton and resuspended in PBS. 10-fold serial dilutions were plated on LB agar plates and incubated overnight at 37°C prior to CFU counts.

Zebrafish macrophage activity and Intra-macrophage bacterial burden analysis: Macrophage response was elicited through infection of *P. aeruginosa* expressing GFP into the muscle

compartment of anesthetized Tg(mpeg1:mcherry-F)ump2 larvae at 3 dpf as previously described (*39*). Macrophage chemotaxis and phagocytosis was evaluated by assessing the number of cells and/or infected cells at infection sites using confocal microscopy. Macrophage chemotaxis was determined at 2 hours post-infection (hpi), and phagocytosis at 4 hpi. Intra-macrophage P. aeruginosa loads, which reflects the bacterial killing, were assessed by quantifying the volume of bacteria inside macrophages at 6 hpi using confocal microscopy.

Zebrafish Microscopy, Image Processing and Analysis: Bright-field and fluorescence microscopy of infected embryos/larvae were performed using an Olympus MVX10 epifluorescent microscope equipped with a X-Cite120Q (Lumen Dynamics) 120-W mercury light source. Images were acquired with a digital color camera (Sony MiC5 Pro) and processed using CellSens (Olympus). Confocal fluorescence microscopy was performed using an ANDOR CSU-W1 confocal spinning disk on an inverted NIKON microscope (Ti Eclipse) and 20x/NA 0.75 air and 40x/NA 1.15 water objectives. Images were acquired with a W&B Zyla 4.2 camera (ANDOR) and processed using IQ3 3.6.5 software (ANDOR).

Overlays of fluorescent and DIC images and 2D reconstructions of image stacks were assembled using FIJI freeware. 3D reconstitution was produced using Imaris 9.0 software (Bitplane).

Mutational burden analysis

PAO1-rooted clones trees were used a input to Treetime 0.8.1 (*102*) to infer a maximum likelihood ancestral character state reconstruction of every nucleotide position. We then implemented a pipeline (https://github.com/aweimann/PhyloEffects) to identify single nucleotide changes and annotate variant effect in their phylogenetic context using the gene annotation from Pseudomonas.com (PAO1 107) (*103*) and the ancestral character state reconstructions. Parsimony ancestral character state reconstruction was used to infer ancestral insertions and deletions. Only unique insertions or deletions were kept. SNPeff version 4.3.1 (*55*) was used to annotate the effect of indels on gene function.

We performed a mutational burden test assuming a Poisson distribution of the mutational burden per gene. We compared the observed number of non-synonymous SNPs and indels within a gene across all clones with the expected number of variants in that gene based on the total number of variants across all clones. *Panaroo* (*35*) was used to estimate the prevalence of every orthologous gene family and infer an adjusted number of expected variants. Multiple-testing correction was used to account for the number of tests (namely the number of genes in the reference genome *PAO1*) using the Benjamini-Hochberg method (*76*) to control the false discovery rate at 5%. The 224 genes passing the adjusted p-value threshold were used to query the STRING 11.5 database (*77*) of protein-protein interaction. STRING reports the statistical significance of the number of interactions found among the input set of genes compared with the number of expected interactions by chance.

Pathoadaptive genes were assigned to 17 functional categories based on the gene products description on Pseudomonas.com (*78*) (Figure S1). Genes across the PAO1 genome were also stratified by whether they had an assigned gene product name. A Fisher exact test was used to compare the number of assigned with the number of unassigned genes among pathoadaptive genes and non-hits.

Genotype-TF regulon analysis

To assess the association between genetic variants and the expression of transcription factor (TF) regulons, gene expression data from (*86*) were pseudoaligned to strain-specific gene sets using Kallisto (*104*). Sample-specific size scaling factors for normalisation were derived from core gene expression using DESeq2 (*84*) and size-scaled counts were $\log_2$ transformed for variance stabilisation. We performed binary scoring of TFs based on the presence (1) or absence (0) of

missense, nonsense or INDEL variants. The normalised expression levels of TF regulons were compared between strains with and without genetic variants using Welch's two-sample *t*-tests. P-values from *t*-tests were adjusted using the Benjamini-Hochberg method.

Phenotyping of pathoadaptive gene transposon mutants

PAO1 mutants with transposon insertions in 154 pathoadaptive genes (selected from the Manoil library (*85*) were arrayed in 96 well plates containing 120ul of no-salt LB broth. Plates were incubated overnight at 37C in a static incubator, after which, DMSO was added to a final concentration of 5% and plates were stored at -80°C. For all assays performed on arrayed mutants, the inoculum was in the form of a fresh over-night culture from the frozen stock plate. All agar plate-based phenotyping assays were performed in duplicate Plus Plates (Singer Instruments) that were inoculated using the ROTOR and 96-Long re-pads (Singer Instruments). Results from all agar plate assays were imaged using the Phenobooth Imager (Singer Instruments).

Swimming motility: the inoculum was stabbed through the agar until the pins contacted the base of the plate resulting in a column of bacteria within the agar layer. Agar plates were incubated at 37°C for between 4 and 6 hours for the formation of an opaque halo surrounding the point of inoculation which was indicative of swimming motility.

Twitching motility: Plus Plates were filled with 12.5ml of low salt LB with 1.5% agar to create a thin, even layer of agar on the bottom of the plate. Inoculum was stabbed through the agar layer into the plastic-agar interface by increasing the pressure applied by the pins on the ROTOR to 100%. Plates were incubated at 37°C for 6 hours under humid conditions, after which they were treated for 30 minutes with chilled TM developer solution (40% methanol and 10% acetic acid) at room temperature. Liquid was decanted from the plates and the agar layer was carefully removed to expose the adherent growth on the plastic of the plate. Plates were allowed to air dry which caused the halo to turn white and become more visible and easier to visualise using the Phenobooth.

Siderophore production: Chrome-azurol (CAS) agar for assaying siderophore production was produced as previously (*104*). To prepare the deferrated cas-amino acid solution, 10g cas-amino acids were dissolved in 100ml of sterile distilled water, and 3g 8-Hydroxyquinilone were dissolved in 100ml chloroform, the 2 mixtures were combined and shaken for around 10 minutes. After shaking, the mixture was transferred to a separating funnel and allowed to rest for 10 minutes. The bottom fraction was discarded, and the top layer collected in a fresh bottle containing 100ml of fresh chloroform. The process was repeated, and the final eluate was left to rest overnight on the bench with the lid loosely attached to allow residual chloroform to evaporate. The deferrated cas-amino acids were then filter sterilised and stored at 4°C. All glassware used for the preparation of CAS agar was rinsed once with 6M HCl followed by washing three times with sterile de-ionised water before use, to eliminate interference. Before the addition of agar, the medium was adjusted to pH 6.8 and after autoclaving, the agar was allowed to cool down to 50 °C before the addition of blue dye, glucose and deferrated cas amino acids. CAS agar plates were inoculated with overnight liquid culture and incubated at 37°C overnight before imaging using the Phenobooth. Formation of an orange halo around the point of inoculation indicated siderophore activity.

Caseinase activity: Skim milk agar was prepared by combining sterile, molten tryptic soy agar (40g dissolved in 1 litre of water) with sterile 10% marvel milk (50ml per litre of agar). Plates were inoculated with overnight culture and incubated at 37°C overnight after which they were observed for the formation of a clear halo.

Gelatinase activity: Medium consisting of nutrient broth (made according to the manufacturer instructions), 1.5% agar and 3% porcine gelatine was dispensed into plus plates. The temperature was maintained at 55°C and the medium was stirred continuously using a magnetic stirrer to ensure a uniform distribution of gelatine throughout the medium. Agar plates were inoculated with overnight culture and incubated at 37°C for 6-8 hours after which they were stored at 4°C overnight to enable the halo to develop fully. After incubation at 4°C, plates were flooded with a solution of 4.1M ammonium sulphate and incubated at room temperature to reveal clear halos that are the result of gelatine degradation by gelatinase. After 30 minutes, the liquid was discarded, and the plates were imaged.

Rhamnolipid production: Agar was prepared containing Trizma base (14.5g), peptone (10g), glucose (5g), NH_4Cl (0.7g), KCl (1.5g), $MgSO_4$ (0.39g), ddH2O (1l), adjusted to pH 7.4. before the addition of agar (15g) and autoclaving. The agar was then allowed to cool to approximately 50°C before it was supplemented with 10ml of sterile CTAB solution and 10ml of sterile methylene blue. Agar plates were inoculated with fresh overnight culture and incubated overnight at 37°C. On the second day, plates were transferred to 4°C for a further 48 hours to allow halos to develop.

Impact of amino acid changes on protein stability and structural analysis

All amino acid changes were analysed with SIFT 4G 6.2.1 (*105*) using the UniProt90 database (*84*). Since *Pseudomonas* does not have a pre-computed database, we built our own SIFT database using the genome and gene annotations from Pseudomonas.com. All non-synonymous mutations were then annotated with the *annotator* command. FoldX 5 (*105*) was used to predict the difference of total energy between the variant and wild type allele for every variant. AlphaFold models were downloaded from Uniprot and the *RepairPDB* command of FoldX was used to repair residues with bad torsion angles or total energy and van der Walls' clashes (*55*). The repaired PDBs were then used as input to the *BuildModel* FoldX command. FoldX and SIFT scores were then averaged per gene/protein and a two-tailed t-test was used to compare the average scores between mutational burden test hits and non-hits.

Mutational frequencies were mapped on the structural models of the identified hotspot genes (*algU*, *ladS*, *pcnA* and *betT2*) in *P. aeruginosa* using the Chimera molecular modelling package (*74*). Coordinates for AlgU were downloaded from the Protein Data Bank. Predicted AlphaFold models (*75*) were downloaded from UniProt for *ladS, pcnA* and *betT2*.

Transmission and host selectivity of pathoadaptive mutations

To assess the transmissibility of pathoadaptive changes, the number of mutations that had been observed in at least two isolates (from different patients) was compared with mutations that had only been observed once using a Fisher exact test. TopGO 2.4.6 was used for functional enrichment analysis of the host-specific Gene Ontology biological pathway annotation compare to the background (*62*) enrichment. Gene Ontology biological process annotations were downloaded for the PAO1 strain from *Pseudomonas.com* (*51*).

Mutations in pathoadaptative genes were stratified by the (ancestral) infection type (CF or non-CF) of every branch based on outgroup-rooted rooted clone trees. To assess host-specific pathoadapdation, the number of CF vs non-CF mutations were compared using a Fisher exact test. Mutations on branches with non-concordant ancestral infection types were discarded. Multiple-testing correction to account for the number of tests (the number of pathoadaptive genes) was achieved using the Benjamini-Hochberg method to control the false discovery rate at 10%.

Pathoadaptive trajectories

Trajectories were inferred as the sequence of mutations in pathoadaptive genes since the emergence of the clone ancestor (random assignment was used where several mutations

coincided on one branch) as implied by the PAO1-rooted tree stratified by cystic fibrosis (CF) and non-CF infection types. Mutation frequencies were position normalised and the frequency plots of the 40 genes with the lowest p-value from the mutational-burden test were manually assigned into five groups of genes with similar frequency curve shapes. Trendlines were generated by locally-weighted smoothing.

Transmission analysis

We established a relatedness cut-off to define potential transmission links using pairwise SNP differences between pairs of isolate genomes from the same patient (n = 81 patients). Potential hypermutators were removed prior to this analysis as described above. We defined a patient-level cut-off as the 95th percentile of the distribution of within-host SNP distances for every patient as described previously (*39*). A global cut-off was derived as the 95th percentile of the distribution of all patient cut-offs. We then identified potential transmission events as isolates from the same clone sampled from different patients that differed by 26 SNPs or fewer. Transmission clusters were annotated and laid out using Cytoscape.

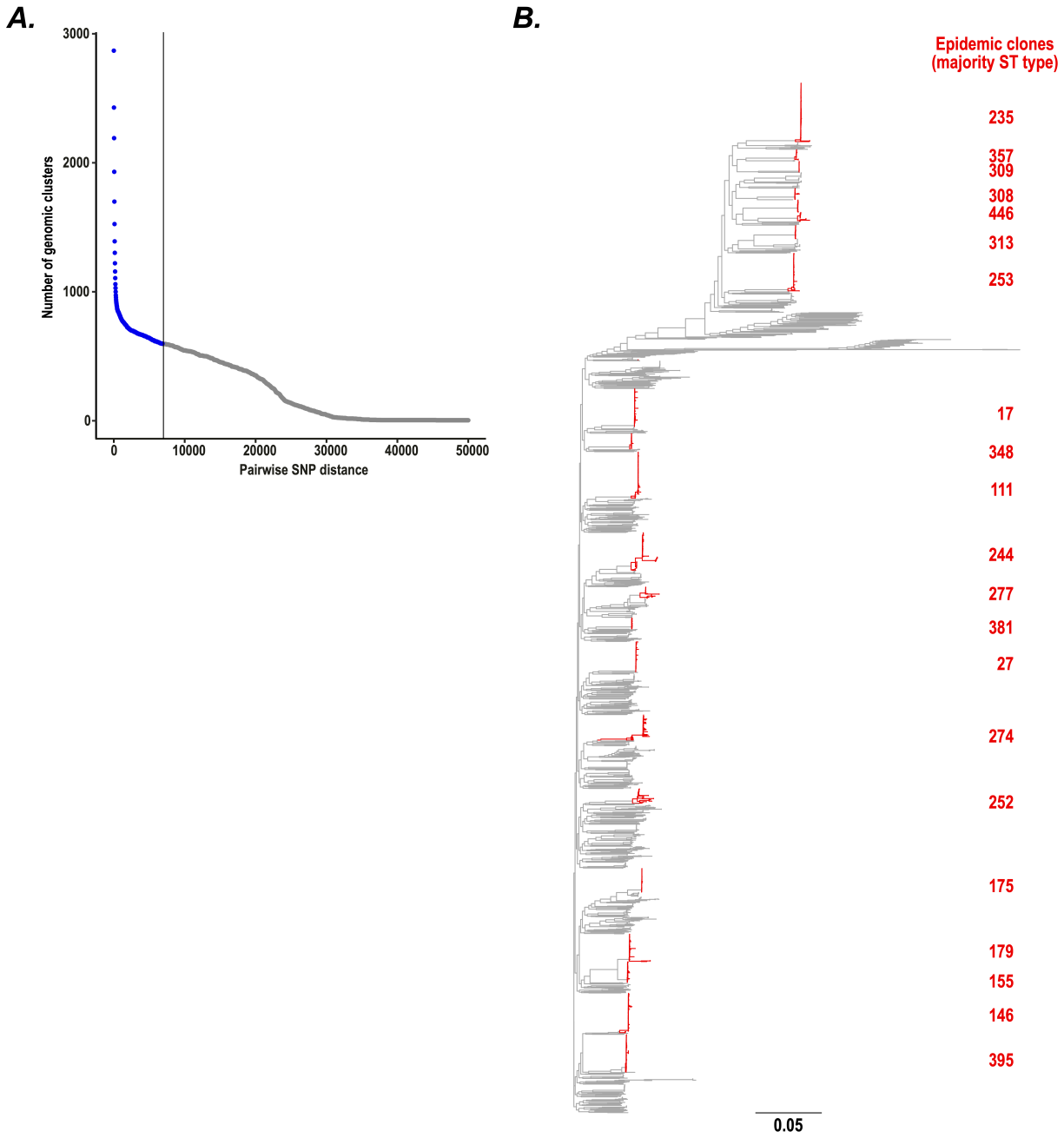


Figure S1. (A) Relationship between pairwise short nucleotide polymorphism (SNP) distance threshold and implied number of genomic clusters. Isolates genomes were clustered based on their pairwise SNP distance using the unweighted pair group method with arithmetic means (UPGMA). UPGMA infers a sample dendrogram which is then separated into cluster by applying a discrete SNP distance threshold. A vertical black line at 7000 SNPs denotes the threshold that was chosen to define the genomic clusters (clones; *blue*) in our dataset. **(B)** Maximum likelihood phylogenetic tree generated from genomes of all study isolates (major epidemic clones labelled in red).

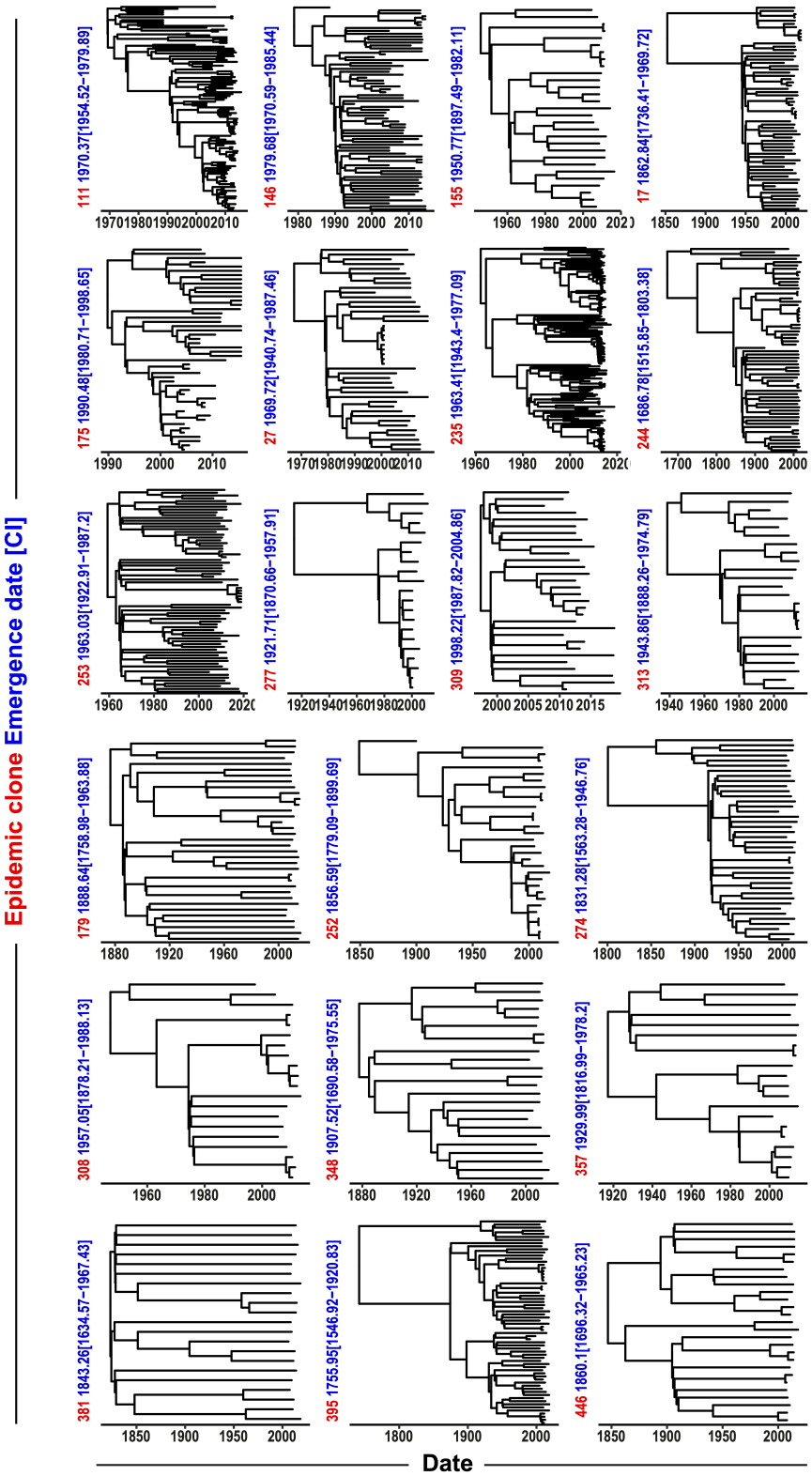
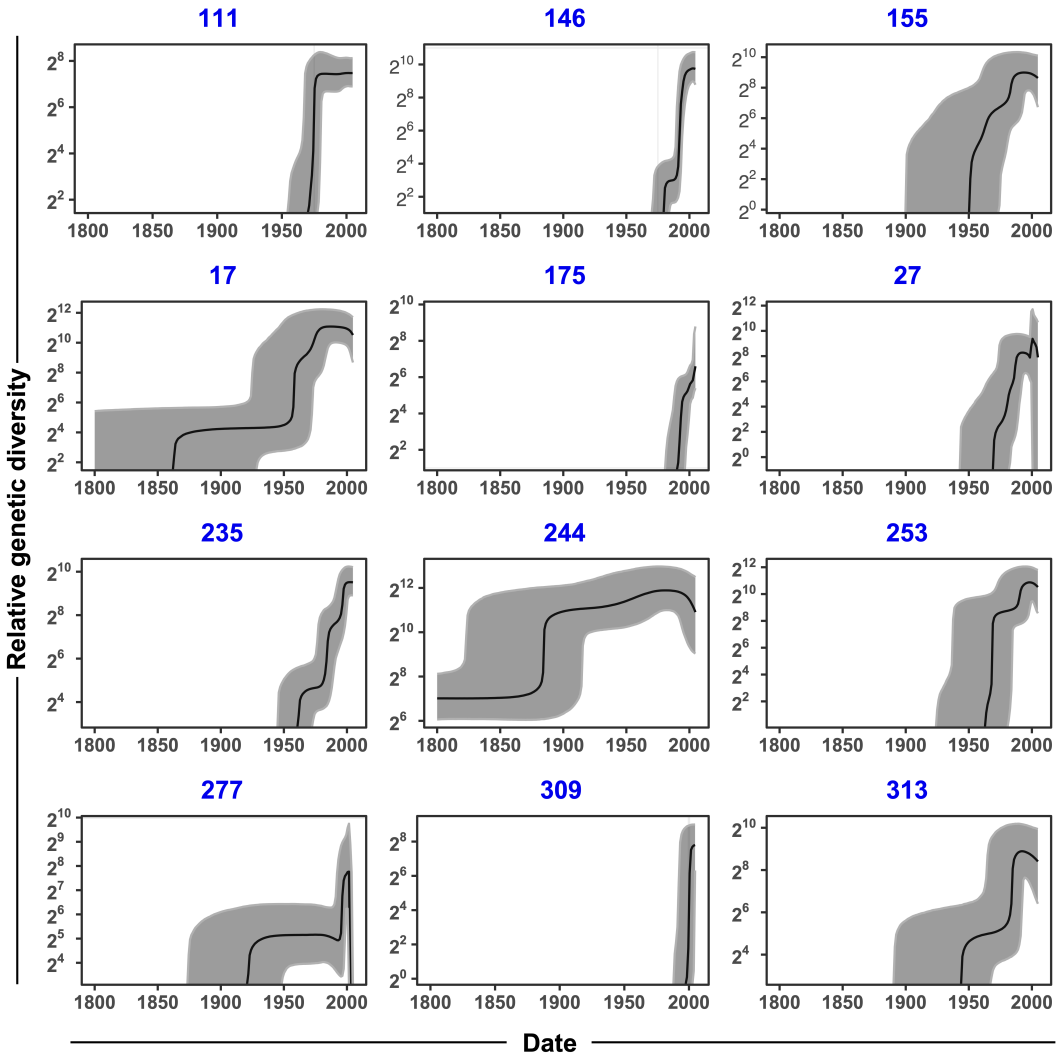


Figure S2. Emergence dates of epidemic clones. Bayesian inferred phylogenetic trees for all epidemic clones with mean and 90% highest posterior density interval estimates of the emergence date.


Epidemic clones (Majority ST)
111
146
155
17
175
27
235
244
253
277
309
313
Relative genetic diversity
Date
1800
1850
1900
1950
2000

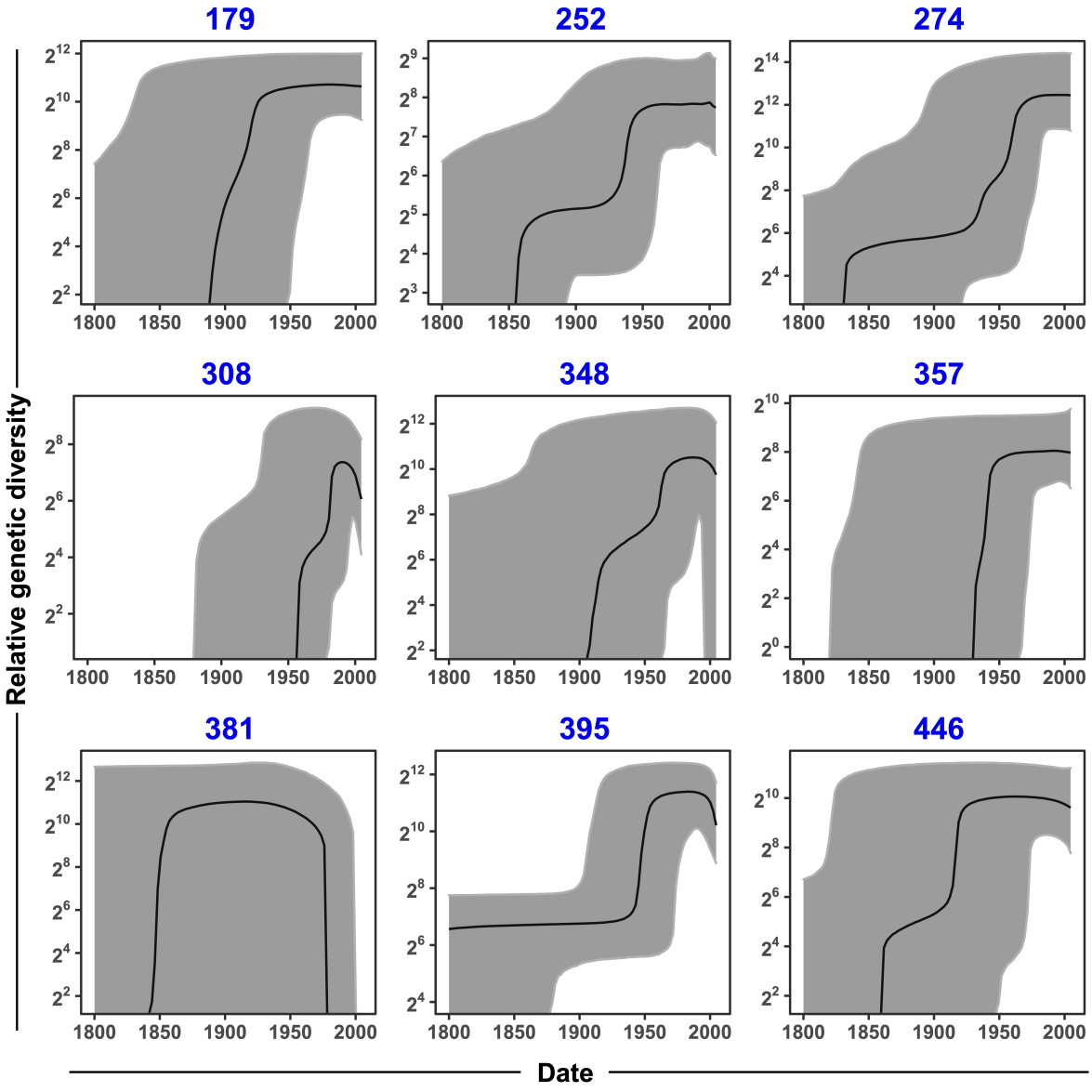


Figure S3. Population expansion of epidemic clones. Relative genetic diversity estimates (as a proxy for population size) based on Bayesian Skyline plot demographic model were inferred for all epidemic clones. 90% highest posterior density intervals are shown in grey.

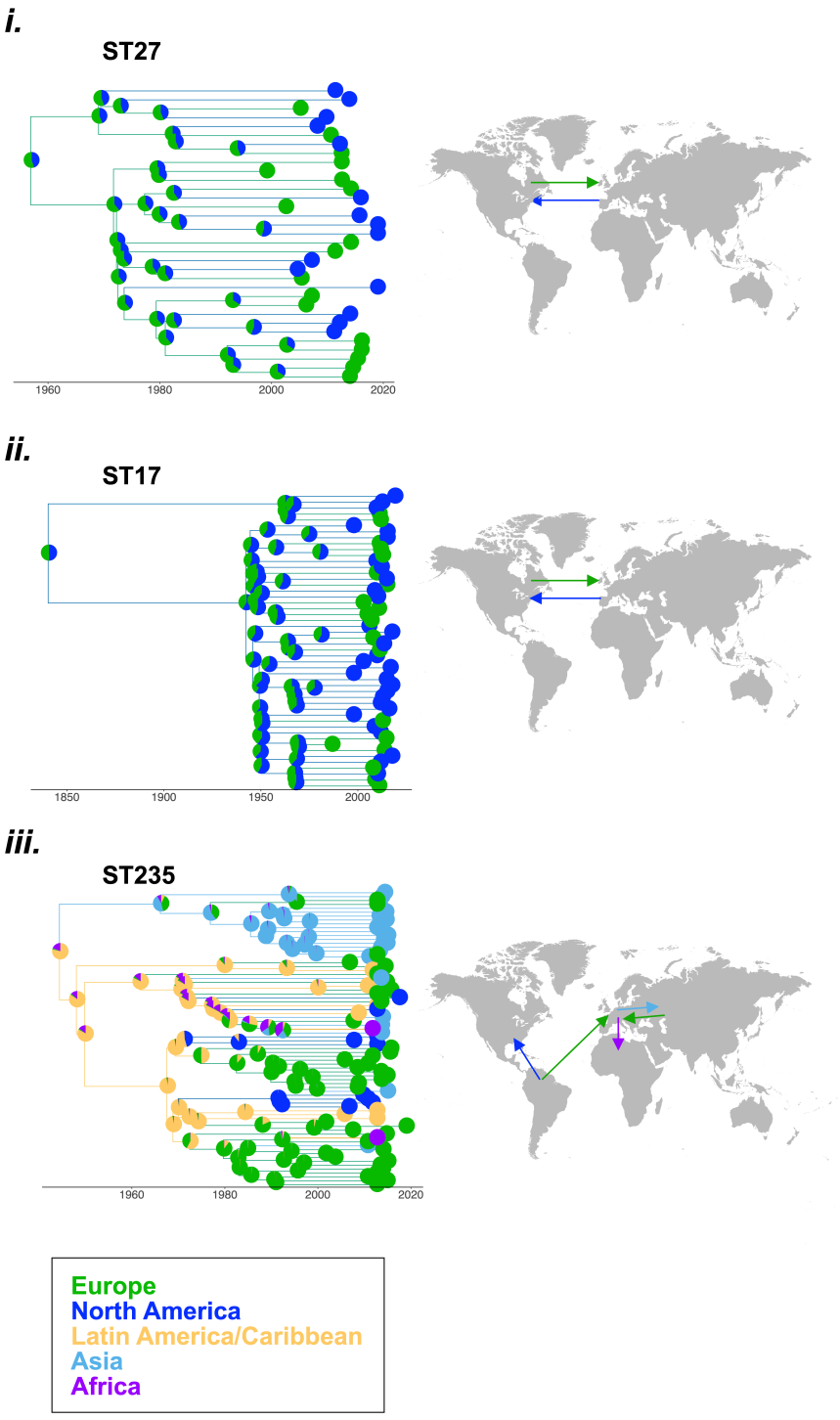


Figure S4. Phylogeography of *P. aeruginosa*. *Left* Bayesian inferred phylogeographic trees shown for three epidemic clones (i) ST27, (ii) ST17, (iii) ST235. Edges were colour-coded by the most probable continent based on the full distribution of trees (*N* = 27,000). Proportion of trees supporting individual continent are shown at nodes. *Right* Arrows show the statistically-supported direction of inter-continental migrations routes (Bayes Factor > 3). Arrows are colour-coded by the recipient continent.

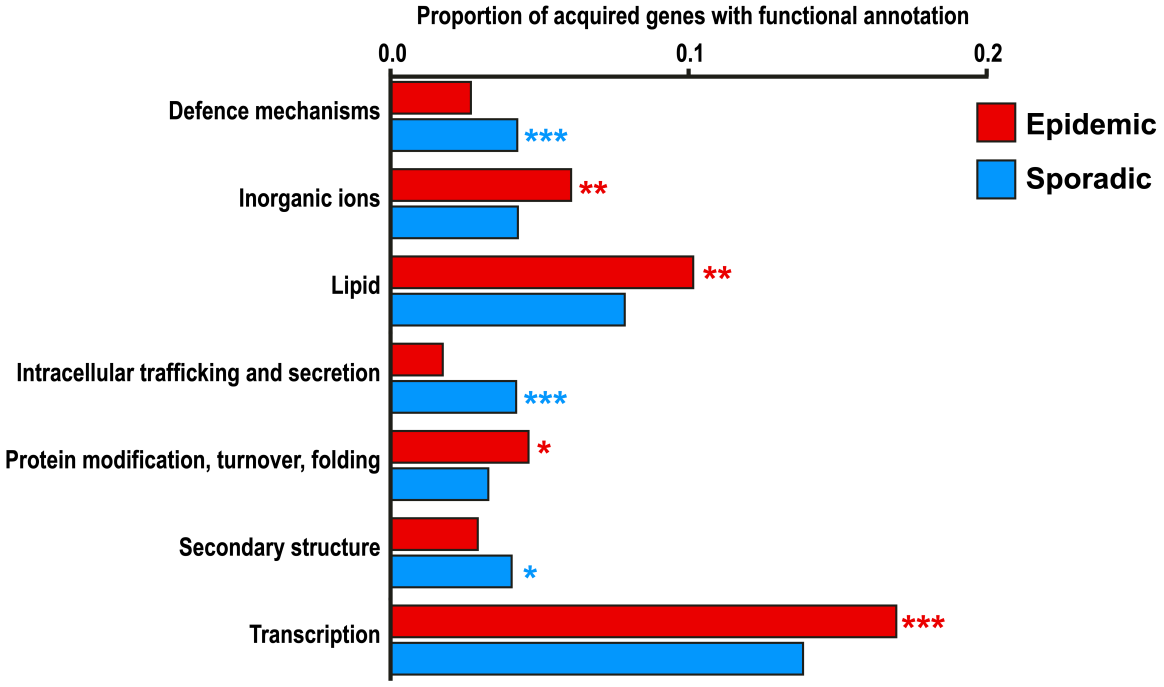


Figure S5. Gene gain events in epidemic and sporadic *P. aeruginosa* clones. Proportion of acquired genes with functional annotations (based on COG database), inferred by maximum parsimony ancestral genome reconstruction within epidemic (*red*) and sporadic (*blue*) clones. All COG categories with significantly different numbers of acquired genes between epidemic and sporadic clones are shown. * p value < 0.05; ** p value < 0.01; *** p value < 0.001.

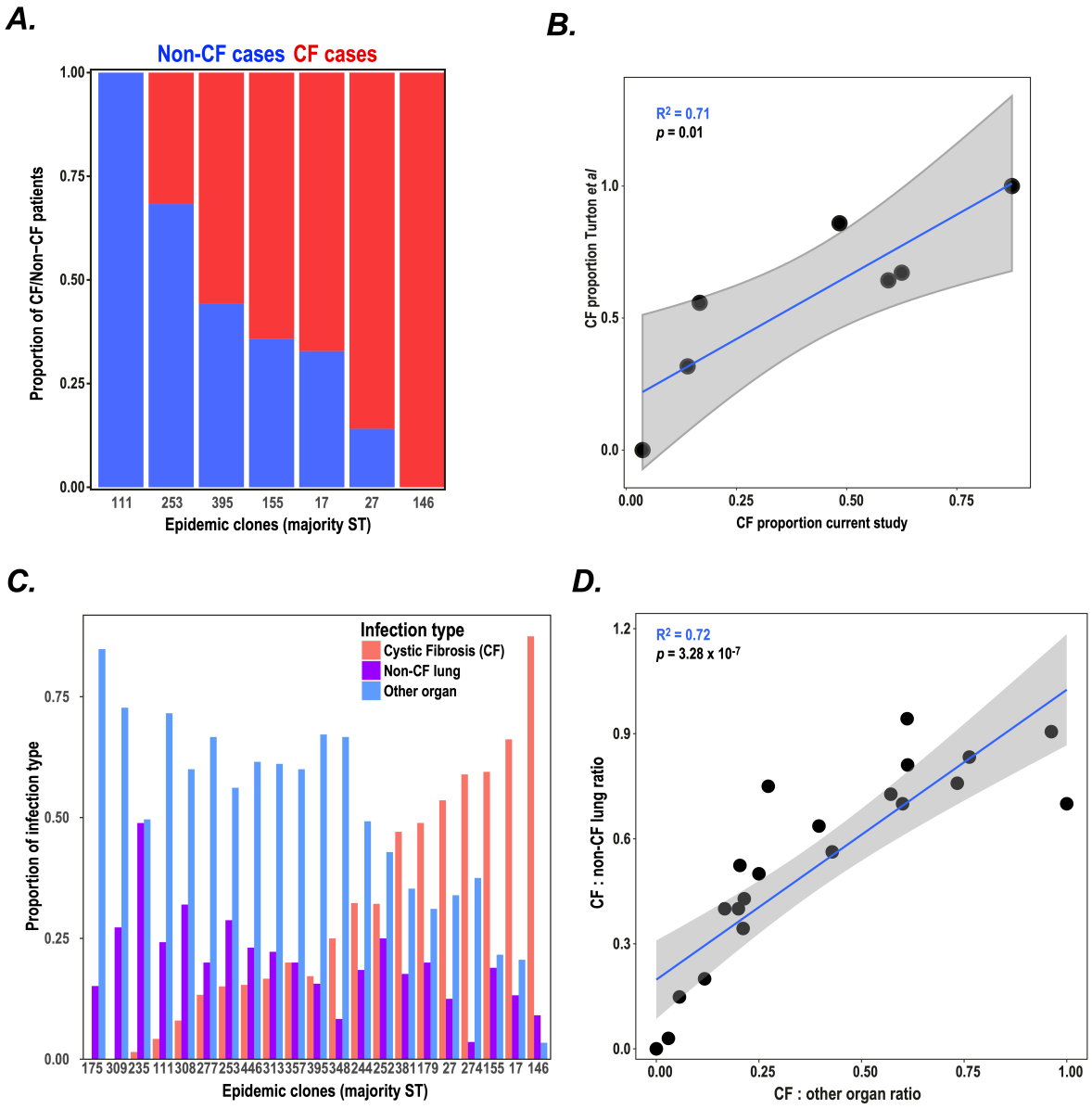


Figure S6. Further analysis of varying host preference of epidemic *P. aeruginosa* clones. (**A**) Proportion of cystic fibrosis (CF; *red*) and non-CF infections (*blue*) caused by epidemic clones also represented in our study data, analysed from data collected as part of a surveillance study with UK hospitals between 2010 and 2012 (*63*). (**B**) Comparison of CF proportion in individual epidemic clones between our study and previous surveillance study *(33)* (Adjusted $R^2 = 0.71$, $p = 0.01$, F-test). (**C**) Proportion of infections caused by epidemic clones (labelled by their majority multi-locus sequence type, ST) in patients with cystic fibrosis (CF; *red*), non-CF lung infection (*purple*), and non-lung infections (other organ; *blue*). (**D**) Comparison of ratio of CF to non-CF lung infections compared to the ratio of CF to other organ infections for major epidemic clones (Adjusted $R^2 = 0.72$, $p = 3.28 \times 10^{-7}$, F-test).

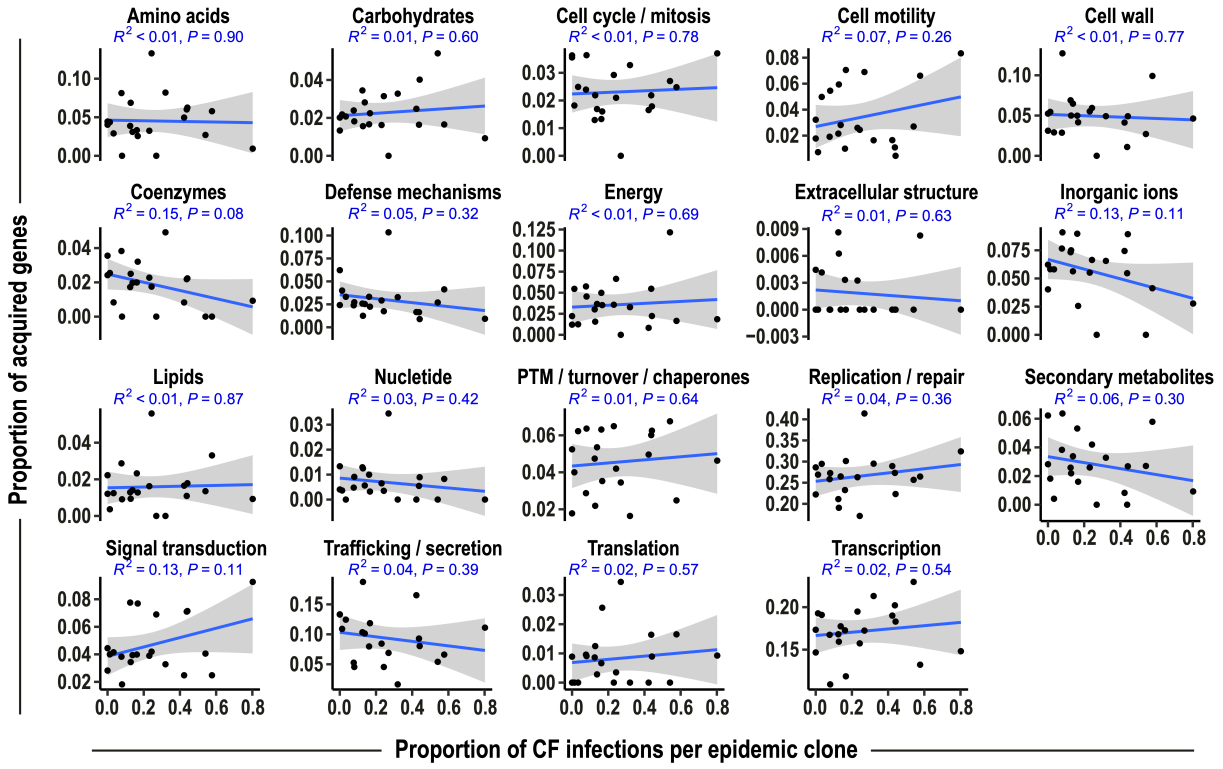


Figure S7. Gene gain events in epidemic clones across the *P. aeruginosa* CF affinity spectrum. Ancestrally-acquired genes were annotated (based on COG database) using maximum parsimony ancestral genome reconstruction across ancestral genome representatives. Each panel shows the proportion of acquired genes annotated within a specific functional category across the CF proportions of the underlying clones. Linear trendlines are shown in blue with shaded area in grey denoting the 95% confidence level interval. Adjusted R^2 and p-values are shown based on fitting a linear model for every category (F-test).

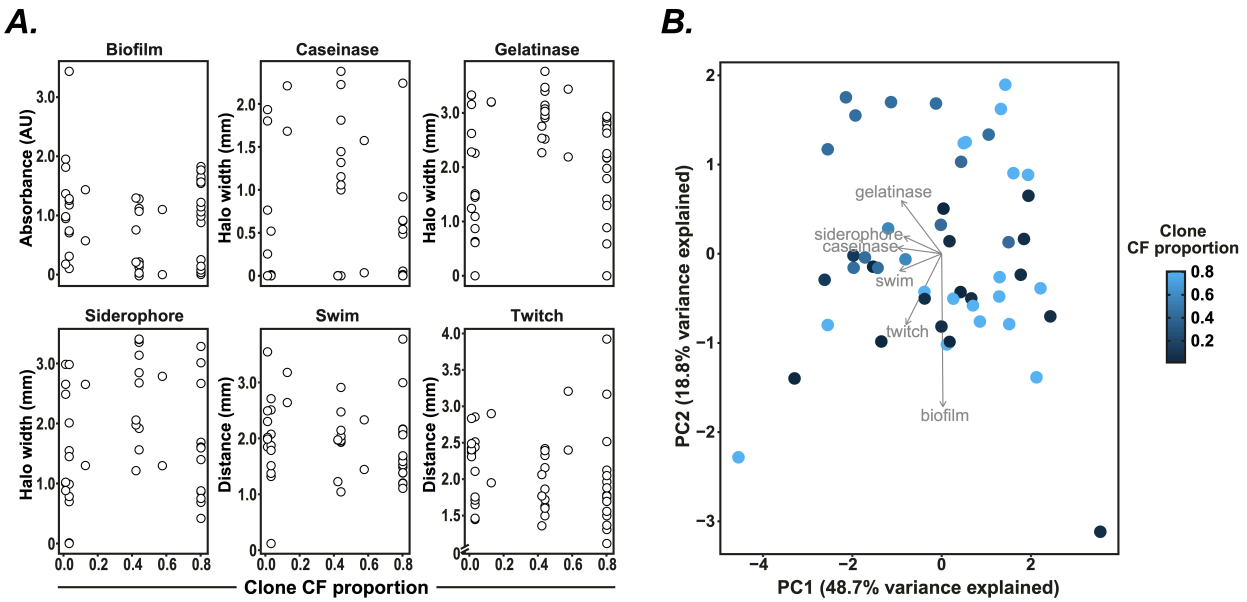


Figure S8. Multi-dimensional virulence phenotyping of *P. aeruginosa* clinical isolates. Virulence phenotypes of 49 clinical isolates as determined using a series of phenotypic assays (see Method section). (**A**) Virulence phenotype quantification for each isolate plotted against the clonal CF proportion of each clinical isolate. (**B**) Principal component (PC) analysis of virulence phenotype quantification for all tested clinical isolates, visualised using the two principal component (PC) axes explaining the most variation (variance explained annotated on the axes) based on PC analysis of all virulence factor measurements. Clonal CF proportions annotated for each isolate using blue colour scale.

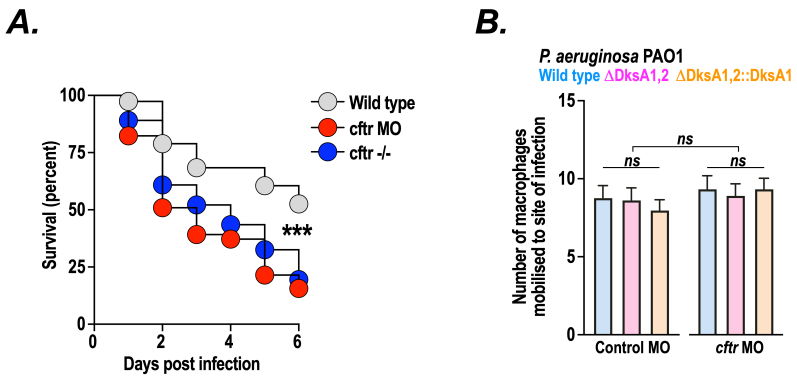


Figure S9. Modelling *P. aeruginosa* infection in zebrafish. (**A**) Wildtype (*grey*), cftr knockout (*cftr* -/-; *blue*) (*43*) and *cftr* morphant (*cftr* MO; *red*) zebrafish were intravenously infected with 250-350 colony forming units (cfu) wildtype *P. aeruginosa* PAO1. Survival analysis of *P. aeruginosa*-infected larvae. Data plotted as percentage of surviving animals over 6 days (average of 2 independent experiments; n=66). *** $p < 0.001$ (Mantel-Cox Log-rank test). (**B**) Control and *cftr* morphant zebrafish larvae with mCherry-labelled macrophages (*Tg(mpeg1:mcherry-F)ump2* (*44*)) were intramuscularly infected with 250-350 GFP-labelled *P. aeruginosa* PAO1 wildtype (*blue*), ΔDksA1,2 (*pink*) or PA01ΔDksA1,2::DksA1 complemennted (*yellow*) strains and the infection tracked using real-time intravital confocal microscopy. Mean ± SEM (standard error of the mean) number of macrophages mobilized to the infected muscle at 2 hours post infection (hpi). (n=14 fish; 2 independent experiments). *ns* non-significant (two-way ANOVA with Tukey's post-test).

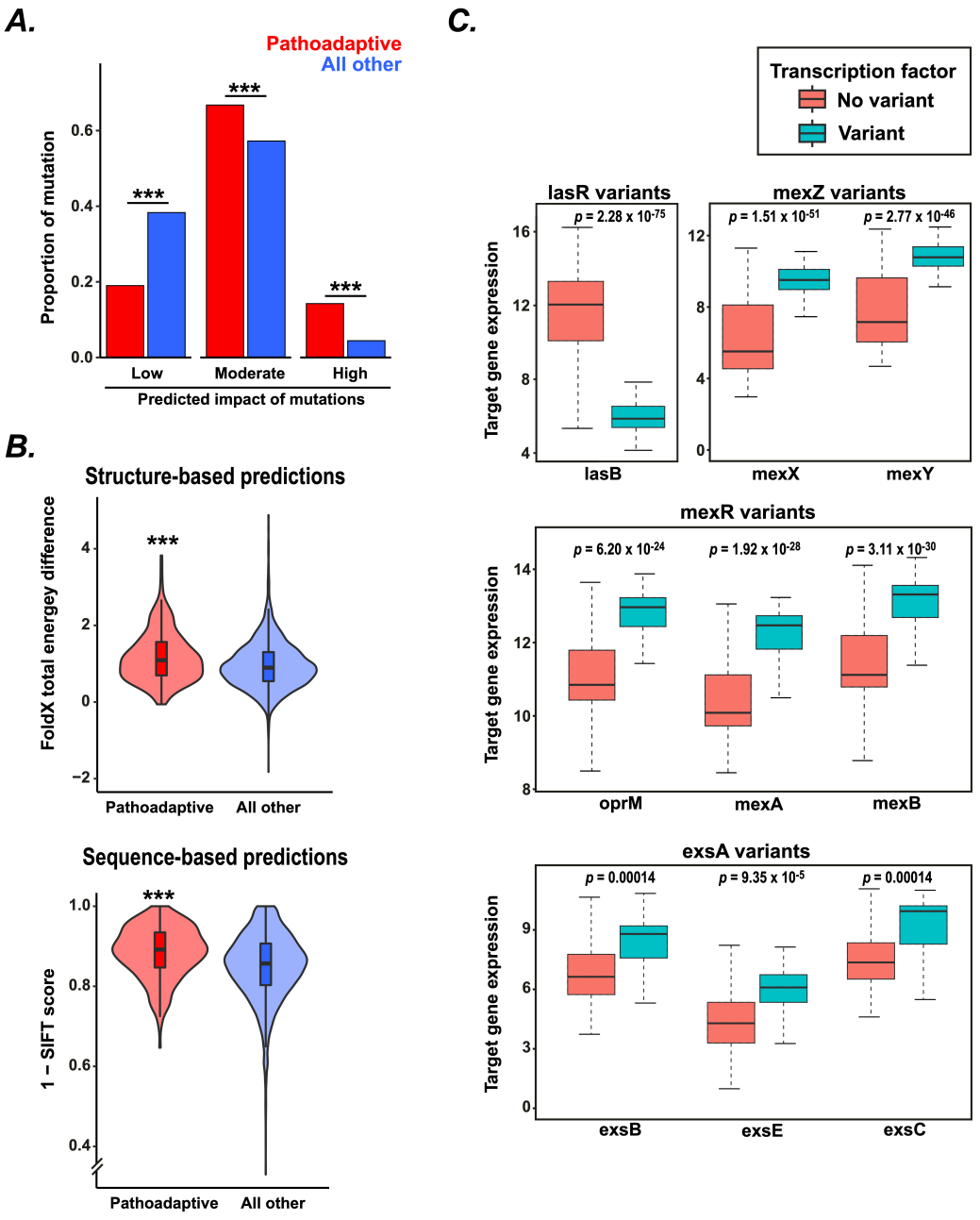


Figure S10. Predicted impact of mutations on protein function (**A**) Proportion of mutations with low, moderate, and high predicted impact in pathoadaptive (*red*) or all other (*blue*) genes (other; *blue*) estimated using SNPeff (*52*). (**B**) The estimated impact of missense mutations on protein stability predicted by (top) FoldX (*59*) (*** p = 1.34 x 10^{-6}) or (*bottom*) protein function predicted by Sorting Intolerant from Tolerant (SIFT) (*57*) analysis (*** p = 9.04 x 10^{-15}) in pathoadaptive (red) or all other (*blue*) genes. (**C**) Boxplots of pathoadaptive transcription factor regulon expression of clinical isolates with variants (*green*) or no variants (*red*) (Benjamini-Hochberg adjusted p-values from a two-sample Welch t-test shown separately for every target gene). Observed changes in target expression all consistent with the presence of loss of function mutations affecting transcription factors.

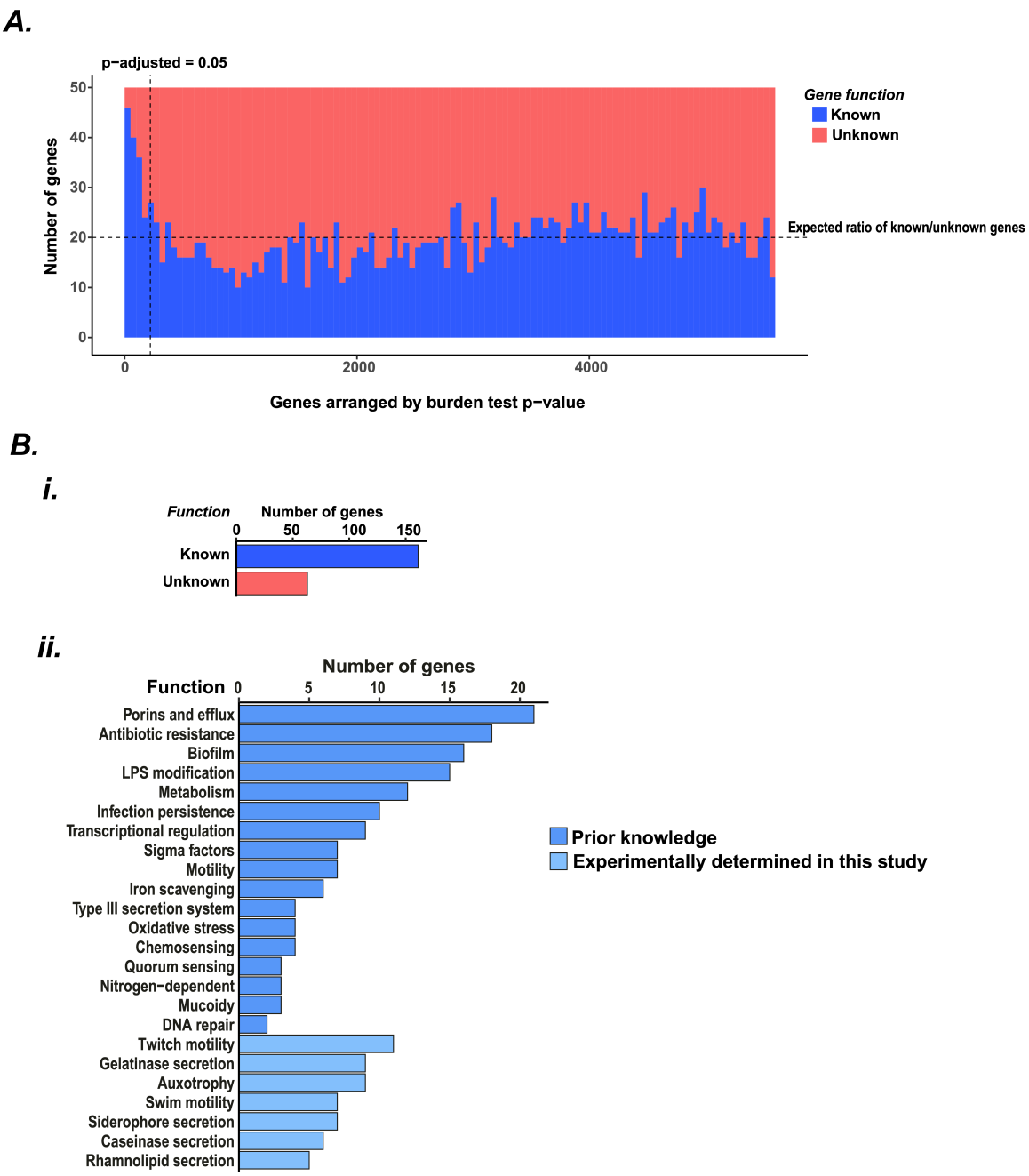


Figure S11. Functional characterisation of pathoadaptive genes. (**A**) Histogram of genes with known (*blue*) or unknown (*red*) function arranged by their adjusted p-value from a Poisson test comparing the mutational burden as expected with the observed number of mutations, with pathoadaptive genes having adjusted *p* values of <0.05. (**B**) *(i)* Fraction of pathoadaptive genes with known (*blue*) and unknown (*red*) function. (ii) Number of pathoadaptive genes with (*dark blue*) known functions (across 17 categories) based on prior knowledge or (*light blue*) experimentally determined in this study using sequence-confirmed transposon mutants from the Manoil library (*64*) representing isogenic disruption of 154 of the 224 pathoadaptive genes.

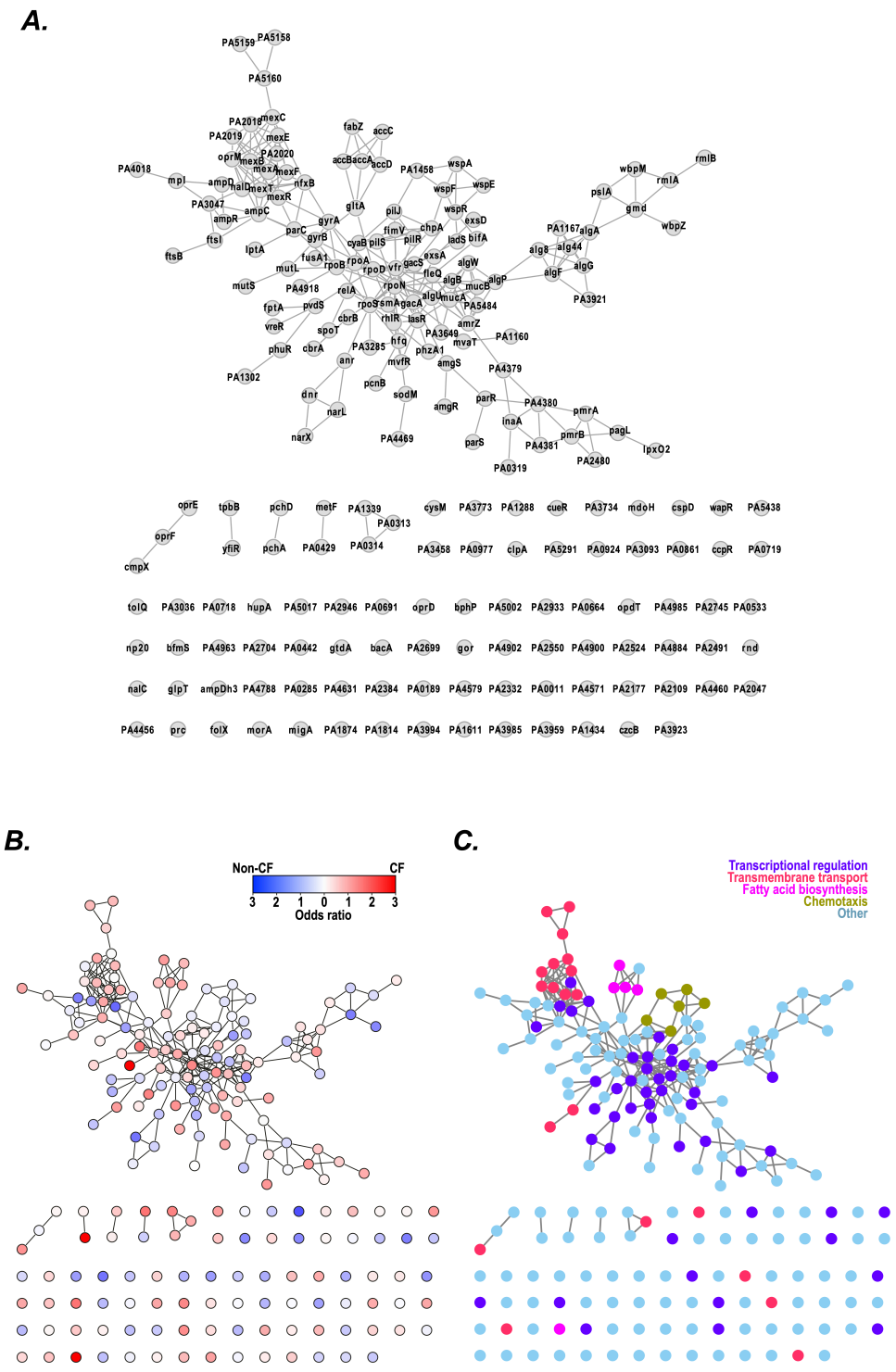


Figure S12. Protein-protein interaction networks of pathoadaptive genes. Genes are depicted as nodes and connected by an edge if their proteins are known to interact according to the STRING database (confidence > 0.7). Protein-protein interaction networks for all pathoadaptive genes (A) labelled by gene name or annotated by (B) the odds-ratio of host-specific pathoadaptation (assessed by comparing the number of CF vs non-CF mutations using a Fisher exact test (FDR = 0.1), or (C) biological process based on Gene Ontology (GO) terms.

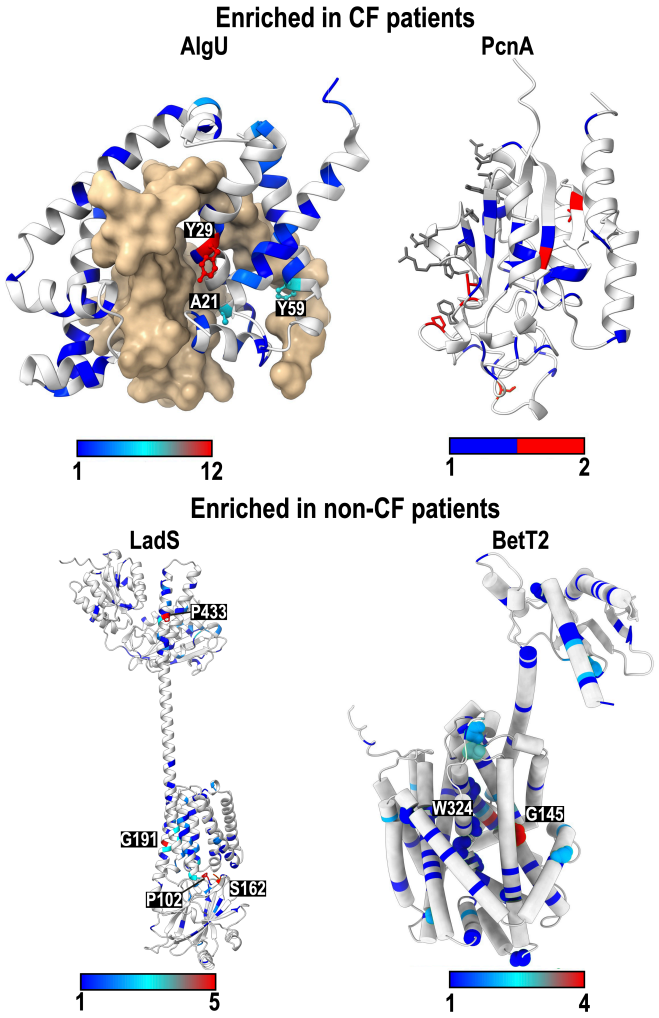


Figure S13. Examples of structural analysis of pathoadaptive mutations. Examples of structural analysis of gene products where mutations are enriched in CF patients (AlgU, PcnA) or in non-CF patients (LadS, BetT2). Frequency of mutations (information entropy) were colour-coded and mapped onto each structure (the most frequent positions were labelled). AlgU, an RNA polymerase sigma-H factor known to regulate mucoidy (*65*), is shown in complex with its negative regulator MucA as cartoon and molecular surface representations respectively with mutations occurring at the interface between these two proteins. PcnA, a putative nicotinamidase, had mutations within the protein core or at sites of protein-protein interaction. LadS, a calcium-responsive histidine kinase (*66*), acquired mutations in the N terminal (sensor) and transmembrane domains. BetT2 (*67*), a putative choline transporter, showed helix-breaking mutations within its transmembrane domain. Models based on experimentally-derived structures (for AlgU, PDB 6IN7) or AlphaFold predictions (for others).

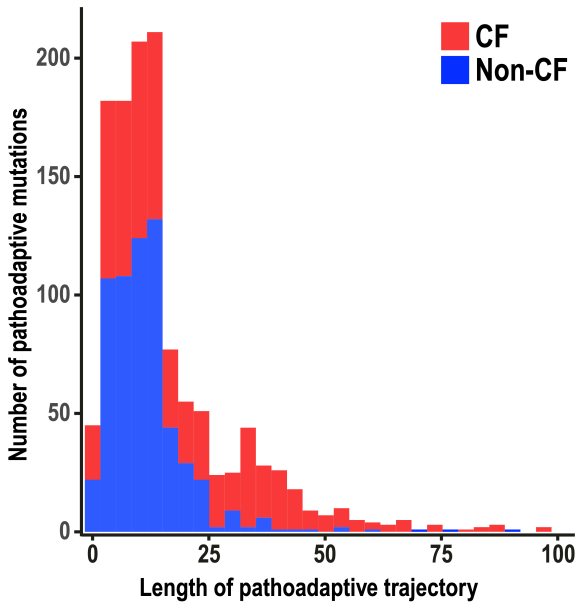


Figure S14. Evolutionary trajectory lengths in *P. aeruginosa.* Comparison of the lengths of evolutionary trajectories between isolates infecting CF (*red*) and non-CF (*blue*) individuals. Trajectories were inferred as the sequence of mutations in pathoadaptive genes since the emergence of the clone ancestor as defined by the mid-point-rooted tree stratified by CF and non-CF infection types. Evolutionary trajectories were down-sampled to contain an equal number of samples from CF and non-CF infections.

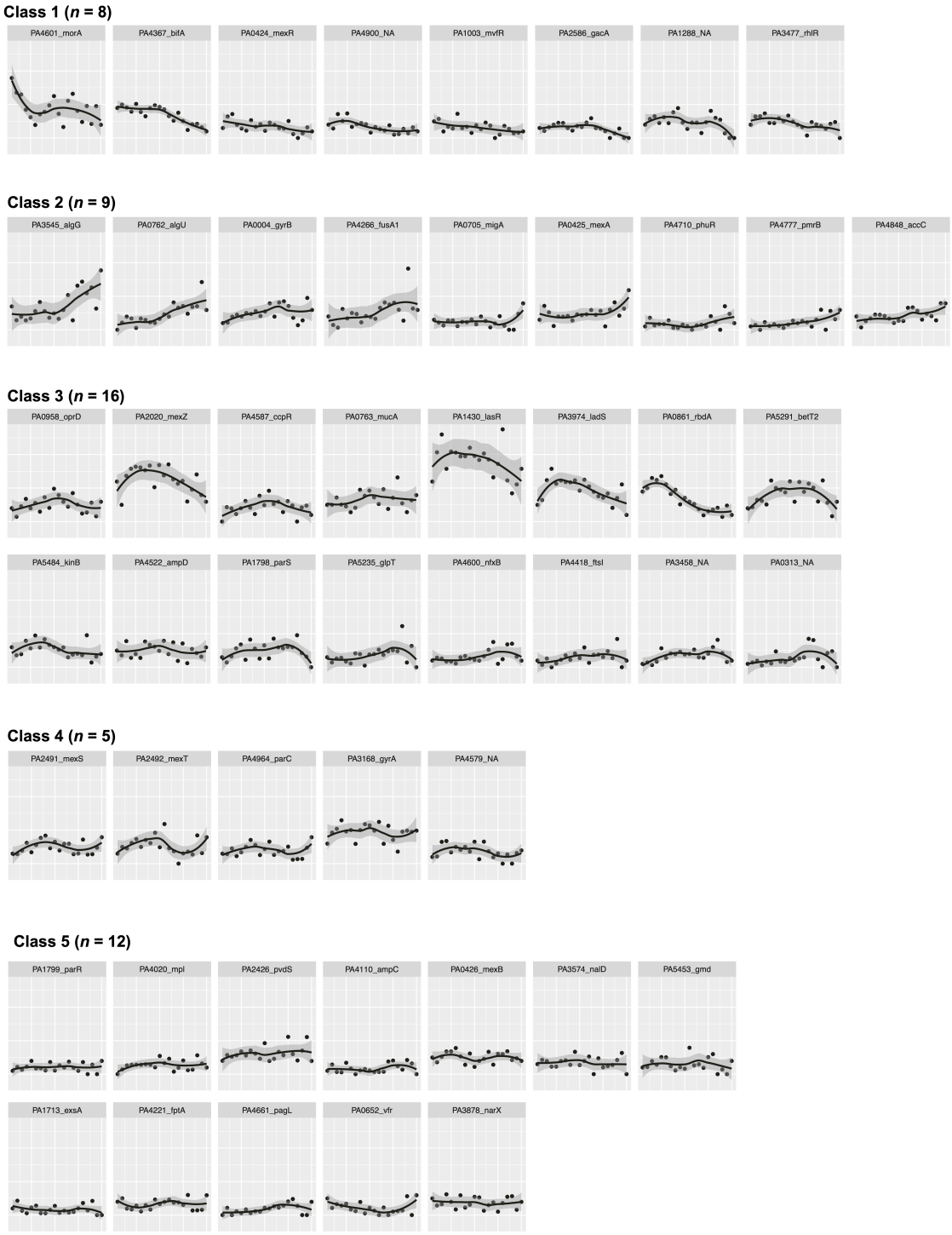


Figure S15. Frequency of mutations over evolutionary time in pathoadaptive genes. Mutation frequencies were position-normalised and the trajectories of the 50 most frequently mutated genes were manually assigned to one of five classes of genes with similar frequency curve shapes. Trendlines from locally weighted smoothing are shown. X axis shows evolutionary time.

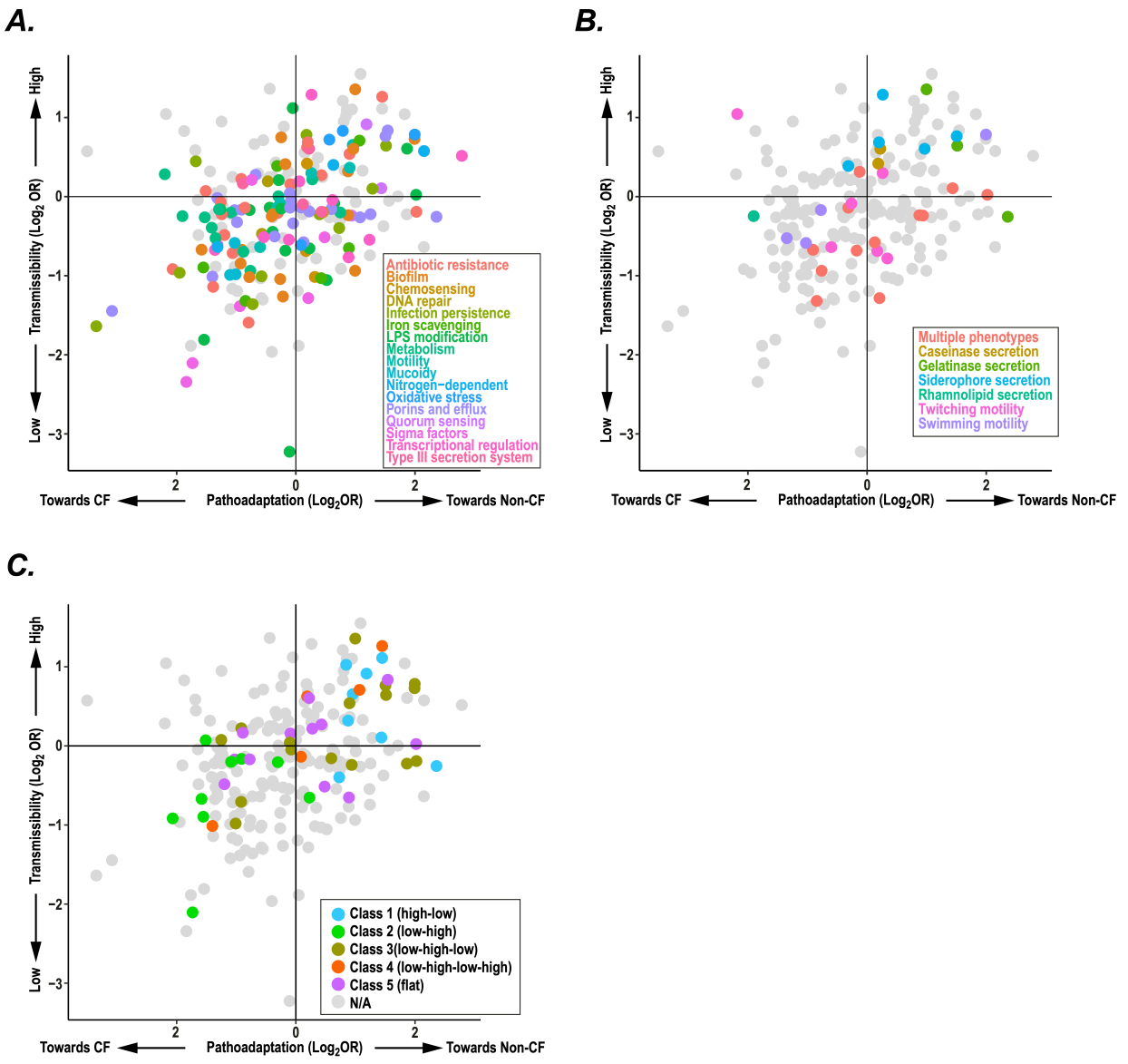


Figure S16. The relative transmissibility and host-specific adaptation of pathoadaptive genes. To estimate host-specific pathoadaptation, the number of cystic fibrosis (CF) vs non-CF mutations (determined by stratifying mutations in pathoadaptive genes on terminal branches by the infection type of isolates) were compared using a Fisher exact test (FDR = 0.1) and expressed as an odds ratio. To assess the transmissibility of pathoadaptation, the number of mutations that had been observed in at least two isolates were compared with mutations that had only been observed once using a Fisher exact test (FDR = 0.1). Genes were colour-coded (A) by established function, (B) based on experimentally-derived functions based on our analyses of corresponding transposon mutants, or (C) by the class of mutation frequency change over evolutionary time.

Table S1. (separate file) Sample information including infection type, isolate clone type, mutational burden, host-specificity and transmissibility test results.

Table S2. (separate file) Gene families gained with COG functional annotations.

Table S3. (separate file) Differentially expressed genes across CF proportions.

Table S4 (separate file) THP1 pooled infection assay DNA sequencing data.

Table S5. (separate file) Gene mutation stats including number of synonymous and non-synonymous SNPs, burden test p-value, host-specificity p-value, transmissibility p-value

Table S6. (separate file) Amino acid variants in LadS, BetT2, PcnA and AlgU.

Table S7. (separate file) Potential transmission links between patients based on a SNP distance threshold of 26 SNPs.