

The diverse trajectories of bacterial heterogeneity and adapted subpopulations in invasive *Staphylococcus aureus* infections

Stefano G. Giulieri

`stefano.giulieri@unimelb.edu.au`

The University of Melbourne

Diane Daniel

The University of Melbourne

Maria Duarte Aguilera

The University of Melbourne

Diana Velasquez-Reyes

The Royal Melbourne Hospital

Romain Guérillot

The University of Melbourne

Ashleigh Hayes

The University of Melbourne

Abderrahman Hachani

The University of Melbourne

Andrew Gador-Whyte

The University of Melbourne

Emma Goodall

Austin Health

Paula Fuge-Larsen

Austin Health

Louise Judd

The University of Melbourne

Jason Trubiano

Austin Health

Marcel Leroi

Austin Health

Katherine Bond

The Royal Melbourne Hospital

Joshua B. Parsons

Duke University School of Medicine

Timothy P. Stinear

The University of Melbourne

Torsten Seemann

The University of Melbourne

Sebastian Duchene

Institut Pasteur

Norelle L. Sherry

The University of Melbourne

Steven Y. C. Tong

The Royal Melbourne Hospital

Jason C. Kwong

The University of Melbourne

Benjamin P. Howden

The University of Melbourne

Research Article

Keywords:

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11122774/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Within-host *Staphylococcus aureus* adaptation can lead to antibiotic treatment failure through the emergence of mutations that cause antibiotic resistance or immune evasion. Yet, given the inherent heterogeneity of bacterial populations, it is not known whether these adaptive mutations emerge at low frequency and subsequently sweep to dominance as infection progresses.

Methods

To assess bacterial within-host evolution during invasive *S. aureus* infections and assess the mutational trajectory of infecting bacterial sub-populations, we collected 669 clinical samples from 103 patients with *S. aureus* bacteraemia or other sterile-site infections and sequenced up to five colonies per sample. Within-patient *S. aureus* lineages were identified based on a threshold of fewer than 100 mutations. We identified adaptive mutations by tracking convergent evolution among lineages.

Results

Nasal swabs displayed significantly higher genetic diversity than clinical samples (median nucleotide diversity 3.6×10^{-7} vs 0). Pathoadaptive mutations involved 16 genes and were found in 26/122 (21%) within-host lineages. Tracking adapted subpopulations revealed divergent trajectories, with a minority of selective sweeps and frequent non-sweeping events or adaptations at low frequency, particularly within the known adaptive locus *agr*. Similar evolutionary trajectories were found when we investigated *de novo* transposon insertions, which were sporadic and contributed marginally to within-host adaptation.

Conclusions

Our findings show that the expansion of the genomic analysis to subpopulations improves the power of within-host evolution studies and reveals unexpected adaptive trajectories that might reflect the complex selective environment of invasive bacterial infections.

Trial registration

ACTRN12623001222639.

Introduction

Staphylococcus aureus infections, ranging from skin infection to pneumonia, sepsis and bone and joint infections, have been linked to > 1 million deaths per year worldwide ¹. This heavy burden is explained by a combination of high prevalence ², high case fatality ratio for the most severe manifestations ³ and antibiotic resistance. Treatment failure is common in severe infections such as *S. aureus* bacteraemia (SAB). Despite prolonged antibiotic treatments and frequent need for source control interventions, microbiologically documented treatment failure (persistent or recurrent infections) occurs in about one in five cases of SAB ⁴. Treatment failure has often been associated with host factors (e.g. diabetes, kidney failure) or treatment issues (e.g. lack of source control). However, the central feature of treatment failure is the ongoing presence of *S. aureus* despite antibiotic treatment, implying that bacterial factors either present at the beginning of the infection or acquired during treatment contribute to treatment failure.

Over the last two decades, sequencing of longitudinal isolates obtained from persistent or recurrent *S. aureus* infections (within-host evolution analysis) has revealed the extent of bacterial adaptive evolution during infection ⁵⁻⁷. These studies have shown that adaptive mutations typically emerge within a network of genes associated with antibiotic resistance ^{8,9}, virulence regulation ¹⁰ and immune evasion ¹¹. In addition to clinical studies, within-host evolution studies of nose-colonising *S. aureus* isolates have revealed distinctive adaptive trajectories ¹², however, the same adaptive mutations observed in invasive isolates can be acquired at the colonisation stage and possibly promote invasiveness ¹³, underscoring the need to collect both infection site and colonisation isolates to capture the full picture of within-host adaptation. While there is robust support for a statistical association between adaptive changes and persistent infection ^{5,7}, it remains unclear whether these pathoadaptive mutations drive treatment failure or are simply selected for in the context of treatment failure that is primarily caused by host factors (e.g. immunosuppression)¹⁴.

While early within-host evolution studies relied on single-colony sequencing, there is now growing evidence of genetic heterogeneity within human samples ¹⁵. By investigating these heterogeneous populations, it is now possible to detect adaptive mutations even when occurring at low frequency. In this study we investigated two hypotheses underlying treatment failure in *S. aureus* infections: 1) that bacterial genetic heterogeneity is associated with treatment outcomes and 2) that adaptive mutations acquired at low frequency before or early during invasive infection become fixed and lead to treatment failure (selective sweeps). To this end we designed BARISTA (Bacterial Adaptation during Invasive *Staphylococcus Aureus* Infections), a large comprehensive clinical-genomic study of invasive *S. aureus* infections. We conceived BARISTA to overcome the limitations of previous within-host evolution studies by including both infection site and colonising samples and capturing bacterial heterogeneity through multi-colony sequencing.

Results

Mapping *S. aureus* within-host evolution and diversity during invasive infections: the BARISTA study

Within-host evolution studies of *S. aureus* infection have relied on opportunistic sampling, with limited information on the colonising population, restricted temporal and anatomical signal, and insufficient sequencing depth to investigate genetic diversity at colonising and infection sites. To overcome these limitations, we tracked *S. aureus* evolution and within-sample heterogeneity through a prospective cohort study of 103 patients hospitalised with invasive (sterile site) infections between 2021 and 2025 (BARISTA, Fig. 1). Median age (interquartile range) was 57 years (47–76); 69 (67%) were male. Eighty-nine patients (86%) were bacteraemic; the number of infectious foci in these bacteraemic episodes ranged from 0 (i.e. bacteraemia without focus) to 5 (median [IQR], 2 [1–2]). We used hierarchical clustering to group individuals with SAB and infection foci (Supplementary Figure S1). Based on the output of this analysis and the literature, we generated broader categories of bacteraemia syndromes, including bone and joint infections (n = 51), intravascular infections (excluding line infections, n = 15) and skin/soft tissue infections (n = 14). Among 14 non-bacteraemic infections, 12 were bone and joint infections. We collected 485 clinical samples (64% blood cultures, 36% other sterile sites) and 184 nasal swabs, over a timespan of up to 10 years (clinical samples) and 1.6 years (nasal swabs) (Fig. 2C-E). Using density-based spatial clustering of applications with noise (DBSCAN) of collection dates, we defined 130 distinct invasive infection episodes; in 13 cases there was at least one recurrence (Supplementary Figure S2). The duration of microbiologically documented episodes ranged from 1 to 34 days (median 2 days, interquartile range 1–6); thirty-two episodes (31%) had a duration of at least 7 days. We therefore classified 39 (38%) cases as refractory-recurrent *S. aureus* infections (RR-SAI) and the remaining 64 as treatment-sensitive *S. aureus* infections (TS-SAI).

Among 1,971 genomes included in BARISTA, 57 belonged to *Staphylococcus argenteus*. The population structure of the *S. aureus* genomes broadly reflected the diversity of circulating clinical *S. aureus* clones, including globally disseminated clones such as sequence type (ST) 2, 30, 45 and 398 (Fig. 1G). Three hundred and sixty-one strains (18.3%) were predicted to be methicillin-resistant *S. aureus* (MRSA) due to the presence of *mecA*, while 388 (19.6%) were penicillin-susceptible (PSSA, *blaZ* negative), and 1,222 (62.0%) were methicillin-susceptible/penicillin-resistant (MSSA). A detailed resistome is listed in Supplementary Table 1. While 45% of nasal swabs grew *Staphylococcus aureus*, we found variable colonisation patterns. Over half (57/101, 56%) of the swabs collected during the initial admission were negative, likely reflecting concurrent antibiotic treatment. Similarly, following discharge, only 39/83 (47%) were positive (Fig. 1E).

Co-infection and re-infection are rare but may be missed without genomic investigations

While evidence from clinical strains¹⁰ and animal models¹⁶ suggests that invasive *S. aureus* infections are clonal (i.e. they are established by a single founding population due to bottleneck effects), there are anecdotal reports of infection with multiple lineages. Similarly, while it is assumed that the transmission bottleneck will establish monoclonal colonisation¹⁷, new transmission events are expected to lead to co- and re-colonisation (i.e. clonal replacement), which can in turn lead to re-infection. Detecting these

clonal dynamics is critical in within-host evolution studies and might have clinical implications, as the management of persistent infection usually differs from those of reinfection¹⁸. Phylogenetic analysis can be used to establish the clonality of a within-host population, however, simpler distance-based methods are easier to implement and can be more accurate¹⁹. First, we performed a fast, low-resolution pairwise analysis of all within-host genomes using Mash²⁰ with a 90% similarity threshold to categorise isolate pairs as genetically unrelated (Fig. 2A). We then called variants pairwise and symmetrically to obtain a precise pairwise mutation count. By combining mapping reads on the draft assembly of an internal reference and removing low-quality variants, our approach yielded a higher accuracy than standard variant calling²¹. This is supported by the observation that pairwise mutation counts were highly concordant ($R^2 = 0.98$, $p < 2.2 \times 10^{-16}$) regardless of comparison direction (isolate 1 vs. isolate 2, and vice versa) (Supplementary Figure S4). We therefore assigned a unique mutation count for each isolate pair. Pairs with no more than 100 mutations were considered genetically related, while pairs with higher mutation counts but > 90% similarity were investigated separately to establish relatedness (Supplementary Figure S5). After this analysis, we were able to identify 122 clusters of genetically related strains (micro-evolved resident populations [MRP] (median 1 per patient, range 1–3, 17% with > 1 MRP). Among 17 individuals with > 1 MRP, almost all cases involved nasal swabs (Fig. 2C), either as co-colonisation (> 1 MRP at the same timepoint, $n = 2$), re-colonisation (new MRP replacing previous MRP, $n = 8$), or dissociation between colonising and invasive strains (distinct MRPs in nasal swab and infection site, $n = 6$). Only one case of co-infection was found. Taken together, this indicates that the co-existence of multiple *S. aureus* lineages is more relevant to colonisation than infection.

Nasal swabs and non-bacteraemic sterile sites have higher within-sample diversity

Using pairwise mutation counts and nucleotide diversity π as markers of genetic diversity²², we found most MRPs had minimal diversity (median within-MRP mutations: 1 [0–3]; median nucleotide diversity $\pi = 3.7 \times 10^{-7}$ [9.4×10^{-8} – 1.6×10^{-6}]). This is consistent with the literature, given short evolutionary times and small effective population sizes²³. Nasal swabs had higher diversity than clinical samples (median pairwise mutation count 3 (IQR 1–8) vs 0 (IQR 0–1); π : 3.6×10^{-7} (0 – 1.3×10^{-6}) vs 0 (0 – 1.2×10^{-7}), $p = 1.6 \times 10^{-8}$), and sterile sites other than blood tended to have higher diversity than blood cultures (π : 0 (0 – 1.2×10^{-7}) vs 0 (0 – 0), $p = 0.06$) (Fig. 2E). While higher genetic diversity in nasal swabs was expected, we were intrigued by the presence of higher mutation counts in some clinical samples, however we didn't find any association between nucleotide diversity π and clinical syndromes or time of collection. Aggregated within-MRP diversity remained stable over time (Supplementary Figure S6). To investigate the impact of the sequencing approach, we generated rarefaction curves of genetic diversity according to the number of colonies sequenced per sample (Fig. 2F). This analysis confirmed the higher genetic diversity of nasal swabs, however, high level of sample coverage could be achieved in nasal swabs by sequencing 10 colonies per sample, while for infection sites samples up to 50 colonies were needed.

Signatures of adaptive evolution are associated with treatment-refractory infections

A key strength of within-host evolution studies is their ability to identify signatures of convergent evolution, as each infection episode can be assumed to be independent provided that it was established by a single founding population (i.e., it is monophyletic). We performed a convergence analysis of 89 MRPs with at least one protein-altering mutation in genes with a homolog in reference genome *S. aureus* FPR3757. We first investigated convergence of amino-acid positions and found a frameshift at position 39 of the LysR-type regulator protein *hutR* independently acquired in 2 MRPs. Similarly, there were convergent frameshifts in genes *gltB* and SAUSA300_2330. All convergent frameshift were in single-sequence repeats (SSRs)²⁴. As convergence at mutation and position level is expected to be weak, we used a commonly employed Poisson model to infer within-host adaptation at gene level^{10,25}. After Bonferroni correction, this analysis revealed two genome-wide significant mutational loci. The highest-ranked gene was *agrA* (five MRPs, $p = 1.34 \times 10^{-6}$) (Fig. 3B). The *agrA* finding was consistent with earlier large-scale within-host evolution studies¹⁰. However, the second-ranked mutated gene was *adh1*, an alcohol dehydrogenase (five MRPs, $p = 1.44 \times 10^{-5}$) that has not been previously linked to within-host adaptation. *Adh1* mutations were found in telavancin resistant mutants²⁶ and *adh1* was upregulated in small-colony variants²⁷, suggesting some mechanisms linking *adh1* mutations to infection persistence.

To overcome sample size limitations of within-host evolution studies, we have recently used a more powerful approach – large-scale within-host evolution, enabling statistically robust inference of adaptive signals beyond the *agr* locus^{5,18}. Outputs of these large datasets can be used to reliably track adaptive mutations in clinical cases of persistent *S. aureus* infections². We leveraged an expanded version of our previously published within-host evolution dataset^{18,28} (1,112 infection episodes, including 5,167 within-host genomes) and identified 25 genome-wide significant loci, 15 of which were also mutated in the BARISTA cohort (Fig. 3C), indicating that BARISTA is a reliable representation of within-host pathoadaptation in *S. aureus*. Beside the already identified *agrA*, the additional 14 genes with adaptive mutations included regulators involved in virulence control (*agrC*: $p = 1.11 \times 10^{-29}$; *rsp*: $p = 8.10 \times 10^{-8}$; *saeR*, $p = 9.65 \times 10^{-7}$), antibiotic resistance genes (*fusA*: $p = 3.50 \times 10^{-9}$; *rpoB*: $p = 3.73 \times 10^{-11}$, *parC*: $p = 8.14 \times 10^{-12}$) and genes linked to both antibiotic resistance/tolerance and metabolic functions (*citZ*: $p = 3.00 \times 10^{-5}$, *relA*, $p = 3.24 \times 10^{-5}$). Taken together, our list included 16 adaptive genes that were found in 26 MRPs: there was not a systematic association between adaptation signatures and specific syndromes, antibiotics, and surgical strategies, however, treatment refractory infections were over-represented (36% with adaptive mutations vs 14% among treatment-sensitive infection, $p = 0.02$).

Most adaptive mutations during infection are non-sweeping

While adaptive evolution has been repeatedly documented in invasive *S. aureus* infections^{5,10,29,30}, especially bacteraemia, the trajectories of mutated sub-populations have not been characterised,

generally because of limited sampling and sequencing. We hypothesised that adaptive mutations might be present at lower frequency at baseline (standing variation)³¹ and become dominant later during the infection (selective sweeps), as seen in within-host evolutionary settings dominated by positive selection. To assess this hypothesis, we investigated invasive and colonising samples separately and defined selective sweeps as a change of mutation prevalence from below 40% at baseline to above 60% at the latest timepoint. Among 978 unique mutations in invasive samples, 336 (35%) were sweeps. The proportion of sweeps was constant regardless of the type of mutation. To identify genes preferentially affected by selective sweeps, we ranked mutated genes based on the number of sweeping and non-sweeping events (sweeping score). We noticed that some of the genes with the most significant mutation enrichment such as *agrA* and *adh1* had a low sweeping score. When we focussed the analysis on genes that were classified as adaptive (see above), we noticed that the proportion of sweeps was higher for non-synonymous substitutions as compared to truncations (Fig. 4AB). This indicates that some adaptive mutations, such as truncations, are less likely to become fixed during infection, possibly due to the fitness cost of more severe loss-of-function mutations.

Within-host acquisition of insertion sequences is sporadic and transient

We have previously shown that structural variants in general, and insertion sequences (*IS*) insertions can play a role in within-host adaptation during invasive *S. aureus* infections^{5,21,32}. Surprisingly, new *IS* events were very rare in our dataset, with just eight insertions in seven MRPs (Fig. 5). Five insertions affected coding regions, including *agrA* and three were intergenic (Fig. 5C-F). In addition, all new *IS* were only found in subpopulations within samples and in two out of five cases in a fraction of the reads, consistent with *IS* being rare, low-frequency occurrences. It is unclear whether *IS* events have an adaptive role or are the consequence of non-adaptive processes like relaxed constraint (i.e. ineffective purifying selection due to small effective population size). Due to the low number of insertions, we were not able to identify convergence. Taken together, our results indicate that *IS* insertions can contribute to within-host evolution, however, their role is secondary compared to point mutations.

Discussion

Investigating within-host bacterial evolution is one of the most effective approaches to untangle the impact of bacterial factors of complex, severe bacterial infections, particularly those with higher risk of persistence and recurrence due to limited treatment success. It might inform prevention and management of treatment failure and provide novel clinically relevant mechanistic insights, that can in turn spark new treatment strategies^{33,34}. This field of genomic research started about 20 years ago with the landmark study by Mwangi et al³⁵, followed by a series of case-studies of persistent infections^{8,9,21,36,37} and has evolved into larger studies over the last decade^{5,10,12,38,39}. Large-scale within-host evolution studies have enabled the identification of convergent evolution, pointing to adaptive hotspots,

however they were limited by opportunistic sampling, insufficient representation of colonising strains and of within-sample diversity^{5,12,38,39}.

To overcome these limitations, we undertook a prospective genomic study of invasive *S. aureus* infections, which showed that within-host and within-sample diversity is limited in clinical samples compared to nasal swabs, however, a subset of blood and sterile site samples showed higher levels of genetic diversity that were not explained by coinfection. We identified genome-wide significant convergent evolution in *agrA* and near-significant convergence in the alcohol dehydrogenase *adh1*. By tracking subpopulations carrying adaptive mutations we identified some selective sweeps, however, counterintuitively, mutations in some highly convergent genes don't sweep and are only found at lower frequency. We describe within-host trajectories in the comparatively rare IS insertions, including an insertion disrupting *agrA*.

Our initial hypothesis was that adaptive mutations might be present at low frequency early during infection and become fixed later, driving treatment failure. These selective sweeps have been previously shown in within-host evolution studies. For example, in a ciprofloxacin challenge study, *gyrA* mutations associated with fluoroquinolone resistance swept under antibiotic pressure⁴⁰. Similar sweeps were observed in persistent *Mycobacterium tuberculosis* infections²² or in HIV treatment failure⁴¹. We also hypothesised that baseline or early within-host and within-sample diversity could be associated with infection chronicity, infection syndromes or outcomes. This hypothesis builds on phylodynamic infection models, showing that genetic diversity correlates with the time of acquisition of the infection³⁷ and might indicate antibiotic-induced differentiation¹³.

To assess these hypotheses, we augmented our capacity to detect bacterial sub-populations at a given timepoint by using all available clinical samples (including multiple blood culture bottles) and by applying multi-colony sequencing. We showed that within-host and within-sample diversity is limited in clinical samples compared to nasal swabs, however, a small number of blood and sterile site samples showed higher levels of heterogeneity that were not explained by coinfection. The higher genetic diversity of nasal swabs is consistent with previous reports⁴². First, we found evidence of co-colonisation and lineage replacement in longitudinally collected nasal carriage samples. This clonal diversity is likely underestimated given that the limited number of swabs and the lack of other colonising sites. On the background of this clonal interference, we observed within-clone diversification with up to 152 mutations in some samples. Genetic diversity of clinical samples such as blood cultures and surgical specimens has not been extensively investigated. Animal experiments have indicated that there is a strong bottleneck at invasion and when organs are seeded haematogenously¹⁶; together with small effective population sizes and short evolutionary times²³, these circumstances are expected to limit within-host genetic diversity. Our study confirms these data but with some notable outlier samples and episodes harbouring heterogeneous populations. We have previously suggested that within-host diversity in *S. aureus* bacteraemia can reflect distinctive treatment responses²¹. Here, we build on these early observations and show that higher diversity can be observed in some cases of invasive infections. This

is in line with a report of heterogenous infecting *Staphylococcus epidermidis* populations in prosthetic-joint infections ⁴³.

By increasing the number of samples and sequenced isolates, we show that the power of within-host convergent evolution can be further increased. In our analysis, we identified genome-wide significant convergent evolution in *agrA* (which was expected ^{5,13}). Surprisingly, the second-ranked gene was the alcohol dehydrogenase *adh1*, indicating that our study had the capacity to identify new adaptive hotspots. This finding is consistent with reports of the association between *adh1* mutations and telavancin resistance ²⁶ and *adh1* upregulation in small colony variants ²⁷. A potential hypothesis for this role of *adh1* can be found in a *Streptococcus pneumoniae* study that investigated the role of the alcohol dehydrogenase in regulating bacterial virulence in invasive infections ⁴⁴.

Despite the effect of multi-sample and multi-colony sequencing, the power of our analysis remained limited by the number of episodes included, with only 2 genes reaching statistical significance. We therefore leveraged an expanded version of our genomic meta-analysis of *S. aureus* within-host evolution ^{5,18,28} to identify an additional set of 14 genes that were significantly mutated in the meta-analysis and found in BARISTA. We then tracked the fate of these adaptations at sub-population level. In accordance with our initial hypothesis, we found that subpopulations carrying adaptive mutations can sometimes sweep, as we would expect under selective pressure; however, surprisingly, mutations in some highly convergent genes like the *agr* locus never sweep and are only found at lower frequency. In general, non-sweeping adaptation were more likely to be truncations and affect regulatory genes. This pattern can be explained by non-adaptive mechanisms that are expected to occur within-host, such as relaxed constraint ⁴⁵ (less effective purifying selection related to bottleneck effects and small effective population size). However, the fact that it occurred in the *agr* locus which has strongest evidence for within-host adaptations, suggests that the non-sweeping trajectories reflect changing selective pressures in the complex landscape of clinical infections. This phenomenon has been labelled 'adaptive tracking' and it depicts a situation where apparent neutral evolution (including non-sweeping mutations) hides beneficial mutations that don't become fixed due to changing environmental pressures and antagonistic pleiotropy ^{24,46}. While adaptive tracking offers an intriguing hypothesis to explain adaptive mutations trajectories in within-host evolution, our data don't provide enough evidence to confirm it. For example, we assessed whether the fluctuating prevalence of adaptative mutations could be linked to antibiotic exposures or anatomical niches, but could not find strong correlations. This might be related to insufficient sampling or to more complex interactions that can't be inferred from simple exploratory associations. For example, it has been shown that *agr* mutants undergo specific evolutionary trajectories in invasive infections ⁴⁷, possibly indicating a constant fine tuning between *S. aureus* phenotypes that can't be detected from opportunistically collected clinical samples.

Taken together, these observations offer new insights into the pathogenesis of *S. aureus* infections and its evolutionary correlates, however, it remains unclear whether the detection of adaptive mutations at low frequency has clinical implications. This question is important, because low-frequency mutations are

likely to be missed by standard whole-genome sequencing. In virology, where deep sequencing has been performed for more than a decade, selective sweeps have been anecdotally reported in treatment failure, however, larger studies didn't show an association between the detection of adaptive mutations (e.g. drug-resistant variants) and risk of treatment failure, suggesting that antimicrobial treatment is effective even if there is a resistant sub-population.

We have previously shown that structural variants, and IS insertions in particular, can emerge in invasive infections at a much higher rate than during nasal colonisation^{5,21}. The role of IS insertions in bacterial micro-evolution has been documented in large-scale studies of *Enterococcus faecium*⁴⁸ and *Bacteroides fragilis*⁴⁹, however, it remains unclear whether these IS accumulations promote adaptive evolution in the host bacteria or are rather amplifications without benefit for the host ("selfish genes"). Here we found that there was a limited role of new IS events. This might be due to different sequence types and less active IS classes⁵. Similarly to truncations, IS insertions were not sweeping and in fact sometimes found in subpopulations within single colonies. The analogy with truncating mutations was striking, given that one low-frequency IS insertion was found in *agrA*. Given the low number of IS insertions we couldn't use between-host convergence to assess whether these insertions were adaptive, however, the location of some insertions suggest that at least some of them are adaptive, as previously shown for IS insertions in the *walkR* promoter leading to vancomycin resistance^{21,50} or IS insertions disrupting *mecA*⁵¹ or *agrC*⁵.

Our study has limitations. We sequenced 5 colonies per sample, which can only detect subpopulations at around 20% prevalence and thus miss much rarer variants. Other within-host evolution studies have sequenced a slightly higher number of colonies⁵², however, in our study, the limited number of sequenced colonies was partially compensated by the availability of multiple samples at a given timepoint and for each anatomical niche, thus lowering the threshold for variants detection per timepoint to as low as 2.5%. Another limitation of multiple colony sequencing is the bias towards more rapidly growing populations. This could be overcome by deep sequencing of plate sweeps and ideally of the clinical samples^{53,54}. Of note, using deep sequencing of plates sweeps, Mills *et al* didn't find an increased sensitivity for convergent evolution compared to the multi-colony sequencing approach⁵⁴. In addition, deep sequencing might have higher false positive rates due to mapping errors⁵³ and complicates strain-level resolution and the construction of within-host phylogenies. Another limitation is that we relied on short-read sequencing. Long-read sequencing has better performance in calling structural variants, however, work done by us²¹ and others^{48,49,55} indicates that IS insertions from short read data is robust. Finally, our sampling relied on availability of clinical specimens and on the patient to return to the same hospital in case of recurrent infections. This might have limited our power to detect adaptation in relapsing *S. aureus* infections, a key factor driving within-host adaptation⁶. Nasal colonisation investigations were also limited by the inability to collect nasal swabs early during infection (ideally at the start of antibiotic treatment), a relative lack of sampling during follow-up and negative

cultures from half of the swabs. Metagenomics might increase the yield of these samples, although it remains unclear if the small amount of *S. aureus* DNA recovered is sufficient to call variants accurately.

In conclusion, our study BARISTA provides a comprehensive investigation of the role of *S. aureus* subpopulations in shaping the course of invasive infections. We show that within-host diversity is low but can reach high levels in some clinical samples. Whether within-host diversity is associated with infection syndromes or outcomes remains to be determined. We track the fate of subpopulations carrying adaptive mutations and demonstrate that adaptations can be found at low level exclusively and be missed by standard whole-genome sequencing. Future studies should assess the accuracy and yield of more efficient deep sequencing methods, such as plate sweeps sequencing or direct metagenomic sequencing from clinical samples. While these types of clinical studies rely on opportunistic samples with its inherent weaknesses, robust prospective protocols and a multi-centric approach could increase the statistical power to identify associations between clinical outcomes and low-frequency variants, with potential impact on precision management of invasive *S. aureus* infections.

Methods

Ethics approval

The BARISTA study protocol was approved by the Austin Health Human Research Ethics Committee on 3 March 2021 (Reference Number HREC/64786/Austin-2020). Governance approval was obtained at all sites involved. Signed consent was obtained from patients or medical decision makers if applicable. The study was registered in the Australian New Zealand Clinical Trials Registry (ANZCTR) on 28 November 2023 (Registration number ACTRN12623001222639).

Study design

BARISTA was a prospective cohort study of invasive *S. aureus* infections, carried out at two teaching hospitals in Victoria, Australia, between April 2021 and June 2025 (Fig. 1A). Participants older than 18 were enrolled if they were admitted for a sterile site *S. aureus* infection (i.e. at least two culture-positive sterile samples, including blood culture, joint / pleural / peritoneal / cerebrospinal fluid, tissue biopsy, or abscess fluid). Exclusion criteria were polymicrobial infection at time of inclusion (i.e. presence of a bacterial species other than *S. aureus* in any sterile-site sample, excluding species thought to be contaminants) and death deemed to be imminent within the following 48 hours. Upon inclusion, a nasal swab was collected by a study staff member, and the hospital lab was requested to keep all clinical samples with growth of *S. aureus*. Participants were contacted by mail at 6 and 12 months after inclusion and asked to self-collect a nasal swab and to fill in a questionnaire with queries about intercurrent hospitalisations, infections and antibiotic therapies. If participants were re-admitted to the participating hospital for a recurrent sterile-site *S. aureus* infection, the same sample collection procedure described for the initial episode was repeated. If they declared to have been admitted to another institution for a *S. aureus* infection, the external hospital laboratory was contacted and asked to provide stored *S. aureus* isolates for the infection episode.

Infection syndrome classification

For each infection episode at enrolment, we collected on information on any infection site at diagnosis (i.e. within 48 h after admission) and at discharge. Infection sites included following categories: Bacteraemia without focus, Infective endocarditis, Vascular infection, Vertebral osteomyelitis, Septic arthritis, Osteomyelitis other than vertebral, Epidural abscess, Skin infection, Myositis/muscle abscess, Respiratory/chest infection, Intra-abdominal infection, Urinary tract infection, CNS infection. Device infections were further classified in Prosthetic-valve endocarditis, Intracardiac-device infection, Vascular graft infection, Vascular line infection (including PermCath), Prosthetic joint infection, Fracture-fixation device infection (including spondylodesis). Following higher-level categories were defined: intravascular infections (Infective endocarditis, Intracardiac-device infection, Prosthetic-valve endocarditis, Vascular graft infection, Vascular infection), bone and joint infections (Septic arthritis, Epidural abscess, Vertebral osteomyelitis, Osteomyelitis other than vertebral, Prosthetic joint infection, Myositis/muscle abscess), Skin infection and Line infection. Infection sites were clustered and visualised using the heatmap() function in R.

Treatment outcome attribution

We used the collected microbiological data to classify each participant in “refractory/recurrent *S. aureus* infection” (RR-SAI) and “treatment-sensitive *S. aureus* infection” (TS-SAI). For each participant, we clustered the collection dates of clinical samples with growth of *S. aureus* using DBSCAN algorithm (Density-based spatial clustering of applications with noise), as implemented in the R package dbSCAN, v1.2.4⁵⁶, with the *e* (eps) parameter set to 2 and minimum number of points required to form a dense region at 3. We reviewed manually individuals with multiple clusters and those with clusters stretching over more than 15 days. For 19 of these 20 individuals the manual correction involved grouping samples that the DBSCAN algorithm assigned to singletons, while in one case we split a cluster in seven smaller clusters based on the clinical data. Persistent infection clusters were defined by a duration of 7 days or longer. Individuals with RR-SAI were defined by the presence of a persistence infection cluster and/or more than one cluster containing sterile-site samples with growth of *S. aureus*, all other individuals were categorised as TS-SAI.

Sample collection

Nasal swabs at enrolment were collected by inserting a sterile MRSA swab (Copan eSwab, Italy) up to 2.5 cm into both nasal cavities and rotating for 5 seconds. Swabs were stored at 4°C and shipped to the study lab within a week. Native clinical samples with growth of *S. aureus* were stored at 4°C or at room temperature (blood cultures bottles) for up to 2 months before shipment to the study laboratory. Where available, primary plates with growth of *S. aureus* were also stored. If no clinical samples were available, a single isolate stored by the clinical laboratory at -80°C was subcultured onto horse blood agar (HBA) and shipped to the study laboratory.

Sample processing

All samples apart from EDTA blood tubes were processed in the Howden research laboratory at the Doherty Institute in Melbourne, Australia. Nasal swabs were inoculated onto HBA and mannitol-salt plates (MSA). Presence of *S. aureus* was confirmed by, first, growth of yellow colonies on MSA and then using MALDI-TOF. If no colonies grew overnight, the nasal swab liquid was inoculated into brain-heart infusion (BHI) broth. Using a syringe, a drop of the blood culture bottles or liquid clinical samples was inoculated onto HBA plates and incubated at 37°C for 24 hours. Tissue samples were suspended in 5 ml NaCl 0.9% before inoculation. Up to 5 colonies per plate were submitted for bacterial whole-genome sequencing (WGS). Sequenced colonies and the plate sweeps were stored in 20% glycerol storage medium at -80°C. Native clinical fluids and suspended tissue samples were stored in Zymo buffer at -80°C.

Bacterial whole-genome sequencing

Bacterial whole genome sequencing (WGS) was performed at the Microbiological Diagnostic Unit Public Health Laboratory, Melbourne, using a NATA-accredited genomics workflow. DNA extraction was carried out from single colonies placed in 100µl of Gram-positive lysis buffer (20mM Tris-HCL, pH 8.0, 2mM EDTA, 1.2% TritonX-100) on the QIAAsymphony using the QIAAsymphony DSP Virus/Pathogen Mini kit protocol (QIAGEN, Germany), aiming at a DNA concentration of 1-15ng/µl. The Nextera XT (Illumina) was used for library preparation. Libraries were sequenced on the NextSeq 500/550 or MiSeq platforms (Illumina) to generate 150 bp paired-end reads. The short reads are available on ENA with the study accession PRJEB127027.

Quality control and genome assembly

Sequence quality was assessed by calculating coverage metrics obtaining quality data using seqtk v1.4-r122 (<https://github.com/lh3/seqtk>) and classifying reads using kraken2, v2.1.1⁵⁷, Reads were assembled using shovill, v1.0.1⁵⁸ and spades, v3.14.1⁵⁹, and annotated with prokka, v1.14.6⁶⁰. Sequencing was repeated if the quality control (QC) criteria were not met: Q-score ≥ 30 , average reads depth $\geq 40x$, dominant species other than *Staphylococcus aureus* or *Staphylococcus argenteus* in the Kraken report, and draft assembly genome size outside the expected range for *S. aureus* (2.6–3.2 Mb). Sequence types and resistance genes were inferred from the assemblies using Mlst v2.19.0 (<https://github.com/tseemann/mlst>), and Abricate v1.0.1⁶¹ (using the NCBI AMRFinderPlus database⁶²), respectively.

Phylogenetic analysis

Illumina reads were aligned to reference methicillin-susceptible *S. aureus* genome BPH2819 (accession GCF_900620225.1)⁶³ using Snippy v4.6.0 (<https://github.com/tseemann/snippy>). We generated a core genome alignment with all reference positions with $\geq 90\%$ non-missing information (i.e. ≥ 10 coverage depth) using Snippy-core v4.6.0, Goalign v0.3.4⁶⁴ and SNP-sites v2.5.1⁶⁵. We then built a maximum-likelihood phylogenetic tree with IQ-TREE, v3.1.1⁶⁶ using a GTR+G4 model and 1000 bootstraps. The phylogenetic tree and related metadata were processed in R using the ggtree package, v4.0.1⁶⁷.

Within-host and within-sample genetic distance

The pairwise genetic distance between all genomes within each individual participant was first estimated from the draft assemblies using Mash, v2.3²⁰ and a sketch size of 10,000. Pairs with a distance below 9,000/10,000 were retained for the variant calling step. Pairwise variant calling was performed as described in⁶⁸. We used Snippy to map the reads of both isolates to the draft assembly of the first isolate, with the procedure repeated on the symmetric pair (i.e. mapping reads on the draft assembly of the second isolate). For each pair, we excluded variants called from the reference isolate's own reads, as well as variants at positions where the reference had a coverage depth < 10. This filtering followed the procedure described in⁵ and using scripts available at https://github.com/stefanogg/staph_adaptation_paper. A 2-step approach was used to define micro-evolved resident populations (MRP): 1) For each non-symmetric pair, we computed an average mutational distance and included non-symmetric pairs if they had a mean distance ≤ 100 mutations. Pairs with > 1000 mutations were excluded and pairs with 100–1000 mutations were included if visual inspection indicated that the distance distribution was homogenous. 2) Included pairs were grouped using the tidygraph package in R.

Variants annotation

Amino-acid sequences of the mutated coding regions were clustered using CD-HIT, v4.8.1⁶⁹, with default setting. We searched *S. aureus* reference genome FPR3757 for homologs of each cluster representative using Blastp, v2.16.0 +⁷⁰ with an E-value of 10^{-5} , identity filter at 90% and coverage filter at 50%. FPR3757 homologs were annotated using the Aureowiki database⁷¹.

Between-host convergence

For each CD-HIT cluster we computed the number of independent mutations, defined as the number of MRPs with an individual mutation, mutated position or mutated gene. To infer the statistical significance of between host convergence at gene level, we counted the number of MRPs with a non-synonymous substitution or truncating mutation in each FPR3757 homolog (excluding phage and plasmid genes). These counts and the length of each gene were used to build 2 Poisson models of mutation counts as reported in¹⁰: a model 0, with uniform mutation rate for each gene (i.e. neutral evolution) and a gene-specific model based on the actual mutation counts. Likelihood ratio was used to compare the two models and infer statistical significance. Thus, genes with the lowest p value were considered hotspots of between-host convergent evolution, given the extremely low probability that the observed mutation count is due to neutral evolution. Bonferroni correction was applied.

Within-host evolution meta-analysis

To infer adaptive evolution at larger scale, we used an updated version of our curated collection of within-host genomes from *S. aureus* infection and colonisation episodes that was previously reported^{5,18,28}, including 5,167 genomes and 1,112 episodes from 32 studies (see Supplementary Table 2). We

extracted independent mutation counts for each FPR3757 homolog as described above and applied Poisson models to infer significance of between-host convergent evolution.

Analysis of bacterial sub-populations

Nucleotide diversity π was estimated as in ²² but with a simplified approach, since it could be computed from variant calls rather than raw reads. For each mutated site, the variable π was calculated using the formula: $D = 2p(1 - p)$, where p is the alternative allele fraction within the sample, timepoint or MRP. The sum of D was divided by the number of sites (i.e. all reference sites, excluding positions with < 10 reads coverage). Rarefaction curves were created using the R package iNEXT, v3.0.2 ⁷² using aggregated point mutation data for all samples with at least 5 colonies sequenced, and computing the incidence frequency as the number of sequenced colonies per sample carrying each mutation.

Selective sweeps

To investigate the dynamics of mutated subpopulations we selected MRPs with at least three collection timepoints or two timepoints and at least 5 genomes per timepoint. We assessed infection site samples and nasal swabs separately. Selective sweeps were defined by an allele fraction below or equal to 40% at the earliest timepoint and an allele fraction above or equal to 60% at the last timepoint. For each mutated gene with a FPR3757 homolog we calculate a sweep score as the number of sweeping mutations minus the number of non-sweeping mutations.

Detection of de novo IS insertions

To detect new insertion sequences (IS) events within MRPs we applied a modified version of the split-reads approach described in ²¹ and ⁵. Using BWA-MEM ⁷³, all short reads belonging to the same MRP were mapped to the closest available complete genome (based on Mash distance). Reads coverage was assessed using the BEDTools suite, v2.31.1 ⁷⁴. Split reads were extracted from the BAM files using a Python script from Lumpy ⁷⁵. We first extracted split reads breakpoints (junctions) in Python using Samtools mpileup, v1.20 ⁷⁶; candidate insertion sites were obtained using BEDTools to identify regions of no more than 10 bp length with left-sided breakpoint followed by a right-sided breakpoint. We then processed the candidate insertion sites in a separate Python script to extract the supplementary alignment for each split read within the candidate region and annotate both the candidate region and the supplementary alignment using the GFF file of the mapping reference. The candidate insertion sites, annotated candidate insertions and supplementary alignment and reads coverage data were processed in R. IS insertion sites were confirmed if: 1) at least one supplementary alignment region overlapped with a transposase; 2) the ratio between split reads at breakpoint and total split reads coverage was ≥ 0.75 ; 3) only 2 breakpoints were detected in the candidate region. To infer *de novo* IS insertions, we removed insertion sites that were found in the MRP-specific internal reference. We used ggplot2, v3.2.1 ⁷⁷ to inspect the remaining sites manually by comparing total reads coverage and split reads coverage for the internal reference and all remaining MRP genomes over a region spanning 1,000 bp upstream and

downstream. Sites were confirmed if the internal reference displayed zero split reads coverage but conserved total reads coverage.

Declarations

Funding

This work was supported by a University of Melbourne Early Career Researcher Grant (to S.G.G.) and the National Health and Medical Research Council (NHMRC), grant number GNT1196103 (to B.P.H.).

Author Contribution

S.G.G., R.G., T.P.S., B.P.H conceived the study. S.G.G., D.D., A.H., J.C.K, B.P.H. wrote the protocol. S.G.G., D.V.R., A.G.W., E.G., J.T., S.Y.C.T., J.C.K., recruited study participants. S.G.G., D.D, M.D.A, D.W.R., P.F.L., M.L., K.B. collected study samples. D.D, M.D.A, L.J. , N.L.S. performed whole-genome sequencing. T.S. provided bioinformatic tools. J.P.D., T.S., S.D., J.C.K. contributed intellectual input. S.G.G. analysed the data. S.G.G. wrote the manuscript. All authors reviewed the manuscript.

Acknowledgement

We acknowledge all the patients and healthcare providers who were involved in the BARISTA study

Data Availability

All data supporting the findings of this study are available within the paper and its Supplementary Information, except for clinical and demographic characteristics of the study participants, which are available upon request to protect the participants' anonymity. All sequencing data were deposited into the ENA database under accession number PRJEB127027.

References

1. Ikuta KS, et al. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2022;400:2221–48.
2. Imam N, et al. Increased incidence of community-associated *Staphylococcus aureus* bloodstream infections in Victoria and Western Australia, 2011–2016. *Med J Aust*. 2019;210:87–8.
3. Bavaro DF, et al. Anti–Methicillin-Resistant *Staphylococcus aureus* (MRSA) Cephalosporins Plus Daptomycin as Initial Therapy for MRSA Bacteremia: Does a Hit Hard and Fast Strategy Improve Outcomes? *Clin Infect Dis*. 2026;ciag228. 10.1093/cid/ciag228.

4. Choi S-H, et al. Risk Factors for Recurrent *Staphylococcus aureus* Bacteremia. *Clin Infect Dis*. 2021;72:1891–9.
5. Giulieri SG, et al. Niche-specific genome degradation and convergent evolution shaping *Staphylococcus aureus* adaptation during severe infections. *eLife*. 2022;11:e77195.
6. Talbot BM, et al. Methicillin-Resistant *Staphylococcus aureus* Bacteremia Relapses Show Diverse Genomic Profiles but Convergence in Bacteremia-Associated Genes. *J Infect Dis*. 2025;232:1338–50.
7. Elgrail MM et al. Convergent Evolution of Antibiotic Tolerance in Patients with Persistent Methicillin-Resistant *Staphylococcus aureus* Bacteremia. *Infect Immun* 0, e00001–22.
8. Howden BP, et al. Evolution of Multidrug Resistance during *Staphylococcus aureus* Infection Involves Mutation of the Essential Two Component Regulator WalKR. *PLoS Pathog*. 2011;7:e1002359.
9. Peleg AY, et al. Whole Genome Characterization of the Mechanisms of Daptomycin Resistance in Clinical and Laboratory Derived Isolates of *Staphylococcus aureus*. *PLoS ONE*. 2012;7:e28316.
10. Young BC, et al. Severe infections emerge from commensal bacteria by adaptive evolution. *eLife*. 2017;6:e30637.
11. Gao W, et al. The RpoB H481Y Rifampicin Resistance Mutation and an Active Stringent Response Reduce Virulence and Increase Resistance to Innate Immune Responses in *Staphylococcus aureus*. *J Infect Dis*. 2013;207:929–39.
12. Coll F, et al. The mutational landscape of *Staphylococcus aureus* during colonisation. *Nat Commun*. 2025;16:302.
13. Young BC et al. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc. Natl. Acad. Sci. U.S.A.* 109, 4550–4555 (2012).
14. Shepherd MJ, et al. Ecological and evolutionary mechanisms driving within-patient emergence of antimicrobial resistance. *Nat Rev Microbiol*. 2024;1–16. 10.1038/s41579-024-01041-1.
15. Cheng S, et al. Within-Host Genotypic and Phenotypic Diversity of Contemporaneous Carbapenem-Resistant *Klebsiella pneumoniae* from Blood Cultures of Patients with Bacteremia. *mBio*. 2022;13:e02906–22.
16. Pollitt EJG, Szkuta PT, Burns N, Foster SJ. *Staphylococcus aureus* infection dynamics. *PLoS Pathog*. 2018;14:e1007112.
17. Worby CJ, Lipsitch M, Hanage WP. Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. *PLoS Comput Biol*. 2014;10:e1003549.
18. Giulieri SG, et al. A multi-hospital, clinician-initiated bacterial genomics programme to investigate treatment failure in severe *Staphylococcus aureus* infections. *Nat Commun*. 2025;16:4869.
19. Higgs C, et al. Optimising genomic approaches for identifying vancomycin-resistant *Enterococcus faecium* transmission in healthcare settings. *Nat Commun*. 2022;13:509.

20. Ondov BD, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 2016;17:132.
21. Giulieri SG, et al. Genomic exploration of sequential clinical isolates reveals a distinctive molecular signature of persistent *Staphylococcus aureus* bacteraemia. *Genome Med.* 2018;10:65.
22. Nimmo C, et al. Dynamics of within-host *Mycobacterium tuberculosis* diversity and heteroresistance during treatment. *eBioMedicine.* 2020;55:102747.
23. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. *Nat Rev Microbiol.* 2016;14:150–62.
24. Guérillot R et al. Simple sequence repeats power < em>*Staphylococcus aureus* adaptation. *bioRxiv* 2025.07.08.663602 (2025) 10.1101/2025.07.08.663602
25. Lieberman TD, et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat Genet.* 2011;43:1275–80.
26. Roch M, Varela MC, Taglialegna A, Rose WE, Rosato AE. Activity of Telavancin against *Staphylococcus aureus* Isolates, Including Those with Decreased Susceptibility to Ceftaroline, from Cystic Fibrosis Patients. *Antimicrob Agents Chemother.* 2018;62:e00956–18.
27. Cao S, Huseby DL, Brandis G, Hughes D. Alternative Evolutionary Pathways for Drug-Resistant Small Colony Variant Mutants in *Staphylococcus aureus*. *mBio.* 2017;8. 10.1128/mbio.00358 - 17.
28. Chen Y-T, et al. Regulation of airway fumarate by host and pathogen promotes *Staphylococcus aureus* pneumonia. *Nat Commun.* 2025;16:7050.
29. Long DR, et al. Polyclonality, Shared Strains, and Convergent Evolution in Chronic Cystic Fibrosis *Staphylococcus aureus* Airway Infection. *Am J Respir Crit Care Med.* 2021;203:1127–37.
30. Key FM, et al. On-person adaptive evolution of *Staphylococcus aureus* during treatment for atopic dermatitis. *Cell Host Microbe.* 2023;31:593–e6037.
31. Orr HA. The genetic theory of adaptation: a brief history. *Nat Rev Genet.* 2005;6:119–27.
32. Guérillot R et al. Unstable chromosome rearrangements in *Staphylococcus aureus* cause phenotype switching associated with persistent infections. *Proceedings of the National Academy of Sciences* 116, 20135–20140 (2019).
33. Culyba MJ, Tyne DV. Bacterial evolution during human infection: Adapt and live or adapt and die. *PLoS Pathog.* 2021;17:e1009872.
34. Tonkin-Hill G, Ruis C, Bentley SD, Lythgoe KA, Bryant JM. Within-host bacterial evolution and the emergence of pathogenicity. *Nat Microbiol.* 2025;10:1829–40.
35. Mwangi MM et al. Tracking the in vivo evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. *Proceedings of the National Academy of Sciences* 104, 9451–9456 (2007).
36. Wüthrich D, et al. Genomic characterization of inpatient evolution of MRSA resistant to daptomycin, vancomycin and ceftaroline. *J Antimicrob Chemother.* 2019;74:1452–4.

37. Klemm EJ, et al. Emergence of host-adapted *Salmonella* Enteritidis through rapid evolution in an immunocompromised host. *Nat Microbiol.* 2016;1:1–6.
38. Vargas R, et al. In-host population dynamics of *Mycobacterium tuberculosis* complex during active disease. *eLife.* 2021;10:e61805.
39. Gatt YE, Margalit H. Common Adaptive Strategies Underlie Within-Host Evolution of Bacterial Pathogens. *Mol Biol Evol.* 2021;38:1101–21.
40. Yaffe E et al. Brief antibiotic use drives human gut bacteria towards low-cost resistance. *Nature* 1–10 (2025) 10.1038/s41586-025-08781-x
41. Metzner KJ, et al. Minority Quasispecies of Drug-Resistant HIV-1 That Lead to Early Therapy Failure in Treatment-Naive and -Adherent Patients. *CLIN INFECT DIS.* 2009;48:239–47.
42. Stanczak-Mrozek KI, et al. Within-host diversity of MRSA antimicrobial resistances. *J Antimicrob Chemother.* 2015;70:2191–8.
43. Widerström M, et al. Heterogeneity of *Staphylococcus epidermidis* in prosthetic joint infections: time to reevaluate microbiological criteria? *Eur J Clin Microbiol Infect Dis.* 2022;41:87–97.
44. Im H, et al. Targeting NAD⁺ regeneration enhances antibiotic susceptibility of *Streptococcus pneumoniae* during invasive disease. *PLoS Biol.* 2023;21:e3002020.
45. Wertheim JO, Murrell B, Smith MD, Pond K, S. L., Scheffler K. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. *Mol Biol Evol.* 2015;32:820–32.
46. Song S, Chen P, Shen X, Zhang J. Adaptive tracking with antagonistic pleiotropy results in seemingly neutral molecular evolution. *Nat Ecol Evol.* 2025. <https://doi.org/10.1038/s41559-025-02887-1>.
47. Altman Deena R, et al. Genome Plasticity of *agr*-Defective *Staphylococcus aureus* during Clinical Infection. *Infect Immun.* 2018;86. 10.1128/iai.00331 - 18.
48. Grieshop MP, et al. Transposable elements are driving rapid adaptation of *Enterococcus faecium*. *Nature.* 2026. <https://doi.org/10.1038/s41586-026-10373-2>.
49. Kirsch JM, Hryckowian AJ, Duerkop BA. A metagenomics pipeline reveals insertion sequence-driven evolution of the microbiota. *Cell Host Microbe.* 2024;32:739–e7544.
50. McEvoy CRE, et al. Decreased Vancomycin Susceptibility in *Staphylococcus aureus* Caused by IS 256 Tempering of *WalKR* Expression. *Antimicrob Agents Chemother.* 2013;57:3240–9.
51. Proulx MK, et al. Reversion From Methicillin Susceptibility to Methicillin Resistance in *Staphylococcus aureus* During Treatment of Bacteremia. *J Infect Dis.* 2016;213:1041–8.
52. Conwill A, et al. Anatomy promotes neutral coexistence of strains in the human skin microbiome. *Cell Host Microbe.* 2022;30:171–e1827.
53. Jago MJ, et al. High-throughput method characterizes hundreds of previously unknown antibiotic resistance mutations. *Nat Commun.* 2025;16:780.
54. Mills EG et al. High-depth whole genome sequencing of blood culture plates reveals evolutionary dynamics in cases of persistent bacteremia due to methicillin-resistant *Staphylococcus aureus*. 2026.04.25.720776 Preprint at <https://doi.org/10.64898/2026.04.25.720776> (2026).

55. Hawkey J, Monk JM, Billman-Jacobe H, Palsson B, Holt KE. Impact of insertion sequences on convergent evolution of *Shigella* species. *PLoS Genet.* 2020;16:e1008931.
56. Hahsler M, Piekenbrock M, Doran Ddbscan. Fast Density-Based Clustering with R. *J Stat Softw.* 2019;91:1–30.
57. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 2019;20:257.
58. Seemann T, Shovill. Robust Microbial Short-Read Genome Assembly. *Preprints* <https://doi.org/10.20944/preprints202607.1422.v1> (2026) doi:10.20944/preprints202607.1422.v1.
59. Bankevich A, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol.* 2012;19:455–77.
60. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30:2068–9.
61. Seemann T, Abricate. Mass Screening of Bacterial Genome Assemblies Against Curated Gene Sequence Databases. *Preprints* <https://doi.org/10.20944/preprints202607.1215.v1> (2026) doi:10.20944/preprints202607.1215.v1.
62. Feldgarden M, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci Rep.* 2021;11:12728.
63. Mu A, et al. Integrative omics identifies conserved and pathogen-specific responses of sepsis-causing bacteria. *Nat Commun.* 2023;14:1530.
64. Lemoine F, Gascuel O. Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. *NAR Genomics Bioinf.* 2021;3:lqab075.
65. Page AJ, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genomics.* 2016;2:e000056.
66. Wong TKF, et al. IQ-TREE 3: phylogenomic inference software using complex evolutionary models. *Mol Biol Evol.* 2026;43:msag117.
67. Yu G, Lam TT-Y, Zhu H, Guan Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol Biol Evol.* 2018;35:3041–3.
68. Hachani A et al. A high-throughput cytotoxicity screening platform reveals agr-independent mutations in bacteraemia-associated *Staphylococcus aureus* that promote intracellular persistence. *eLife* 12, e84778 (2023).
69. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* 2012;28:3150–2.
70. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
71. Fuchs S, et al. AureoWiki – The repository of the *Staphylococcus aureus* research and annotation community. *Int J Med Microbiol.* 2018;308:558–68.

72. Hsieh TC, Ma KH, Chao A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods Ecol Evol.* 2016;7:1451–6.
73. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
74. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841–2.
75. Layer RM, Chiang C, Quinlan AR, Hall I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* 2014;15:R84.
76. Li H, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25:2078–9.
77. Valero-Mora PM. ggplot2: Elegant Graphics for Data Analysis. *J Stat Soft.* 2010;35:1–3.

Figures

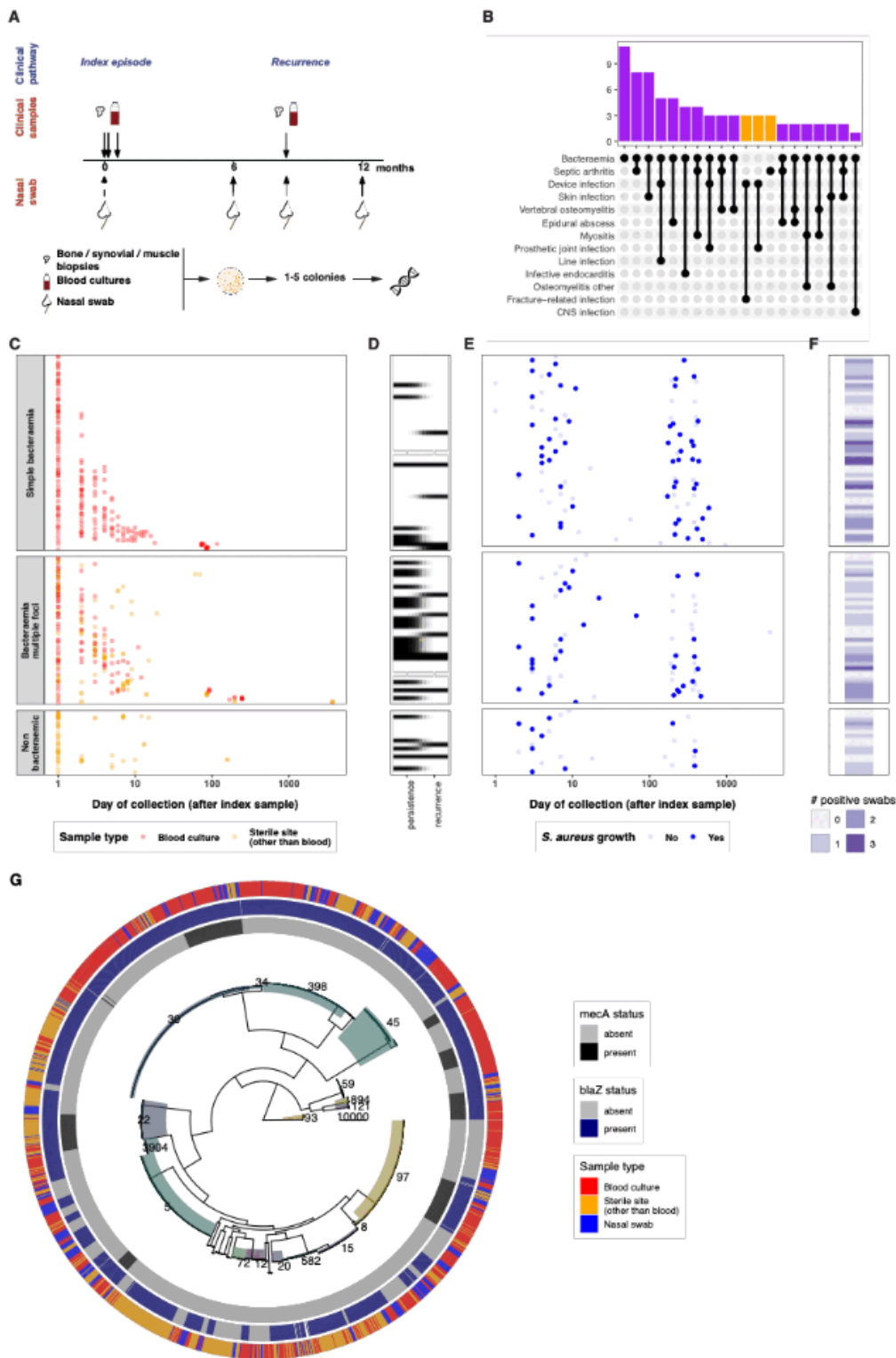


Figure 1

Overview of the study. A: All clinical samples with growth of *Staphylococcus aureus* were collected from hospitalised patients with invasive (sterile site) infections, including at any recurrence over the following years. Nasal swabs were collected at enrolment and at 6 and 12 months follow-up. B: Infection syndromes of 103 included episodes of invasive *S. aureus* infection: the Upset plot depicts the 20 most common combinations of infection foci. Bacteraemic infections are shown in purple, non-bacteraemic

infections in orange. C-F: Sample collection timeline, stratified by clinical syndrome. D: Heatmap with classification of each participant based on whether their *S. aureus* infection was persistent, recurrent or both (see methods for classification approach based on recorded microbiological samples). E: Nasal swabs collection timeline. F: Heatmap showing the number of swabs with growth of *S. aureus*. G: Global phylogeny of 1914 *S. aureus* genomes (57 *Staphylococcus argenteus* genomes are not shown, see Supplementary Figure S3). The 18 most prevalent clonal complexes are highlighted and labelled with their identifying number.

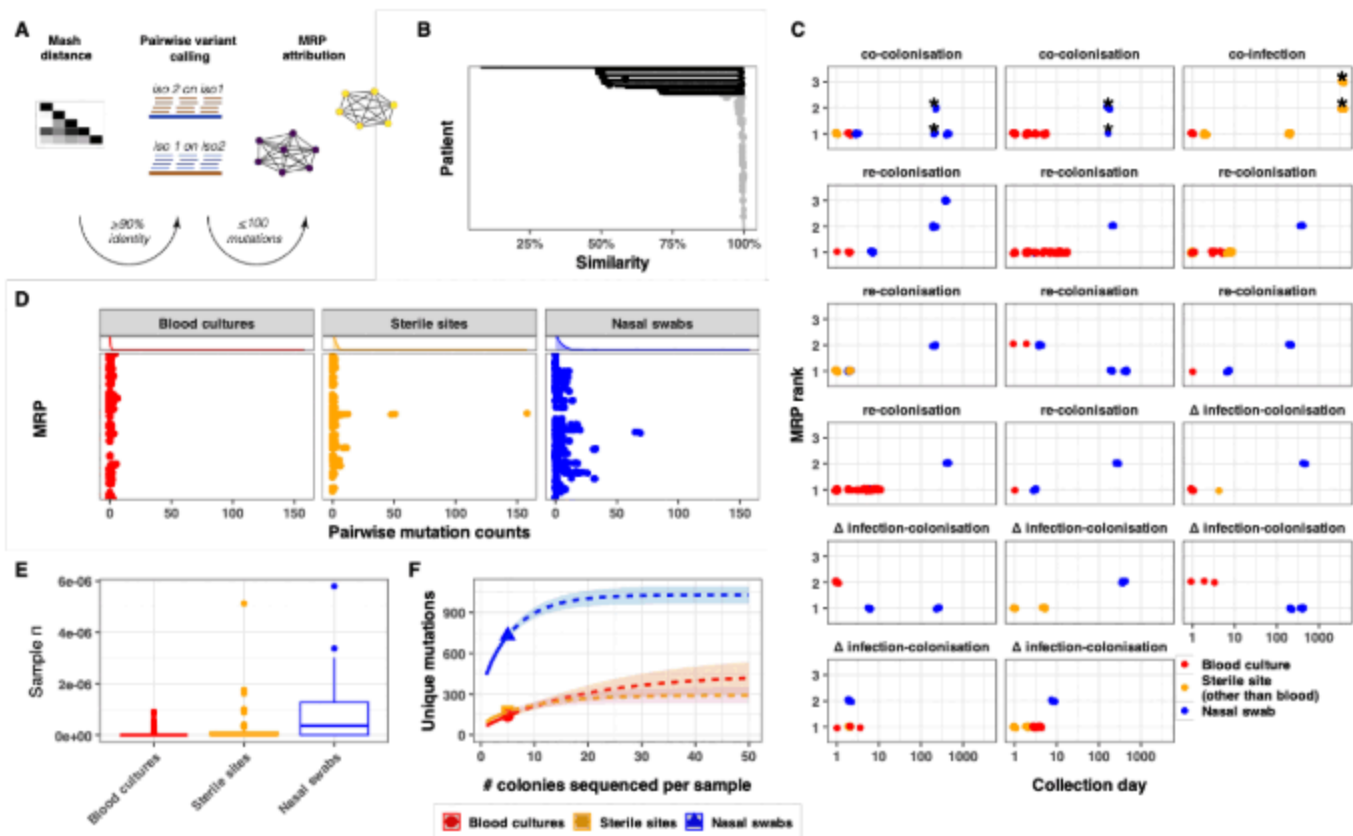


Figure 2

Within-host genetic diversity. A: Approach to calculate genetic distance among within-host genomes and group into micro-evolved resident populations (MRPs). B: Pairwise genetic similarity (expressed as fraction of shared hashes between genomes, as calculated with the Mash algorithm). Each boxplot shows median, interquartile range (IQR), the whiskers indicate 1.5xIQR. Eight episodes with at least one pair with $< 90\%$ similarity are depicted in black, episodes where all pairs have $> 90\%$ similarity are shown in grey. C: Similarity timelines of 17 episodes with multiple MRPs (on the y axis), interpreted as co-colonisation (> 1 MRP within the same nasal swab, $n=2$), re-colonisation or clonal replacement ($n=8$), co-infection ($n=1$) or dissociation between invasive samples and nasal swabs ($n=6$). Collection day is represented in logarithmic scale. Each dots depicts a genome, coloured according to the sample type. Samples with co-occurrence of genetically dissimilar genomes (i.e. mixed samples) are indicated by an asterisk. D: Within-sample pairwise mutation counts, stratified by type of sample. Each row represents an infection episode, each dot represents a pairwise comparison. The top of the plots depicts the

probability density function of the mutation counts for each sample. E: Within-sample diversity p according to the sample group. F: Rarefaction curves of genetic diversity (unique mutations) according to the number of colonies sequenced per sample, stratified by sample type. The continuous lines depicts rarefaction curves and the dash lines the extrapolation to up to 50 colonies sequenced.

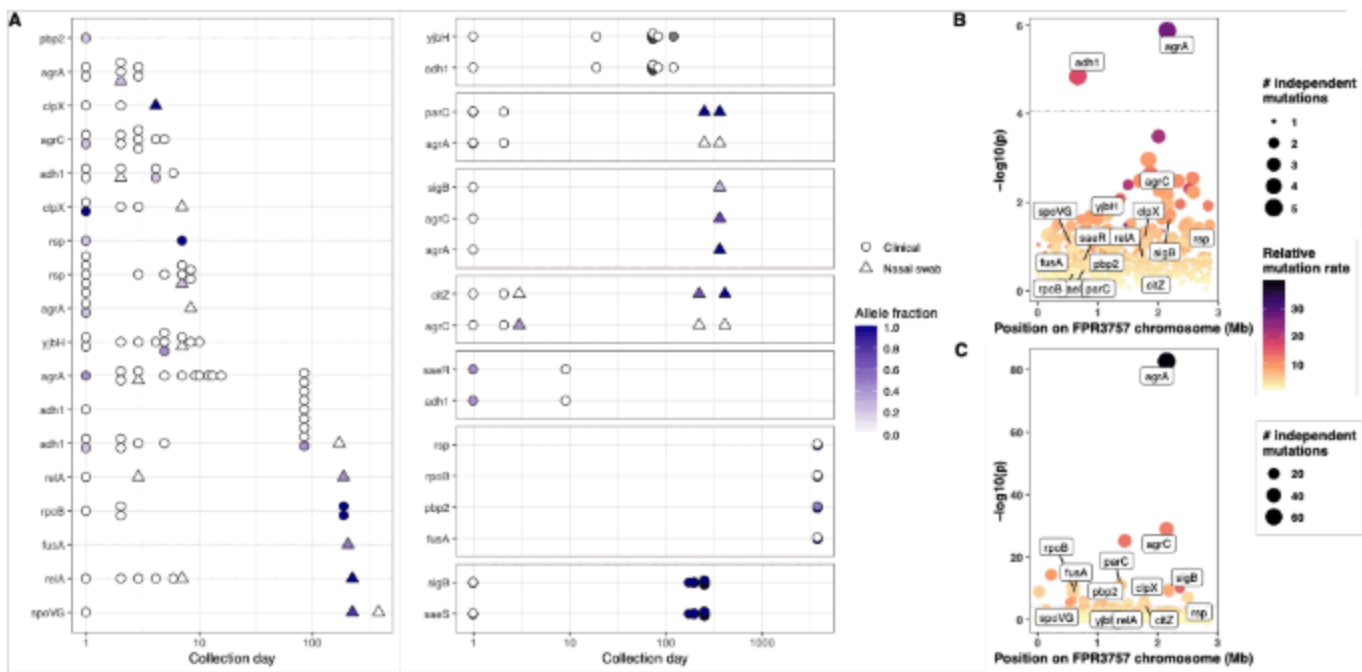


Figure 3

A: Timeline plots with the day of collection and type of samples carrying adaptive mutations. The dots are coloured according to the allele frequency of the adaptive mutations. Multi-allelic mutations are labelled, with colour intensity representing the allele frequency of the alleles (panel A: episodes with one mutated gene; panel B: episodes with more than one mutated gene). B,C: Manhattan plot of the between-host convergent evolution analysis of the BARISTA dataset (B) and the large-scale within-host evolution dataset (C). Genes with significant adaptive signal that are mutated in BARISTA are labelled and shown in detail in panel A.

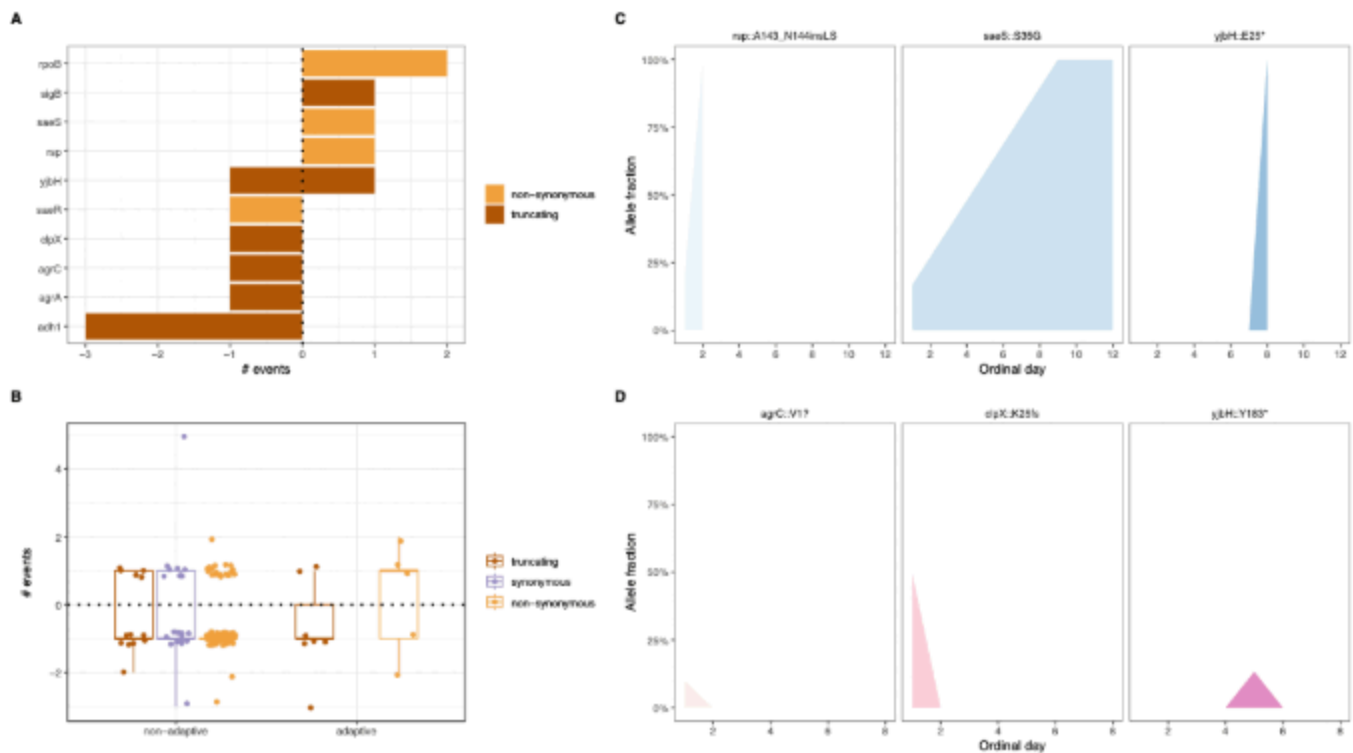


Figure 4

Trajectories of adapted subpopulations and selective sweeps. A: Overview of sweeps in adaptive genes (stratified by mutation type). B: Number of sweeps (positive scale) and non-sweeping events (negative scale) in adaptive genes vs non-adaptive genes, stratified by mutation type. C, D: Examples of frequency trajectories of selected adaptive genes (C: sweeping events; D: non-sweeping events) AF: allele frequency per collection timepoint.

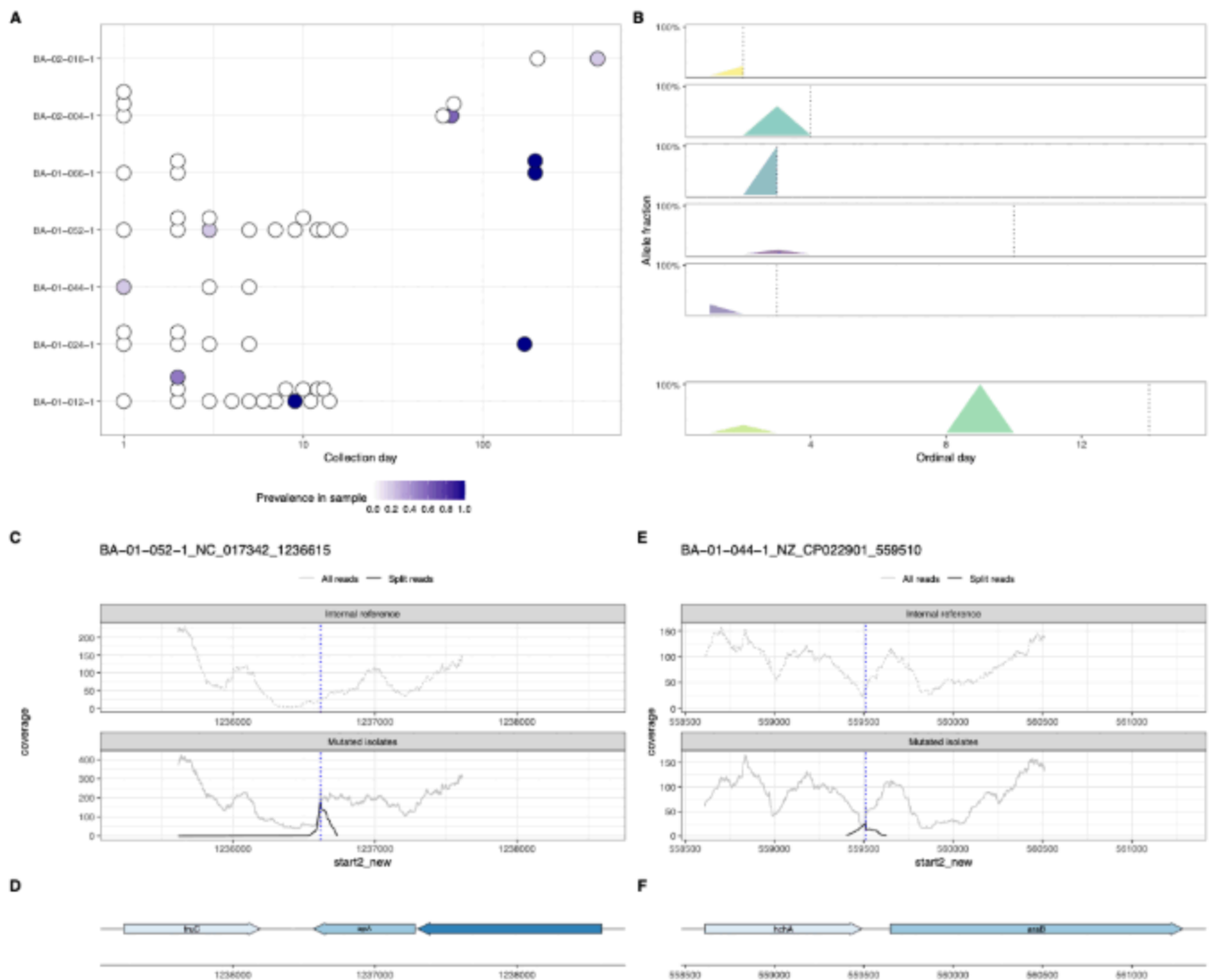


Figure 5

Structural variants. A: *IS* events. B: Evolution of subpopulations carrying the *IS* insertions shown on panel A: for each timepoint, the prevalence of *IS* insertions is represented as allele fraction (AF), each colour represents a unique *IS* insertions. The dotted line shows the last sampling timepoint for each episode (no map is shown for episode BA-01-024, as the sample carrying the *IS* included just one genome). C-D: Coverage reads for *IS* insertions for two representative examples (intragenic *agrA* insertion, panel C; *IS* insertion upstream of *araB*, panel D). E-F gene maps indicating the localisation of the *IS* insertions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementaryappendixv1.docx](#)