Niche-specific genome degradation and convergent evolution shaping *Staphylococcus aureus* adaptation during severe infections

Stefano G. Giulieri^{a,b,c}, Romain Guérillot^a, Sebastian Duchene^a, Abderrahman Hachani^a, Diane Daniel^{a,d}, Torsten Seemann^d, Joshua S. Davis^{e,f}, Steve Y.C. Tong^{f,g}, Bernadette Young^h, Daniel J. Wilsonⁱ, Timothy P. Stinear^{a*}, Benjamin P. Howden^{a,b,d*}

^a *Department of Microbiology and Immunology, The University of Melbourne at the Doherty Institute for Infection and Immunity, Melbourne, Australia*

^b *Department of Infectious Diseases, Austin Health, Heidelberg, Australia*

^c *Victorian Infectious Diseases Service, Royal Melbourne Hospital, Melbourne, Australia*

^d *Microbiological Diagnostic Unit Public Health Laboratory, The University of Melbourne at the Doherty Institute for Infection and Immunity, Melbourne, Australia*

^e *Department of Infectious Diseases, John Hunter Hospital, Newcastle, New South Wales, Australia*

^f *Menzies School of Health Research, Charles Darwin University, Casuarina, Northern Territory, Australia*

^g *Victorian Infectious Disease Service, Royal Melbourne Hospital, and University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia*

^h *Nuffield Department of medicine, Oxford*

ⁱ *Big Data Institute, Nuffield Department of Population Health, Li Ka Shing Centre for Health Information and Discovery, Old Road Campus, University of Oxford, Oxford*

OX3 7LF, UK

*Joint Senior author

#Address correspondence to Benjamin P. Howden, bhowden@unimelb.edu.au

Keywords: within-host evolution, adaptation, *Staphylococcus aureus*, genomics

ABSTRACT

During severe infections, *Staphylococcus aureus* moves from its colonising sites to blood and tissues, and is exposed to new selective pressures, thus potentially driving adaptive evolution. Previous studies have shown the key role of the *agr* locus in *S. aureus* pathoadaptation, however a more comprehensive characterisation of genetic signatures of bacterial adaptation may enable prediction of clinical outcomes and reveal new targets for treatment and prevention of these infections. Here, we measured adaptation using within-host evolution analysis of 2,590 *S. aureus* genomes from 396 independent episodes of infection. By capturing a comprehensive repertoire of single-nucleotide and structural genome variations, we found evidence of a distinctive evolutionary pattern within the infecting populations compared to colonising bacteria. These invasive strains had up to 20-fold enrichments for genome degradation signatures and displayed significantly convergent mutations in a distinctive set of genes, linked to antibiotic response and pathogenesis. In addition to *agr*-mediated adaptation we identified non-canonical, genome-wide significant loci including *sucA-sucB* and *stp1*. The prevalence of adaptive changes increased with infection extent, emphasising the clinical significance of these signatures. These

findings provide a high-resolution picture of the molecular changes when *S. aureus* transitions from colonisation to severe infection and may inform correlation of infection outcomes with adaptation signatures.

INTRODUCTION

While *Staphylococcus aureus* is one of the most important human pathogens (1), its common interaction with the human host is colonisation, usually of the anterior nares (2). Comparatively, severe, life-threatening infections such as bacteraemia or osteomyelitis occur very rarely. This suggests that at the macro-evolutionary level *S. aureus* is primarily adapted to its natural ecological niche (the nasal cavity) and to specific selective pressures arising in this environment, such as competition with the resident microbiota (3). By contrast, during invasive infection, a new fitness trade-off needs to be achieved to adjust to environmental challenges that include innate and acquired immune responses (4), high-dose antibiotics (5) and nutrient starvation (6). These trade-offs could occur across three potentially distinctive dynamics of micro-evolution during colonisation and infection (within the colonising population, from colonising to invasive and within the invasive population), leading to nose-adapted, early infection-adapted and late infection-adapted strains. Identifying infection-adapted strains might assist precision medicine strategies for infection prevention and management, and refine the understanding of *S. aureus* pathogenesis versatility, as mutational footprints of selection mirror functions that are critically important for bacterial survival during invasion.

Emerging genomic approaches for analysis of within-host evolution are among the most powerful means to study bacterial host adaptation (7-9). Studies have shown

the remarkable diversity and evolution of colonising populations of *Streptococcus pneumoniae* (10) and *S. aureus* (11). In *S. aureus* and *Enterococcus faecium* it has been shown that transition from colonisation to invasion favours strains with specific adaptive signatures (12, 13), while evidence of niche adaptation was noted in a within-host study of bacterial meningitis due to *S. pneumoniae* (14). Further, phenotypic and genomic adaptation (often in response to antibiotic pressure) has been investigated during selected episodes of persistent invasive infections due to *S. aureus* (15-17), *Pseudomonas aeruginosa* (18), *Salmonella enterica* (19), and *Mycobacterium tuberculosis* (20). To increase power, bacterial within-host evolution studies have leveraged on large collections of paired samples coupled with statistical models of genome-wide mutation rates (7, 14, 21) and extended the analysis to include chromosomal structural variation (17, 22, 23) as well as intergenic mutations (22, 24).

Convergent evolution among separated (independent) episode of colonisation or infection is a key indication of adaptation in evolution analyses. However, with the notable exception of one study of cystic fibrosis (25), the convergence has generally been weak in within-host studies of *S. aureus* infections, with no convergence at all (17) or significant enrichment limited to the *S. aureus* master regulator *agr* (12). We hypothesised that in addition to the small sample size, the extended range of bacterial functions potentially under selective pressure (each function being potentially targeted by diverse patho-adaptive mutations) has hampered the identification of important adaptation mechanisms. To overcome the limitations of studies to date, we have pooled all publicly available *S. aureus* within-host evolution studies, and complemented this with a new dataset from a recent *S. aureus* clinical

trial (26), in a single large-scale analysis. Rather than focussing on point mutations and small insertions/deletions alone, we leveraged multiple layers of genome annotation (encompassing coding regions, operons, intergenic regions and functional categories) and included chromosome structural variants to compile a comprehensive catalogue of bacterial genetic variation arising during host infection. This strategy enabled the detection of convergent adaptation patterns at an unprecedented resolution. We also outline distinctive signatures of adaptation during colonisation, upon transition from colonisation to infection and during invasive infection.

RESULTS

The *S. aureus* within-host evolution analysis framework.

We compiled a collection of 2,251 *S. aureus* genomes from 267 independent episodes of colonisation and/or infection, reported in 24 genomic studies (12, 17, 23, 27-47) (Table S1). We supplemented this dataset of publicly available sequences with unpublished sequences from 603 serial invasive isolates collected within the CAMERA-2 trial (26).

After excluding sequences failing quality control and genetically unrelated strains within the same episode, 2,590 genomes (1,397 invasive, 1,193 colonising) from 396 episodes were included in our within-host evolution analysis (Figure 1, Table 1, Supplementary files 1 and 2). The most prevalent lineages in the collection were sequence type (ST) 30 (342 strains, 13%), ST 22 (277 strains, 11%) and ST 5 (271 strains, 11%); 1001 strains (39%) were *mecA* positive. The collection was representative of the global *S. aureus* diversity, with an even distribution of colonising and invasive strains across the major clades (Figure 2A). The most

frequent infection syndrome was bacteraemia without focus (152 episodes, 38.4%), while nasal carriage (166 episodes, 42%) was the most prevalent colonisation condition (Table 1).

Our within-host evolution analysis strategy identified 4,556 genetic variants (median 3 per episode, range 0-237) (Supplementary file 3). Importantly, by investigating both point mutations and structural variation, we were able to uncover 214 large deletions (≥ 500 bp), 160 new insertion sequences (IS) insertions and 609 copy number variants, underscoring the role of large chromosome structural variation in within-host evolution. To increase the evolutionary convergence signal by aggregating mutations in functionally consistent categories, we annotated all genetic variants using multiple datasets, including coding sequences, regulatory intergenic regions, operons and gene ontologies (Figure 1C).

Distinctive evolutionary patterns define nose-adapted, early infection-adapted and late infection-adapted strains

Based on the working hypothesis that *S. aureus* host adaptation patterns differ according to whether the strains are nose-colonising, collected at an early stage of infection or at a late infection stage (i.e associated with persistence or treatment failure), we assessed whether it was possible to define i) general paradigms of genetic variation and ii) specific convergence signatures. Thus, we classified within-host acquired variants into three groups according to their most likely location in the within-host phylogeny: (i) between colonising strains (type C>C); (ii) between colonising and early infection adapted strains (type C>I); and (iii) between invasive

strains (type I>I). Overall, the 396 infection episodes included in the analysis allowed us to independently assess 166 type C>C, 118 type C>I and 312 type I>I within-host variants. In 95 cases there were sufficient samples to assess all 3 types within the same episode (Figure 2A).

We have previously shown that invasive strains from persistent or relapsing infections exhibit a high proportion of protein-truncating mutations (17). A similar enrichment of protein-truncating variants was identified within invasive strains as compared to strains from asymptomatically colonised individuals (12, 27). We reasoned that if this indicates genome degradation during infection, infecting strains might also be enriched for other loss-of-function mutations caused by structural variants, such as movement of insertion sequences (48) and large deletions, leading to complete or partial gene loss (49). In addition, we hypothesised that mutations and IS insertions in intergenic regions might contribute to altering gene expression or activity by interfering with the expression of key genes or operons (50).

Therefore, we calculated the prevalence of intergenic mutations, protein-truncating mutations, IS insertions and large deletions among all variants and compared it between type C>C (colonising-colonising), type C>I (colonising-invasive) and type I>I (invasive-invasive) variants. Strikingly, the distribution of mutations according to the predicted effect differed substantially in invasive-invasive pairs when compared to mutations identified between nose-colonising and invasive strains and within colonising strains (Figure 2 and Table 2). As compared to type C>C variants, variants emerging within the infecting strains were enriched for intergenic mutations (neutrality index [NI] 2.5; $p = 1.8 \times 10^{-16}$) and protein-truncating mutations (NI 2.4; $p =$

4.8 x 10⁻¹⁰). In contrast, no significant enrichment was observed among type C>I variants.

While large deletions were significantly more enriched in type I>I variants (NI 4.0, p = 1.1 x 10⁻¹⁵), the strongest evidence for enrichment (NI 19.9, p = 1.6 x 10⁻⁴²) was found for IS insertions. We and others have previously shown that new insertions of IS256 may provide an efficient mechanism of genomic plasticity in invasive *S. aureus* strains (17, 39, 50). Here, we expand this observation in a larger dataset and show that this mechanism is not limited to IS256 (Figure 2 – figure supplement 2). As shown in Figure 2, two invasive strains exhibited a burst of > 10 new IS insertions (IS3 and IS256, respectively). It has been shown that IS activation occurs under stress conditions, such as antibiotic exposure and oxidative stress (51), which is consistent with the selection environment encountered by invasive strains. However, these bursts occurred only in 2/1,068 adapted invasive strains.

Overall, these data support a model, where late infection-adapted strains show an enrichment for variants that are predicted to exert a stronger functional impact, either by producing a truncated protein or by potentially interfering with intergenic regulatory regions, through point mutations or IS insertions. This strong genome degradation signature appears to be specific to type I>I variants and was absent in type C>I variants, suggesting that the bottleneck effect upon blood or tissue invasion doesn't explain it. To assess whether this general enrichment of non-silent evolution represented a signature of positive selection or derived from within-host gene obsolescence occurring during invasive infection, we further investigated signals of

gene, operon and pathway specific enrichment across independent episodes of infection.

Gene enrichment analysis identifies significant hotspots of adaptation.

To identify signatures of adaptation, we first counted how many times each coding sequence was mutated independently across distinct colonisation/infection episodes (Figure 3, Tables 3 and Table 3 – table supplement 1, Supplementary file 3). We considered all protein-modifying mutations either predicted to cause a gain or loss of function to the locus: non-synonymous substitutions, truncations, IS insertions or deletions. To ensure consistency across the dataset, we restricted our analysis to genes with a homologue in reference strain FPR3757 (excluding plasmid genes and phage genes). Mutations were considered independent if they arose in distinct colonisation/infection episodes. To assess whether the convergent signals were a reliable indication of adaptation, we applied a gene enrichment analysis for protein-altering mutations which computes a length-corrected gene-level enrichment of protein-modifying mutations and estimates the significance of the enrichment for each gene using Poisson regression (12).

When applying a Bonferroni-corrected significance threshold (4.6×10^{-5}), mutations in *agrA* were highly significantly enriched across the entire dataset (45 fold enrichment, $p=7.0 \times 10^{-28}$). Other significantly enriched genes were *agrC* (13 fold enrichment, $p=2.8 \times 10^{-10}$), *stp1* (14 fold enrichment, $p=1.1 \times 10^{-7}$) and *mprF* (6 fold enrichment, $p=4.6 \times 10^{-6}$). The gene *sucA* reached near-significance (5 fold enrichment, $p=6.8 \times 10^{-5}$). Mutations in genes most significantly targeted by convergent evolution were evenly distributed across the *S. aureus* phylogeny, indicating that these adaptive

mechanisms were not specific to selected lineages (Figure 3 – figure supplement 1).

Using dN/dS analysis, we confirmed signatures of positive selection in the most significantly enriched genes, although only *agrA* reached statistical significance (Figure 3 – figure supplement 2 and supplementary file 4).

We found that several genes with the most significant enrichment (*agrA*, *agrC*, *stp1*, *sucA*) were recurrently mutated across all three within-host evolutionary scenarios, implying a global role in *S. aureus* adaptation during colonisation and invasion. This suggests partial adaptation of *S. aureus* strains upon invasion. It has been previously shown that adaptive mutations, particularly within the quorum sensing accessory gene regulator (*agr*), are enriched in invasive *S. aureus* strains, suggesting that adapted strains are more prone to be involved in invasive disease (12, 27, 47, 52). While the key role of *agr* was consistent with previous evidence from clinical and experimental studies, the high number of recurrent *sucA* mutations was surprising. This metabolic gene encodes for the α -ketoglutarate dehydrogenase of the tricarboxylic acid cycle, and recent work has revealed the functional basis of its potential role in adaptation. Its inactivation was found to lead to a persister phenotype (53) and *sucA* was a hotspot of metabolic adaptation to antibiotics in a recent *in vitro* evolution study (54).

To confirm that our gene enrichment analysis (focused on point mutations and IS insertions and limited to genes with FPR3757 homologs) captured the large part of adaptation, we analysed variation due to large deletions and copy number variation, which were not included in the gene enrichment analysis. We observed multiple independent deletions and amplifications mainly in phage genes (Figure 3 – figure

supplement 3 and Figure 3 – figure supplement 4). We also repeated the gene enrichment analysis with all mutated genes (with and without FPR3757 homologs) and found very similar results, with only two hypothetical proteins with no FPR3757 homolog among the genes with most significant enrichment (Figure 3 – figure supplement 5).

Combining multiple mechanisms of adaptation and multilayered annotation increases the signal of convergent evolution

To increase our ability to capture signatures of adaptation from convergent evolution, we extended our analysis beyond coding sequences, to integrate the genetic variation signals issued from intergenic mutations and IS insertions in intergenic regions. This multi-layered annotation of mutated regions was shown to increase the amount of information gained from *in vitro* adaptive evolution experiments (55). Such methodology allows for an advanced classification of intergenic mutations based on regulatory sequences including promoters and transcription units based on data acquired from RNAseq experiments (56, 57).

Using this approach, we were able to assign 150/1237 (11%) of all intergenic mutations and IS insertions to a predicted regulatory region. We found that the *agr*, *sucAB*, *walkR* operons had the strongest convergent evolution signal with 28, 13 and 13 independent mutations (Supplementary file 5). Mutations within these loci were significantly (*agr*, 12-fold enrichment, $p=1.1e-21$; *sucAB*, 4-fold enrichment, $p=4.5e-5$) or near-significantly enriched (*walkR*, 3-fold enrichment, $p=1.7e-4$) (Figure 4). Interestingly, promoter mutations represented 2/13 (15%) of the *walkR* operon mutations, indicating that potentially impactful intergenic variants may be missed

when considering only coding regions. Further, new IS insertions were found within type I>I (invasive-invasive) variants: three insertions into *agrC* (predicted to inactivate the gene, as shown previously in staphylococci (58, 59)) and an insertion 159 bp upstream of *walR*, in a region encompassing its cognate promoter. Together with the strong enrichment for IS insertions within type I>I variants, the location of these insertions in recurrently mutated operons suggests that IS insertions contribute to the adaptive evolution of *S. aureus* during invasive infection.

Adaptation within the invasive population is distinctive and strongly driven by antibiotics.

The excess of non-silent evolution (and potentially function-altering) within invasive strains suggested that strong, specific selection pressure occurs within the invasive populations (type I>I variants). We therefore assessed genes that appeared to be specifically mutated or inactivated during infection. We performed our gene- and operon-enrichment analysis for each type of within-host variants separately (i.e. within the colonising population, between colonising and invasive strains and within the invasive population) (Figure 5). We found that *agrA* mutations were highly enriched in any group of variants, and particularly prevalent between colonising and invasive strains (type C>I variants), consistent with a previous study that is included in this analysis (12). Among type I>I variants (between invasive strains), a significant enrichment was observed in *mprF* (18 fold enrichment, $p = 2.8 \times 10^{-9}$), *agrC* (24 fold enrichment, $p = 2.1 \times 10^{-7}$), *rpoB* (10 fold enrichment, $p = 8.8 \times 10^{-6}$). Other genes that were strongly enriched in type I>I variants (below the Bonferroni-corrected threshold, but above the suggestive significance threshold, Figure 5), included *walR* (22 fold enrichment, $p = 3.5 \times 10^{-4}$), *stp1* (20 fold enrichment, $p = 4.2 \times 10^{-4}$), *yjbH* (19 fold

enrichment, $p = 5.4 \times 10^{-4}$), *sgtB* (19 fold enrichment, $p = 5.5 \times 10^{-4}$) and *purR* (18 fold enrichment, $p = 5.8 \times 10^{-4}$).

The enrichment for mutations in *mprF*, *rpoB*, *stp1*, *sgtB* and in the *walKR/yycH* operon (11 fold enrichment, $p = 9 \times 10^{-9}$, see Figure 5 – figure supplement 1 for the operon enrichment analysis) highlights the role of antibiotic pressure in shaping adaptation within the invasive population, since these loci are hotspots of adaptation to key anti-staphylococcal antibiotics that are often used in invasive infections (rifampicin, daptomycin, vancomycin). For example, the essential two-component regulator *walKR/yycFG* (and its associated genes *walH/yycH*) have been shown to have a key role in vancomycin resistance in one of the within-host evolution studies included in this analysis (41), while mutations in both *stp1* and *sgtB* have been observed in vancomycin-adapted strains (60).

Notably, the most significant gene signatures in invasive strains might have been selected in response to other selective pressures, including the host immune response during infection. For example, *rpoB* mutations have been associated with pleiotropic effects, including co-resistance to vancomycin, daptomycin and oxacillin and immune evasion, suggesting a potential role in adaptation beyond the response to the selective pressure from rifampicin (61). This hypothesis is supported by the presence of mutations (such as the *rpoB* R503H substitution and N405 inframe deletion) outside the rifampicin-resistance determining region

Pleiotropic phenotypes are also likely to underlie the enrichment of *yjbH* with invasive strains, which was mutated four times (of which three were truncations), yet

only one mutations was found in colonising strains or early infection-adapted strains. This gene has a cysteine-rich domain that is homologous to *dsbA* in *E. coli*. One of its roles in *S. aureus* is to facilitate the ClpXP-dependent degradation of the transcriptional regulator Spx (62). Inactivation of *yjbH* has been associated with oxacillin (63) and vancomycin (64) resistance, impaired growth (65) and reduced virulence in animal models (66), indicating that *yjbH* mutations may influence both host-pathogen interaction and antibiotic resistance. Finally *purR*, a purine biosynthesis repressor, has been recently characterised beyond its metabolic function: interestingly it was shown to be a virulence regulator (67), where *purR* mutants displayed higher bacterial counts following mice infection, increased biofilm formation (68) and higher capacity to invade epithelial cells (69).

We performed a gene set enrichment analysis (GSEA), using gene ontology and antibiotic resistance gene annotations (70). The GSEA, stratified by variant type, showed significant enrichment only in type I>I (invasive-invasive) variants, further underscoring the higher level of adaptation in this group (Figure 5 – figure supplement 2 and supplementary file 6) and confirmed the broad functional implications of the most enriched genes and operons with the invasive populations, since among the ontologies that were significantly enriched within the invasive population, we found the categories “DNA binding” (normalised enrichment score [NES] = 1.6, false-discovery rate (FDR)-adjusted $p = 9 \times 10^{-4}$), “pathogenesis” (NES = 1.7, adjusted $p = 4 \times 10^{-3}$) and “antibiotic response” (NES = 1.8, adjusted $p = 7 \times 10^{-3}$).

Taken together, these findings point to six key genetic loci that appear to have an important role in *S. aureus* adaptation during during invasive infections. These loci

are associated with either antibiotic resistance (*mprF*, *rpoB*, *stp1*, *sgtB*, *walkR*),
pathogenesis (*agrAC*, *purR*) or both (*yjbH*).

A mutations co-occurrence network defines loci under within-host co-evolutionary pressure

Epistasis, defined as the interaction of multiple mutations on a given phenotype (71), plays a role in adaptive evolution in bacteria, particularly in antibiotic resistance (72-74). Whether epistatic interactions could promote *S. aureus* adaptation during infection remains unknown. Identifying these interactions would enable identification of combinations of mutations underlying bacterial adaption during infection and refine the prediction of infection outcomes. Here, we assessed co-occurrence of mutations and mutated genes across independent episodes of colonisation/infection. While co-occurrence may simply result from co-selection (e.g. simultaneous exposure to two different antibiotics), it may also indicate putative epistatic interactions, that could be explored in terms of potential impact on adaptive phenotypes (75).

First, we explored co-occurrence of mutations, and found only one case where the same mutations co-occurred in more than one independent episode. The two mutations were an inframe deletion within hypothetical protein SAUSA300_2068 and a A60D substitution of the gene *ywIC*. These genes are closely located in FPR3757. While the co-occurrence could be explained by recombination, recombination is expected to be rare amongst within-host *S. aureus* populations in general (43) and even rarer within invasive strains. *YwIC* is a threonylcarbamoyl-AMP synthase in *E. coli*, thus it is possible that SAUSA300_2068 is also a ribosomal protein. Ribosomal

proteins can display regulatory activity (76), and could plausibly be targets of adaptation to both antibiotics and the host/intracellular survival. This specific case of convergent co-occurrence of mutations was detected within type I>I variants.

When assessing interactions at gene level (i.e. co-occurrence of the same altered protein sequences across independent episodes), we found the strongest interaction between the *agrA* and *agrC* genes (Figure 6). While this is consistent with the high convergence of mutations in the *agr* locus, this suggests that strains acquire multiple mutations within the locus, possibly further impacting *agr* activity. Interestingly, no convergent co-occurrence signature compatible with possible epistasis was observed within the *walkR* locus, the other operon with a high number of independent mutations; which could be due to the essentiality of *walkR* in *S. aureus* (77). Collectively, *agr* locus mutations interacted with 17 other mutated genes, the strongest interaction being with *stp1*. Since *stp1* (a serine/threonine phosphatase) has been previously associated with virulence regulation (78), this interaction potentially indicates another mechanism by which adapted strains fine-tune the gene expression profile that is already altered by *agr* mutations.

Another moderately strong interaction was observed between *rpoB* and *parC*, which were co-mutated in three independent episodes. Given the association of *parC* mutations with fluoroquinolone resistance (79), this interaction is likely to be an example of co-selection due to co-exposure to fluoroquinolones and rifampicin.

Clinical correlates of adaptive signatures within colonising and invasive populations

Genetic signatures of bacterial adaptation have been associated with infection extent, for example enabling the prediction of extraintestinal infection with *Salmonella enterica* (80). We have previously shown that adaptive mutations are enriched in invasive infections, however, it is unclear whether bacterial adaptation is more likely to be associated with distinctive clinical syndromes. To identify clinical correlates of adaptive signatures we classified colonisation and infection episodes based on the sites of collection and on clinical data obtained from the publications (table 1 and Figure 7 – figure supplement 1). We then used the Jaccard index and network analysis to compute node centrality as a global measure of adaptation for each independent episode. The Jaccard index can be used as a simple marker of the proportion of shared mutated genes between pairs of colonisation or infection episodes (81). Node centrality allows to simultaneously take into account the strength of similarity between independent episodes (Jaccard index) and the number of pairs with shared mutated genes (number of connections). Here, we limited the analysis to the 20 most significantly enriched genes with each type of variant.

Our network analysis showed that adaptation was present in only a minority of episodes within each type of variant (Figure 7 – figure supplement 2). With a definition of adaptation based on a centrality value of more than 0, we found that the proportion of adaptive episodes was 43%, 20% , 22% with type C>C, C>I and I>I variants, respectively. In addition certain clinical syndromes were more strongly associated with adaptation. Within the colonising population (type C>C variants), almost 80% of cystic fibrosis episodes were adaptive, as opposed to one third of

episodes of skin colonisation in atopic dermatitis (Figure 7AB). This is consistent with within-host evolution studies showing strong convergent evolution signals among bacterial populations colonising individuals with cystic fibrosis, not only in case of *S. aureus* colonisation (25), but also *Pseudomonas aeruginosa* (7) and *Mycobacterium abscessus* (82), however, one study found adaptive evolution signals in atopic dermatitis (83). We also observed that adaptation among infection episodes correlated with infection extent. Episodes of infective endocarditis episodes displayed higher adaptation metrics (46% with centrality > 0) than bacteraemia with additional infection foci (28%) and bacteraemia without focus (17%) (Figure 7 D-E).

To explore the syndrome-specificity of adaptation signatures, we mapped mutations in the most significantly enriched adaptive genes to clinical syndromes of colonisation and infection (Figure 7 panels C and F). As expected, syndromes with high prevalence of adaptation had higher numbers of episodes with adaptive mutations; however, some genes appeared to be preferentially mutated. For example, *rpsJ*, *stp1* and SAUSA300_1230 were over-represented in cystic fibrosis, while no clear pattern of mutations was discernible for nasal carriage episodes. Within infection syndromes, *mprF* and *purR* mutations were more prevalent in endocarditis, and *yjbH* mutations were only found in severe infections (bacteraemias with additional foci and endocarditis). Some genes appeared to be distinctive for low adaptation groups (atopic dermatitis, skin infections), however the low number of adaptative mutations prevented an accurate assessment of these profiles.

DISCUSSION

Within-host evolution of bacterial pathogens such as *S. aureus* is thought to be governed by a combination of positive selection for variants that confers an advantage within the host, and fixation of mutations owing to genetic drift and relaxation of purifying selection (19, 84). Sudden changes in the effective population size (bottlenecks) can cause genetic drift, for example when a a single or few bacterial cells invade the bloodstream or when a secondary infection foci is established in tissues and organs. Consistent with this view, animal studies have shown that after infecting the blood with a polyclonal population, bacteraemic infection is established stochastically by a single clone (85, 86). On the other hand, several lines of evidence support the role of positive selection and adaptive evolution during *S. aureus* infection. First, adaptive phenotypic features appear to be acquired during infection. The most obvious adaptative phenotype is secondary resistance to anti-staphylococcal antibiotics such as rifampicin, vancomycin (16), daptomycin (87), oxacillin (15). Crucially, these resistance phenotypes can be associated with pleiotropic, patho-adaptative phenotypes such as small colony variant and immune evasion (61, 88, 89). Further, phenotypic adaptation (e.g. loss of toxicity) has been observed upon transition from colonisation to infection (90), supporting the concept that invasive infection is linked to patho-adapted strains. At the molecular level, an excess of protein-truncating mutations in invasive strains (17) and in late colonising strains leading to infection (27) have been noted. While this observation alone could be explained by relaxed constraint resulting from reduced population size (84), it has been suggested that loss of gene function might be a common adaptation mechanism of within-host evolution (21), as supported by evidence of gene- or

pathway-specific enrichment of mutations across independent infection episodes (12).

Despite support for adaptive evolution from previous studies, it has been difficult to identify specific molecular signatures of adaptation during infection, due to the limited power of previous within-host studies of bacteraemia and other serious *S. aureus* infections, that were often limited to a restricted number of episodes. To increase our ability to identify signatures of adaptation and find significantly enriched loci, we analysed multiple sources of genetic variation (point mutations, large deletions, IS insertions, copy number variants) in a large collection of independent episodes of *S. aureus* colonisation and infection from 25 studies. We predicted that the main advantage of our approach would be to increase the ability of detecting convergence of genetic variants arising during invasive infections as opposed to those detected during the colonisation and upon transition from colonisation to infection. To test this hypothesis, we classified within-host variants based on their likely position in the within-host phylogeny (Figure 1D).

Bacterial adaptation is promoted by genomic plasticity, however, within-host evolution is characterised by low genetic variation (17). Based on our previous genomic studies of *S. aureus* bacteraemia, we reasoned that chromosome structural variants may provide an additional mechanism to increase genetic variation during infection. Here, we found that new insertions of insertion sequences (IS) are strongly enriched during invasive infection. However, despite this 20-fold enrichment, IS insertions remained a rare source of variation even within invasive strains and appeared to have a selective contribution to adaptation (i.e. limited to specific loci

such as *agrC*). Similarly, large deletions and copy number variants appeared to play a less prominent role in adaptation, although we didn't include them in our enrichment analysis,

Together with the enrichment for loss-of function mutations, which is another feature of evolution within the invasive population in our analysis, these IS movements suggest that a pattern of reductive evolution emerges during within-host evolution of invasive *S. aureus*. This excess of genome degradation might be related to less effective purifying selection in the invasive population due to a decrease in effective population size and a shorter evolutionary timescale (84). However, our data indicate that these changes converge to specific genes, operons and pathways, suggesting an adaptative benefit. Reductive evolution has been described in several “commensal-to-pathogen” settings (49). Although niche adaptation through reductive evolution has been described in extracellular pathogens (91), a smaller (reduced) genome is an hallmark of obligate intracellular endosymbiotic bacteria (92). Since it appears that invasive *S. aureus* is able to reside intracellularly (promoting dissemination through mobile phagocytes during *S. aureus* sepsis (93, 94), and immune evasion), it is plausible that this pattern of reductive evolution reflects an adaptation of invasive *S. aureus* to an invasive and intracellular lifestyle, as it has been shown for other facultative intracellular pathogens such as non typhoidal *Salmonella* (19). However, it is possible that these signatures of reductive evolution might be only temporary, as genome degradation might be present only in a minority of strains or be reversible; moreover, loss-of-function mutations are expected to be more likely than gain-of-function mutations.

Beside reductive evolution, another distinctive feature of within-host evolution during invasive infection were intergenic mutations (both point mutations and IS insertions). In a within-host evolution study of *S. pneumoniae* colonisation, it was shown that intergenic sites were under convergent evolution (10). Mutations in promoter sequences of some core genes can play an important role in antibiotic resistance as it was repeatedly shown for *pbp4* and resistance to beta-lactam antibiotics (95). The role of intergenic mutations in within-host evolution was shown in a study of *P. aeruginosa* infection, where convergent evolution targeted several intergenic regulatory regions including upstream of antibiotic resistance genes (24).

Previous work on within-host evolution by our group (included in this analysis) has established that *agrA* mutations are significantly enriched upon transition from colonisation to infection (12). In addition, we have shown through genomics and targeted mutagenesis that mutations in key genes such as *walkR* (41) and *rpoB* (96) play a key pathoadaptive role in selected cases of persistent *S. aureus* infections. In this study, we increased our ability to discover potential targets of adaptation by analysing several mechanisms of genetic variation and applying several layers of annotation. As compared to previous work on *S. aureus*, this approach provides a higher-resolution picture of within-host evolution and adaptation. Importantly, this analysis remains robust after removing more than 1,000 sequences from the largest within-host study included (Figure 3 – figure supplement 6). We increase here the generalisability of our findings. We expand the list of genes targeted by convergent evolution and show that there are distinctive adaptation pathways in colonising and invasive populations. We confirm that the dominantly mutated loci belonged to the *agr* locus, in particular *agrA* and *agrC*. This finding is consistent with a large body of

literature that predated the genomic era (97), that supports the role of the *agr* locus as the master regulator of gene expression in *S. aureus*. *Agr*-mediated adaptation was so important that we found a highly significant enrichment of *agr* mutations across all type variants, including within colonising strains (type C>C variants). Shopsin *et al.* showed that ~10% of healthy *S. aureus* carriers held an *agr*-defective strain and that prior hospitalisation was significantly associated with *agr*-defective status, suggesting prior adaptive pressures (98).

Consistent with the distinctive general patterns of evolution displayed during invasive infection, some genes and loci were specifically mutated within invasive strains. Some of these genes were emerging targets of *S. aureus* pathogenesis *in vivo*, such as *purR* and *yjbH*, that were not singled out in previous within-host evolution investigations. Others were known determinants of antibiotic resistance, including *mprF*, *rpoB* and the *walKR* operon. This underscores the crucial role of antibiotic exposure in shaping adaptive evolution during invasive infection. A recent study of within-host evolution during cystic fibrosis found that resistance genes were hotspots of convergent evolution in this population, which is frequently treated with antibiotics and shows features of phenotypic adaptation (25).

While most mutations in adaptive loci were substitutions within the coding sequence, about 40% of *walKR* operon mutations were located outside of the coding regions of *walR* and *walk*, emphasising the need to study intergenic mutations and mutations throughout an operon to capture adaptive signatures. This observation highlights the importance of expanding the analysis of intergenic mutations for the detection of adaptive mutations, in particular those linked to antibiotic resistance.

568

569 If within-host evolution represents adaptative evolution it is possible that adaptation
 570 involved an accumulation of mutations and possibly epistatic mechanisms. Our data
 571 show that some mutations are specific for invasive strains; these mutations may
 572 reflect late adaptation, occurring after evolution during colonisation in the nose and
 573 upon transition from colonisation to infection, and thus occurring after adaptive
 574 mutations were acquired during earlier stages. This evolutionary pattern of stepwise
 575 adaptation (or adaptive continuum) encompassing the entire within-host evolutionary
 576 arch has been well described for cancer (99) and has been also investigated in a
 577 study of transition from colonisation to infection (27). One way to capture this is
 578 mutation co-occurrence analysis. Here, we show mutation co-occurrence within the
 579 *agr* operon, but also co-occurrence of the same mutations in two uncharacterised
 580 proteins in *S. aureus*. Our network of mutation co-occurrence linked to *agrAC*
 581 mutations might suggest a potential pathway of stepwise adaptation following initial
 582 mutations acquisition in the *agr* locus, an hypothesis that has been explored in one
 583 of studies included in our analysis (47).

584

585 While combining multiple studies allowed us to increase statistical power in order to
 586 detect genome-wide convergent signals of adaptation, this approach has some
 587 limitations. The quality of the publicly available sequences and metadata can be
 588 heterogenous, despite performing quality control assessment, for example due to
 589 different read coverage across studies. In addition, detection of structural variants
 590 from short reads is not as accurate as from long reads; for example, chromosomal
 591 inversions can be missed if they inversion site span is larger than the insert site of
 592 the paired-end reads (89).

Ultimately the goal of detecting adaptive signals is to identify new mechanisms of pathogenesis or resistance to therapeutic targets and to inform prediction of clinical outcomes. So far, studies have failed to show consistent associations between specific clinical outcomes and genetic features of infecting (or colonising) *S. aureus* strains. This might be related to the dominance of host / environmental factors, but it could also be linked to the limited power of studies performed so far. By contrast, within-host evolution studies are an elegant approach to identify signatures of adaptation that might be candidate markers of important clinical outcomes, such as infection risk in case of colonisation or treatment failure in case of infection. Here, we show that adaptation signatures are at least partially specific to colonisation, infection and upon transition from colonisation to infection, and that adaptive changes are more frequent in distinctive infection episodes (complicated bacteraemia, endocarditis). These findings suggest that adaptive signatures might be indicative of important clinical outcomes. In the future, precision medicine in infectious diseases could follow the lead of cancer genomics, where within-host evolution studies have tracked the evolution of cancer clones and enabled the detection of high risk mutations early.

MATERIAL AND METHODS

Literature search

We conducted a search of articles indexed in PubMed before the 11th August 2020 using the keyword “aureus” in combination with either “genomics” or “whole genome sequencing” and with either “within-host evolution” or “in vivo evolution” or “adaptation” or “bacteraemia”. The records retrieved through this search were

combined with additional citations identified through other sources. After removing duplicates, this resulted in 815 citations that were screened based on following inclusion criteria: i) Whole-genome sequencing of human *S. aureus* isolates; ii) > 1 *S. aureus* isolates sequenced per individual; iii) Sequences (reads or assemblies) publicly available; iv) Minimum sequences metadata available (either with the manuscript or linked to the sequences): patient ID, date of collection (or collection interval in reference to a baseline isolate), source of the sample. After excluding studies not satisfying the inclusion criteria (730 based on the title, 46 based on the abstract and 15 after reviewing the full text), we kept 24 within-host evolution studies.

Extraction of sequences metadata

For each of the included studies, following variables were extracted either from reads metadata (when available) or from the publication/supplementary data: identifier linking the sequences to a patient or an episode of infection, date of collection (when available) or collection interval in reference to a baseline isolate), site of collection of the isolate. Isolates were broadly categorised in “colonising” and “invasive” based on the site of collection, when the information was unambiguous (e.g. “nose” for “colonising” or “blood” for “invasive”). When the information on the body site was not sufficient (e.g “skin” or “lung”), the categorisation was based on further details provided in the publication. When available, phenotypic metadata and antibiotic treatments were also extracted from the publication. We used clinical and site data to classify colonisation episodes in “nasal carriage”, “atopic dermatitis” and “cystic fibrosis” and infection episodes in “skin infection” (skin infection site surgical site infection without other foci), “osteoarticular infection” (bone/join infection without other foci), “bacteraemia without focus” (bloodstream infection, no other foci, expect

for vascular catheter or skin), “bacteraemia with focus (bloodstream infection with other focus involving the lung, nervous system, bone and joints or internal organs), and “endocarditis” (based on diagnosis reported in the publication or in the clinical metadata).

Sequences processing

Sequences (reads and assemblies) and metadata were downloaded from the European Nucleotide Archive (ENA) and the National Center for Biotechnology Information (NCBI), respectively using the BioProject accession or the genome accession. Quality assessment of the reads was performed by calculating mean read depth and the fraction of *S. aureus* reads using Kraken 2, v2.0.9-beta (100) and by extracting metrics from reads assemblies constructed using Shovill, v1.1.0 (<https://github.com/tseemann/shovill>) and annotated using Prokka, v1.14.6 (101). Sequence type (ST) was inferred from the assembly using Mlst, v2.19.0 (<https://github.com/tseemann/mlst>) and resistance genes were detected using Abricate, v1.0.1 (<https://github.com/tseemann/abricate>). Reads were discarded if the mean coverage depth was below 35, the majority of reads were not *S. aureus* or the size of the assembly was below 2.6 megabases. Assemblies downloaded from the NCBI repository were discarded if the genome size was below 2.6 megabases.

Sequences from the CAMERA2 trial

We collected *S. aureus* strains from bacteraemia episodes included in the CAMERA2 trial (Combination Antibiotics for Methicillin Resistant *Staphylococcus aureus*), where at least 2 strains per episode were available. The CAMERA2 trial was performed between 2015 and 2018 in Australia, New Zealand, Singapore and

Israel, and randomised participants with methicillin-resistant *S. aureus* (MRSA) bacteraemia to either monotherapy with vancomycin or daptomycin or combination therapy with vancomycin or daptomycin plus an antistaphylococcal beta-lactam (flucloxacillin, cloxacillin or cefazolin) (26). Strains were isolated from -80C glycerol onto horse-blood agar. Species was confirmed using Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (MS). Bacterial whole-genome sequencing was performed from single colonies on the Illumina NextSeq platform. Reads were checked for quality, assembled and annotated as described above.

Global phylogeny

To generate a global alignment of all sequences, reads and shredded assemblies were mapped to reference genome USA300 FPR3757 (assembly accession: GCF_000013465.1) (using Snippy, v4.6.0 (<https://github.com/tseemann/snippy>)). The core genome alignment was obtained using Snippy; sites with > 10% gaps were removed using Galign (102) and constant sites were removed using SNP-sites (103), for a final length of 186,825 bp. A maximum-likelihood phylogenetic tree of 2,590 sequences (those kept in the analysis after excluding genetically unrelated strains, see below) was inferred using IQ-TREE, v2.0.3 (104).

Variant calling

We have previously shown that the accuracy of variant calling in within-host evolution analyses is improved when mapping reads to an internal draft assembly as opposed to a closely related closed genome (17). Here, we applied the same approach, where we selected the internal reference among the sequences from the

same patient or episode. When available, the oldest colonising strain was selected. When only sequences from invasive strains were available, the oldest strain (baseline strain) was selected. When multiple sequences were available per sample (e.g multiple colonies sequenced per plate) or at the same date, the reference was randomly selected among them. Snippy with default parameters (minimum reads coverage 10, minimumread mapping quality 60, minimum base quality 13) was used for variant calling. To further improve the accuracy of the calls, we masked variants called from reference reads and those at positions where reference reads had a coverage below 10 (using the BEDTools suite (105)).

Filtering of genetically unrelated sequences

The threshold for removing genetically unrelated sequences was set empirically at 100 episode-specific variants based on the upper Tukey's fence of the distribution of the number of variants in same-episode isolates belonging to the same ST (Figure 2 – figure supplement 1).

Detection of chromosome structural variants

Using BWA-MEM, reads and shredded contigs were aligned to the closest available complete genome (either internal to the dataset or selected from the NCBI repository based on the mash distance). To detect large deletions (≥ 500 bp), reads coverage was compared between the internal reference and the sequences within every episode using BEDTools, as described in (17). To detect new insertion sequences (IS) insertins, split reads were extracted, filtered and annotated as described in (17). We used the R package CNOGpro (106) to detect 1000 bp windows with an estimated copy number above one as compared to the internal reference. The

package calculates the reads coverage per sliding windows of the chromosome, performs a GC bias normalisation and infers copy number state using a Hidden Markow Model. We ran the package with default parameters, with the exception of the length of the sliding window that was set at 1000 bp. As with deletions, we used BEDTools to subtract regions of the internal reference that had already an increased copy number.

Internal variant annotation

To ensure a consistent annotation of mutated genes across independent episodes, we clustered amino-acid sequences using CD-HIT, v4.8.1 with an identity threshold of 0.9. The BEDTools suite was used to annotate mutated intergenic regions with upstream and downstream coding regions and the distance separating the mutation from the start or the end of the gene. For the operon analysis, intergenic mutations were classified according to their location within a presumed promoter based on blasting the sequence of unique promoters (as determined in (57)) on the draft assembly of the internal reference. Phage genes were annotated using blastp and the PhageWeb database (<http://computationalbiology.ufpa.br/phageweb/>).

Variant annotation using reference strain FPR3757

To compare mutated genes across separated episodes we used blastp to identify homologs of each CD-HIT cluster of mutated genes in USA300 FPR3757. Genes in FPR3757 were further annotated using the database provided in the AureoWiki repository (107), operon annotations of FPR3757 were retrieved from Microbes Online (108). Only protein-altering variants in genes with FPR3757 homologues

(excluding plasmid genes and phage genes) were kept for the analysis of convergence at gene and operon level and the gene enrichment analysis.

Classification of variants.

Mutational and structural variants were classified in to type C>C (within colonising strains), type C>I (between colonising and invasive strains) and type I>I (within invasive strains) as follows: all variants arising in colonising strains were classified as type C>C, while variants among invasive strains were classified as type C>I if they were found in a baseline invasive strain (defined as the oldest invasive strain; when multiple sequences were available at same time, the baseline invasive strain was selected randomly), and as type I>I if they were found between invasive strains but not on the baseline invasive strain. This approach is based on the assumption that co-infection or superinfection is rare, as we have shown previously for bacteraemia (17).

Calculation of the neutrality index.

A modified McDonald-Kreitman table was compiled as described in (109), where a ratio was calculated between non-synonymous, protein-truncating, IS insertions, intergenic and deletion variants and synonymous variants. The neutrality index was obtained by dividing the ratio calculated above for type C>I (colonising-invasive) and type I>I (invasive-invasive) by the ratio for type C>C (colonising-colonising) variants, that were used as reference group. Significance was tested by Fisher test.

dN/dS analysis

We used the R package dNdScv (110) to obtain dN/dS ratios for non-synonymous mutations, indels and missense mutations (stop codons) for all FPR3757 genes, based on variants called when mapping all reads on FPR3757 and after subtracting variants from the internal reference reads and variants in positive where internal reference reads had a low coverage. Since this analysis could be hampered by potential false-positive variants resulting from mapping reads on a single reference (17), we also used our curated list of within-host mutations obtained from episode-specific variant calling to calculate crude dN/dS ratios by dividing the number of protein modifying mutations by the number non-synonymous mutations and computed p values by Fisher exact test as in (25).

Gene and operon enrichment analysis.

We calculated the enrichment of protein-altering mutations across all coding regions of FPR3757 (excluding plasmid genes and phage genes) using the approach described in (12). The variant enrichment per gene i was calculated as follows: $(N_i/L_i)/(\Sigma_n/\Sigma_l)$, where N_i is the number of variants per gene i , L_i is the length of gene i , Σ_n is the total number of variants and Σ_l the total length of the genes. We used Poisson regression to model the number of variants per gene j under the null hypothesis (no enrichment), as defined by the equation $\lambda_0 L_j$, where λ_0 is the expected number of variants in any gene and L_j is the the gene length. Under the alternative hypothesis (enrichment of variant in gene i), the estimated number of variants is $\lambda_i L_i$ for gene i , and $\lambda_1 L_j$ for any other gene j . The model parameters $\lambda_0, \lambda_1, \lambda_i$ were estimated using maximum likelihood and tested for significance using the likelihood ratio test. The genome-wide significance cut-off was calculated using the Bonferroni correction (0.05 divided by the number of unique genes or operons) and

the suggestive significance cut-off (1 divided by the number of unique genes or operons), as implemented for bacterial genome-wide associated studies (GWAS) in (22).

Gene set enrichment analysis

For used the PANNZER platform (111), to retrieve a gene ontology annotation of FPR3757 based on the GO terms. We modified the “antibiotic response” category by adding a curated list of antibiotic resistance genes downloaded from the NCBI AMR gene reference database (70). The gene set enrichment analysis (GSEA) was performed as implemented in the R package clusterProfile (112). Genes with a FPR3757 homolog were ranked according to the significance of the enrichment of protein-modifying mutations (gene enrichment analysis, see above) and the GSEA was carried out with a minimum gene set size of 10 and using the false discovery rate (FDR) method for adjustment for multiple testing.

Mutation co-occurrence analysis

To detect co-occurrence of mutations and mutated genes across independent episodes, we constructed a co-occurrence matrix using the R package cooccur (113). A co-occurrence of mutations or mutated genes in at least two independent episodes was interpreted as convergent and as a sign of potential epistatic interaction. The network of co-occurrence of mutated genes was visualised using the R package ggraph (<https://cran.r-project.org/web/packages/ggraph/index.html>)

Network analysis of adaptation signatures

The pairwise calculation of the Jaccard index between set of mutated genes was performed in R. The calculations were performed both with the entire set of mutated FPR3757 genes and with the 20 most significantly enriched genes in each group of variants. A network of shared mutated genes between independent episodes was constructed using ggraph, where edges represented episode connections based on the Jaccard index. We used the R package tidygraph to extract the node centrality (function centrality_degree()) as a summary measure of the degree of adaptation of the episodes. The network graph and analysis was performed for each group of variants separately.

Data availability

The CAMERA2 isolates reads included in this study are available in the European Nucleotide Archive under Bioproject accession no. PRJEB50796.

Code availability

Scripts used to call mutations and structural variants, annotate all variants and perform the gene enrichment analysis are deposited on github at https://github.com/stefanogg/staph_adaptation_paper

FIGURES

Figure 1. Overview of the *S. aureus* within-host evolution analysis framework. (A) Simulated phylogenetic tree illustrating within-host evolution of *S. aureus* colonisation and infection. This model assumes two genetic bottlenecks (dotted lines); upon transmission, and upon transition from colonisation to invasive infection.

(B) Sites and timing of within-host samples and number of genomes per sample define five prototypes of within-host evolution studies, each with colonising-colonising, colonising-invasive or invasive-invasive comparisons in different combinations: from top to the bottom: multiple colonising samples and one invasive samples; one colonising and one invasive sample; multiple colonising samples; multiple invasive samples; multiple colonising and invasive samples. (C) Approach to capture signals of adaptation across multiple independent episodes of colonisation/infection through detection of multiple genetic mechanisms of adaptation from short reads data and multi-layered functional annotation of the genetic variants using multiple databases including characterisation of intergenic regions (promoters), operon prediction, and gene ontology (GO). (D) Statistical framework for the gene, operon and gene set enrichment analysis. Counts of independent mutations with likely impact on the protein sequence (non-synonymous substitutions, frameshifts, stop codon mutations, IS insertions) were computed for each genes with a FPR3757 homolog. Gene counts (with the addition of intergenic mutations in promoter regions) were aggregated in operons and gene ontologies. Gene and operon counts were used to fit Poisson regression models to infer mutation enrichment and significance of the enrichment. Gene ontologies counts and gene enrichment significance were used to run a gene-set-enrichment analysis (GSEA). To illustrate the approach, the example of the gene *walR* is provided in italic.

Figure 2. (A) Maximum-likelihood phylogenetic tree of 2,590 *S. aureus* sequences included in the study. The tree is annotated (starting from the inner circle) with the most prevalent sequence types (ST), presence/absence of the *mecA* gene, compartment of isolation (colonising or invasive) and year of publication. (B) Summary of 396 independent episodes of *S. aureus* colonisation or infection

categorised according to whether they allowed comparing colonising-colonising (C>C), colonising-invasive (C>I) or invasive-invasive (I>I) strains, or a combination of them. (C) Evidence of a distinctive pattern of adaptation in late infection-adapted strains (type I>I variants). For each type of comparison (type C>C, colonising-colonising; type C>I, colonising-invasive; type I>I, invasive-invasive), the cumulative curves display the accrued number of intergenic mutations, truncating mutations, IS insertions and large deletions as a function of the total number of mutations. Genetic events were counted once per episode, regardless of the number of strains with the mutation. The sequence of mutations events in the cumulative curves is random.

Figure 2 - figure supplement 1. Number of episode-specific variants in same-episode strains having the same sequence type (ST) as the internal reference vs. isolates with a different ST. The dashed line represents the mutation threshold used to remove genetically unrelated strains with the same episode.

Figure 2 - figure supplement 2. Distribution of new IS insertions by classification of the transposase and by major sequence types (ST). (A) Distribution of the 9 major ST among 2,590 strains. (B) Number of independent IS insertions by ST group and type of transposase.

Figure 3. Top 20 genes with the most significant mutation enrichment across the entire dataset. (A) Significance of the enrichment for protein-altering mutations. The dashed line depicts the Bonferroni-corrected significance threshold, red circles and blue circles represent genes with p values below and above the Bonferroni threshold, respectively. (B) Enrichment of protein-altering mutations (C) Bar plots of independent mutations separated in 3 panels according to the type of variant (type C>C: colonising-colonising; type C>I: colonising-invasive; type I>I: invasive-invasive)

and coloured according to the class of mutation. (D) Gene maps with type and positions of mutations.

Figure 3 - figure supplement 1. Mapping of mutations in the 10 most significantly enriched mutated genes across the entire dataset. The maximum-likelihood phylogenetic tree was inferred from the core genome alignment of 2,590 isolates.

Figure 3 - figure supplement 2. dN/dS values for non-synonymous mutations (A), indels (B) and nonsense mutations (stop codons) (B) for FPR3757 genes. Only the 20 most significant genes with positive selection (dN/dS for missense mutations > 1) are shown.

Figure 3 - figure supplement 3. Most frequently deleted genes in large deletions.

Figure 3 - figure supplement 4. Most frequently enriched genes in copy number variations.

Figure 3 - figure supplement 5. Gene convergence analysis of all mutated genes (i.e. including both genes with FPR3757 homolog and no FPR3757 homolog). Top 20 genes with the most significant mutation enrichment across the entire dataset. (A) Significance of the enrichment for protein-altering mutations. The dashed line depicts the Bonferroni-corrected significance threshold, red circles and blue circles represent genes with and without FPR3757 homolog, respectively. (B) Enrichment of protein-altering mutations (C) Bar plots of independent mutations separated in 3 panels according to the type of variant (type C>C: colonising-colonising; type C>I: colonising-invasive; type I>I: invasive-invasive) and coloured according to the class of mutation. (D) Gene maps with type and positions of mutations.

Figure 3 - figure supplement 6. Gene convergence analysis after removing variants in strains included in Young et al, eLIFE 2017, the largest collection of this analysis (1,078 strains, 105 episodes). Top 20 genes with the most significant

mutation enrichment across the entire dataset. (A) Significance of the enrichment for protein-altering mutations. The dashed line depicts the Bonferroni-corrected significance threshold, red circles and blue circles represent genes with p values below and above the Bonferroni threshold, respectively. (B) Enrichment of protein-altering mutations (C) Bar plots of independent mutations separated in 3 panels according to the type of variant (type C>C: colonising-colonising; type C>I: colonising-invasive; type I>I: invasive-invasive) and coloured according to the class of mutation. (D) Gene maps with type and positions of mutations.

Figure 4. Top 20 operons with the most significant mutation enrichment across all dataset. (A) Significance of the enrichment for protein-altering mutations. The dashed line depicts the Bonferroni-corrected significance threshold, red circles and blue circles represent operons with p values below and above the Bonferroni threshold, respectively. (B) Enrichment of protein-altering mutations (C) bar plots of independent mutations separated in 3 panels according to the type of variant (type C>C: colonising-colonising; type C>I: colonising-invasive; type I>I: invasive-invasive) and coloured according to the class of mutation. Mutations were considered independent if they occurred in separate episodes of either colonisation or invasive infection. (D) operon maps with positions of the mutations (relative to the start of the first gene of the operon). Operons are labelled with the names of the genes included, longer labels were shorted for clarity (see Supplementary file 5 for details).

Figure 5. Modified volcano plot displaying enrichment (x axis) and significance of enrichment (y axis) within type C>C, type C>I and type I>I variants. The horizontal dashed line depicts the Bonferroni-corrected significance threshold and dotted line shows the suggestive significance threshold. Labels indicate genes with significance

of enrichment below the suggestive threshold . Genes are coloured in red if the p value is below the Bonferroni-corrected threshold and in blue otherwise.

Figure 5 - figure supplement 1. Modified volcano plot displaying enrichment (x axis) and significance of enrichment (y axis) for FPR3757 operons across the entire dataset, type CC, type CI and type II variants. The horizontal line depicts the Bonferroni-corrected significance threshold. Genes are coloured in red if the p value is below the Bonferroni-corrected threshold and in blue otherwise. Operons are labelled if they were significantly enriched or reached near significance.

Figure 5 - figure supplement 2. Gene-set enrichment analysis (GSEA) for protein-modifying mutations in type CC, type CI and type II variants. (A) Gene ontologies (minimum set size 10 for a total of 110 categories) ordered by normalised enrichment score (NES). Ontologies with negative enrichment were excluded. Dark blue bars indicate a significant p value after false discovery rate (FDR) correction (B) Dot plot of 9 significantly enriched ontologies among type II variants.

Figure 6. Network of mutations co-occurrence. The width and colour of the edges represent the strength of the co-occurrence of mutated genes on the same strain (thin and blue, 2 independent co-occurrences; thick and orange, 3 independent co-occurrences).

Figure 7. Clinical correlates of adaptive signatures within colonising (type C>C, panels A-C) and invasive (type I>I, panels D-F) bacterial populations. Adaptation was inferred by computing the Jaccard index of shared mutated genes between independent episodes, followed by network analysis of infection episodes pairs. The node centrality measure was used as an indicator of adaptation. To avoid overinflation of mutated genes the calculation was limited to the 20 most significantly enriched genes within each group of mutations. (A, D) Density of centrality values

across colonisation (panel A) and infection categories (panel D). (B, E) Number and proportion of adaptive episodes. An adaptive episode was defined by a centrality > 0. (C, F) Distribution of mutations in the 20 most significantly enriched genes across categories of colonisation (panel C) and infection (panel F).

Figure 7 - figure supplement 1. Clinical manifestations and infection sites of invasive episodes, grouped by the infection syndromes classification used for the adaptation analysis.

Figure 7 - figure supplement 2. Network of colonisation/infection episodes for type CC (panel A), type CI (panel B) and type II variants (panel C). Nodes indicate independent episodes, colored based on the clinical syndrome, edges show connections based on shared mutated genes (the width of the connection is proportional to the Jaccard index).

975 TABLES

976 **Table 1.** Microbiological and clinical characteristics of the colonisation and infection
977 episodes included in the within-host evolution analysis

	Strains (n= 2590)	Episodes (n = 396)
Sequence type		
30	342 (13.2%)	43 (10.9%)
22	277 (10.7%)	44 (11.1%)
5	271 (10.5%)	42 (10.6%)
45	198 (7.6%)	38 (9.6%)
15	156 (6.0%)	4 (3.5%)
1	133 (5.1%)	14 (3.5%)
93	110 (4.2%)	29 (7.3%)
8	107 (4.1%)	18 (4.5%)
239	100 (3.9%)	29 (7.3%)
Other	896 (34.6%)	125 (31.6%)
<i>MecA</i> positive	1001 (38.6%)	207 (52.3%)
Infection syndrome		
Skin infection	204 (7.9%)	32 (8.1%)
Osteoarticular infection	77 (3.0%)	17 (4.3%)
Bacteraemia without focus	588 (22.7%)	152 (38.4%)
Bacteraemia with focus	331 (12.8%)	85 (21.5%)
Endocarditis	197 (7.6%)	44 (11.1%)
No invasive strains		66 (16.7%)
Colonisation syndrome		
Nasal carriage	974 (37.6%)	166 (42%)
Cystic fibrosis	57 (2.2%)	9 (2%)

Atopic dermatitis	162 (6.3%)	9 (2%)
No colonising strains		212 (54%)

978

979 **Table 1 – table supplement 1.** List of within-host studies included in the analysis.

980

Author, year	PMID	N independent episodes	N sequences	Ref
Gao W, 2015	28348811	1	18	(23)
Howden BP, 2011	22102812	5	10	(41)
Young BC, 2012	22393007	3	167	(27)
Golubchik, 2013	23658690	13	120	(43)
Burd EM, 2014	24850355	1	3	(44)
Rishishwar L, 2016	27446992	1	3	(33)
Trouillet-Assant S, 2016	26918656	3	6	(29)
Young BC, 2017	29256859	105	1078	(12)
Rouard C, 2018	30275089	1	3	(32)
Langhanki L, 2018	30348081	1	2	(38)
Altman DR, 2018	30061376	10	23	(47)
Giulieri SG, 2018	30103826	49	111	(17)
Benoit JB, 2018	29723202	8	42	(45)
Suligoy CM, 2018	29456969	1	3	(31)
Harkins CP, 2018	28951239	27	248	(42)
Tan X, 2019	30753350	4	8	(30)
Loss G, 2019	31696062	1	2	(36)
Kuroda M, 2019	31474962	1	16	(39)
Azarian T, 2019	31244886	3	44	(46)
Wuethrich D, 2019	30726929	1	5	(28)

Ji S, 2020	32176794	3	20	(40)
Miller CR, 2020	31932377	1	2	(35)
Liu J, 2020	31919223	2	12	(37)
Petrovic Fabijan A, 2020	32066959	12	57	(34)
Tong SYC, 2020	32044943	138	585	(26)

981

982 **Table 2.** Modified McDonald-Kreitman table displaying counts of variants (point
983 mutations and structural variants) and the neutrality index for type C>I and type I>I
984 variants (both compared to type C>C variants).

Classification of variant	Number of variants (<i>Neutrality index</i>)		
	Type C>C	Type C>I	Type I>I
synonymous	381	130	155
non-synonymous	978	300 (<i>0.9</i>)	503 (<i>1.3</i>)*
intergenic	544	197 (<i>1.1</i>)	549 (<i>2.5</i>)**
truncating	197	58 (<i>0.9</i>)	190 (<i>2.4</i>)**
IS insertion	17	6 (<i>1.0</i>)	137 (<i>19.8</i>)**
large deletion	76	17 (<i>0.6</i>)*	122 (<i>3.9</i>)**

985

986

987

988

Values are counts of independent mutations. The neutrality index is shown in brackets in italic.
Significance testing: *: $p < 0.05$; **: $p < 0.005$

989 **Table 3. Genome-wide significant gene signatures of within-host evolution.**

990 The genes shown reached genome-wide significance in the entire dataset or in
991 either type C>C, type C>I or type I>I variants.

Gene	p value (whole dataset)	Description	N independent mutations			Significance
			Type C>C	Type C>I	Type I>I	

<i>agrA</i> **	7.04×10^{-28}	accessory gene regulator protein A	5**	9**	8**	Part of the agr quorum sensing system, which is the master regulator of virulence factors expression in <i>S. aureus</i> . Recurrent mutations associated with invasive disease.
<i>agrC</i> **	2.84×10^{-10}	accessory gene regulator protein C	4	2	6**	Histidine kinase, receptor for extracellular autoactivating peptide (AIP). Phosphorylates <i>agrA</i> .
<i>stp1</i> **	1.13×10^{-7}	protein phosphatase 2C domain-containing protein	3	2	3	Associated with vancomycin resistance
<i>mprF</i> **	4.55×10^{-6}	oxacillin resistance-related FmtC protein	2	0	9**	Main determinant of daptomycin resistance. Association with persistence and immune evasion.
<i>rpoB</i>	7.24×10^{-3}	DNA-directed RNA polymerase subunit beta	1	1	7**	Association with rifampicin resistance, but selection in the absence of rifampicin exposure can happen (R503H). Co-resistance to vancomycin, daptomycin, oxacillin. Association with persistence.

** significant enrichment (above the Bonferroni-corrected cut-off, see methods)

992

Table 3 – table supplement 1. Gene signatures of within-host evolution with suggestive significant enrichment. The genes shown reached the suggestive significance threshold in the entire dataset or in either type C>C, type C>I or type I>I variants.

Gene	p value (whole dataset)	Description	N independent mutations			Significance
			Type C>C	Type C>I	Type I>I	
<i>sucA</i> *	6.82×10^{-5}	2-oxoglutarate dehydrogenase E1 component	6	2	2	Encodes a subunit of the α -ketoglutarate dehydrogenase of the tricarboxylic acid cycle (TCA cycle)
<i>saeR</i> *	1.83×10^{-4}	DNA-binding response regulator SaeR	2	1	2	Regulator component of the saeRS two-component system. Virulence regulation.
<i>accB</i>	4.27×10^{-4}	Biotin carboxyl carrier protein of acetyl-CoA carboxylase	3*	1	0	Part of the fatty acid synthesis pathway of <i>S. aureus</i>
SAUSA300_1856	6.41×10^{-4}	hypothetical protein	4*	0	0	Putative intracellular protease. No data on <i>S. aureus</i>
<i>xpaC</i>	1.38×10^{-3}	hypothetical protein	4*	0	0	Predicted 5-bromo-4-chloroindolyl phosphate hydrolysis protein, no data on <i>S. aureus</i>
<i>rpsJ</i>	1.58×10^{-3}	30S ribosomal protein S10	3*	0	0	Mutations at residues 53-60 are associated with tigecycline resistance, at no apparent fitness cost.
SAUSA300_2399	1.68×10^{-3}	ABC transporter ATP-binding protein	4*	0	0	No data on <i>S. aureus</i>
<i>walR</i>	2.10×10^{-3}	DNA-binding response regulator	1	0	3*	Part of walKR two-component response regulator. Associated with vancomycin resistance.

<i>yjbH</i>	3.55×10^{-3}	Dsba-family protein	1	0	3*	Negative regulator of spx (directs its ClpXP-dependent degradation). Association with antibiotic resistance, virulence regulation and oxidative stress resistance.
<i>purR</i>	3.86×10^{-3}	Pur operon repressor	0	1	3*	<i>purR</i> mutants: increased biofilm formation and virulence in animal model; higher capacity to invade epithelial cells.
<i>era</i>	5.34×10^{-3}	GTP-binding protein Era	0	1	3*	Involved in ribosome assembly and stringent response.
<i>pbp2</i>	7.75×10^{-3}	penicillin binding protein 2	6*	0	0	Role in methicillin resistance (PBP2a synergism). Increased expression after oxacillin exposure.
<i>fakA</i>	9.90×10^{-3}	hypothetical protein	5*	0	0	Fatty acid kinase. Deletion mutant displayed increased virulence in a murine model of skin infection.
<i>sgtB</i>	2.65×10^{-2}	glycosyltransferase	0	0	3*	<i>sgtB</i> mutations in adaptive laboratory evolution experiments upon vancomycin exposure.

* suggestive significant enrichment (above the suggestive significance cut-off, see methods)

999 SUPPLEMENTARY FILES

1000 **Supplementary file 1.** List of colonisation/infection episodes included with
1001 publication data (first author, year, PubMed id), number of strains, sites of collection,
1002 clinical characteristics, classification of colonisation and infection episodes.

1003 **Supplementary file 2.** List of strains included with site and date of collection,
1004 sequence type, presence of the *mecA* gene, information on whether the strain was
1005 designed as internal reference or baseline index strain, mash distance to the internal
1006 reference, number of variants called (as compared to the internal reference),
1007 sequencing metrics.

1008 **Supplementary file 3.** List of variants identified annotated with gene, gene
1009 sequence, FPR3757 homolog, FPR3757 operon. Point mutations, IS insertions,
1010 large deletions and copy number variants are presented separately.

1011 **Supplementary file 4.** Gene enrichment analysis for all mutated genes with a
1012 FPR3757 homolog with number of mutations, gene length, mutation enrichment, p
1013 value based on a Poisson regression to model the number of variants per gene.
1014 Results are presented separately for the complete dataset and for type C>C, type
1015 C>I, and type I>I variants.

1016 **Supplementary file 5.** Operon enrichment analysis for all FPR3757 operons (i.e.
1017 mutated genes that could be assigned to a FPR3757 operon) with number of
1018 mutations, operon length, mutation enrichment, p value based on a Poisson
1019 regression to model the number of variants per operon. Results are presented
1020 separately for the complete dataset and for type C>C, type C>I, and type I>I
1021 variants.

1022 **Supplementary file 6.** Gene set enrichment analysis for mutations in genes
1023 aggregated in gene ontologies (GO) categories with enrichment score, normalised

enrichment score (NES), unadjusted and false-discovery rate (FDR) adjusted p value. Results are presented separately for the complete dataset and for type C>C, type C>I, and type I>I variants.

REFERENCES

1. Tong SY, Davis JS, Eichenberger E, Holland TL, Fowler VG, Jr. *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clin Microbiol Rev.* 2015;28(3):603-61.
2. Wertheim HF, Melles DC, Vos MC, van Leeuwen W, van Belkum A, Verbrugh HA, et al. The role of nasal carriage in *Staphylococcus aureus* infections. *Lancet Infect Dis.* 2005;5(12):751-62.
3. Krismer B, Weidenmaier C, Zipperer A, Peschel A. The commensal lifestyle of *Staphylococcus aureus* and its interactions with the nasal microbiota. *Nat Rev Microbiol.* 2017;15(11):675-87.
4. Proctor RA. Immunity to *Staphylococcus aureus*: Implications for Vaccine Development. *Microbiol Spectr.* 2019;7(4).
5. Kuehl R, Morata L, Meylan S, Mensa J, Soriano A. When antibiotics fail: a clinical and microbiological perspective on antibiotic tolerance and persistence of *Staphylococcus aureus*. *J Antimicrob Chemother.* 2020;75(5):1071-86.
6. Hood MI, Skaar EP. Nutritional immunity: transition metals at the pathogen-host interface. *Nat Rev Microbiol.* 2012;10(8):525-37.
7. Marvig RL, Sommer LM, Molin S, Johansen HK. Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. *Nature Genetics.* 2015;47(1):57-64.
8. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, et al. *Helicobacter pylori* genome evolution during human infection. *Proceedings of the National Academy of Sciences.* 2011;108:5033-8.
9. Lieberman TD, Wilson D, Misra R, Xiong LL, Moodley P, Cohen T, et al. Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated *Mycobacterium tuberculosis*. *Nat Med.* 2016;22(12):1470-4.
10. Chaguza C, Senghore M, Bojang E, Gladstone RA, Lo SW, Tientcheu PE, et al. Within-host microevolution of *Streptococcus pneumoniae* is rapid and adaptive during natural colonisation. *Nat Commun.* 2020;11(1):3442.
11. Paterson GK, Harrison EM, Murray GGR, Welch JJ, Warland JH, Holden MTG, et al. Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. *Nature Communications.* 2015;6(1):6560.
12. Young BC, Wu C-H, Gordon NC, Cole K, Price JR, Liu E, et al. Severe infections emerge from commensal bacteria by adaptive evolution. *eLife.* 2017;6:e30637.
13. Van Tyne D, Manson AL, Huycke MM, Karanicholas J, Earl AM, Gilmore MS. Impact of antibiotic treatment and host innate immune pressure on enterococcal adaptation in the human bloodstream. *Sci Transl Med.* 2019;11(487).

14. Lees JA, Kremer PHC, Manso AS, Croucher NJ, Ferwerda B, Serón MV, et al. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microb Genom.* 2017;3(1):e000103.
15. Giulieri SG, Guerillot R, Kwong JC, Monk IR, Hayes AS, Daniel D, et al. Comprehensive Genomic Investigation of Adaptive Mutations Driving the Low-Level Oxacillin Resistance Phenotype in *Staphylococcus aureus*. *mBio.* 2020;11(6).
16. Howden BP, Johnson PD, Ward PB, Stinear TP, Davies JK. Isolates with low-level vancomycin resistance associated with persistent methicillin-resistant *Staphylococcus aureus* bacteremia. *Antimicrob Agents Chemother.* 2006;50(9):3039-47.
17. Giulieri SG, Baines SL, Guerillot R, Seemann T, Gonçalves da Silva A, Schultz M, et al. Genomic exploration of sequential clinical isolates reveals a distinctive molecular signature of persistent *Staphylococcus aureus* bacteraemia. *Genome Medicine.* 2018;10(1):65.
18. Wheatley R, Diaz Caballero J, Kapel N, de Winter FHR, Jangir P, Quinn A, et al. Rapid evolution and host immunity drive the rise and fall of carbapenem resistance during an acute *Pseudomonas aeruginosa* infection. *Nature Communications.* 2021;12(1):2460.
19. Klemm EJ, Gkrania-Klotsas E, Hadfield J, Forbester JL, Harris SR, Hale C, et al. Emergence of host-adapted *Salmonella* Enteritidis through rapid evolution in an immunocompromised host. *Nature Microbiology.* 2016;1(3):15023.
20. Vargas R, Freschi L, Marin M, Epperson LE, Smith M, Oussenko I, et al. In-host population dynamics of *Mycobacterium tuberculosis* complex during active disease. *Elife.* 2021;10.
21. Gatt YE, Margalit H. Common Adaptive Strategies Underlie Within-Host Evolution of Bacterial Pathogens. *Molecular Biology and Evolution.* 2020.
22. Lees JA, Croucher NJ, Goldblatt D, Nosten F, Parkhill J, Turner C, et al. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *Elife.* 2017;6.
23. Gao W, Monk IR, Tobias NJ, Gladman SL, Seemann T, Stinear TP, et al. Large tandem chromosome expansions facilitate niche adaptation during persistent infection with drug-resistant *Staphylococcus aureus*. *Microb Genom.* 2015;1(2):e000026.
24. Khademi SMH, Sazinas P, Jelsbak L. Within-Host Adaptation Mediated by Intergenic Evolution in *Pseudomonas aeruginosa*. *Genome Biology and Evolution.* 2019;11(5):1385-97.
25. Long DR, Wolter DJ, Lee M, Precit M, McLean K, Holmes E, et al. Polyclonality, Shared Strains, and Convergent Evolution in Chronic CF *S. aureus* Airway Infection. *Am J Respir Crit Care Med.* 2020.
26. Tong SYC, Lye DC, Yahav D, Sud A, Robinson JO, Nelson J, et al. Effect of Vancomycin or Daptomycin With vs Without an Antistaphylococcal β -Lactam on Mortality, Bacteremia, Relapse, or Treatment Failure in Patients With MRSA Bacteremia: A Randomized Clinical Trial. *JAMA.* 2020;323(6):527-37.
27. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, et al. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc Natl Acad Sci U S A.* 2012;109(12):4550-5.
28. Wuthrich D, Cuenod A, Hinic V, Morgenstern M, Khanna N, Egli A, et al. Genomic characterization of inpatient evolution of MRSA resistant to daptomycin, vancomycin and ceftaroline. *J Antimicrob Chemother.* 2019;74(5):1452-4.

29. Trouillet-Assant S, Lelievre L, Martins-Simoes P, Gonzaga L, Tasse J, Valour F, et al. Adaptive processes of *Staphylococcus aureus* isolates during the progression from acute to chronic bone and joint infections in patients. *Cell Microbiol.* 2016;18(10):1405-14.
30. Tan X, Coureuil M, Ramond E, Euphrasie D, Dupuis M, Tros F, et al. Chronic *Staphylococcus aureus* Lung Infection Correlates With Proteogenomic and Metabolic Adaptations Leading to an Increased Intracellular Persistence. *Clin Infect Dis.* 2019;69(11):1937-45.
31. Suligoy CM, Lattar SM, Noto Llana M, Gonzalez CD, Alvarez LP, Robinson DA, et al. Mutation of Agr Is Associated with the Adaptation of *Staphylococcus aureus* to the Host during Chronic Osteomyelitis. *Front Cell Infect Microbiol.* 2018;8:18.
32. Rouard C, Garnier F, Leraut J, Lepointeur M, Rahajamananav L, Languépin J, et al. Emergence and Within-Host Genetic Evolution of Methicillin-Resistant *Staphylococcus aureus* Resistant to Linezolid in a Cystic Fibrosis Patient. *Antimicrob Agents Chemother.* 2018;62(12).
33. Rishishwar L, Kraft CS, Jordan IK. Population Genomics of Reduced Vancomycin Susceptibility in *Staphylococcus aureus*. *mSphere.* 2016;1(4).
34. Petrovic Fabijan A, Lin RCY, Ho J, Maddocks S, Ben Zakour NL, Iredell JR, et al. Safety of bacteriophage therapy in severe *Staphylococcus aureus* infection. *Nat Microbiol.* 2020;5(3):465-72.
35. Miller CR, Dey S, Smolenski PD, Kulkarni PS, Monk JM, Szubin R, et al. Distinct Subpopulations of Intravalvular Methicillin-Resistant *Staphylococcus aureus* with Variable Susceptibility to Daptomycin in Tricuspid Valve Endocarditis. *Antimicrob Agents Chemother.* 2020;64(3).
36. Loss G, Simoes PM, Valour F, Cortes MF, Gonzaga L, Bergot M, et al. *Staphylococcus aureus* Small Colony Variants (SCVs): News From a Chronic Prosthetic Joint Infection. *Front Cell Infect Microbiol.* 2019;9:363.
37. Liu J, Gefen O, Ronin I, Bar-Meir M, Balaban NQ. Effect of tolerance on the evolution of antibiotic resistance under drug combinations. *Science.* 2020;367(6474):200-4.
38. Langhanki L, Berger P, Treffon J, Catania F, Kahl BC, Mellmann A. In vivo competition and horizontal gene transfer among distinct *Staphylococcus aureus* lineages as major drivers for adaptational changes during long-term persistence in humans. *BMC Microbiol.* 2018;18(1):152.
39. Kuroda M, Sekizuka T, Matsui H, Ohsuga J, Ohshima T, Hanaki H. IS256-Mediated Overexpression of the WalkR Two-Component System Regulon Contributes to Reduced Vancomycin Susceptibility in a *Staphylococcus aureus* Clinical Isolate. *Front Microbiol.* 2019;10:1882.
40. Ji S, Jiang S, Wei X, Sun L, Wang H, Zhao F, et al. In-Host Evolution of Daptomycin Resistance and Heteroresistance in Methicillin-Resistant *Staphylococcus aureus* Strains From Three Endocarditis Patients. *J Infect Dis.* 2020;221(Suppl 2):S243-S52.
41. Howden BP, McEvoy CR, Allen DL, Chua K, Gao W, Harrison PF, et al. Evolution of multidrug resistance during *Staphylococcus aureus* infection involves mutation of the essential two component regulator WalkR. *PLoS Pathog.* 2011;7(11):e1002359.
42. Harkins CP, Pettigrew KA, Oravcova K, Gardner J, Hearn RMR, Rice D, et al. The Microevolution and Epidemiology of *Staphylococcus aureus* Colonization during Atopic Eczema Disease Flare. *J Invest Dermatol.* 2018;138(2):336-43.

43. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Lerner-Svensson H, et al. Within-host evolution of *Staphylococcus aureus* during asymptomatic carriage. *PLoS One*. 2013;8(5):e61319.
44. Burd EM, Alam MT, Passalacqua KD, Kalokhe AS, Eaton ME, Satola SW, et al. Development of oxacillin resistance in a patient with recurrent *Staphylococcus aureus* bacteremia. *J Clin Microbiol*. 2014;52(8):3114-7.
45. Benoit JB, Frank DN, Bessesen MT. Genomic evolution of *Staphylococcus aureus* isolates colonizing the nares and progressing to bacteremia. *PLoS One*. 2018;13(5):e0195860.
46. Azarian T, Ridgway JP, Yin Z, David MZ. Long-Term Intrahost Evolution of Methicillin Resistant *Staphylococcus aureus* Among Cystic Fibrosis Patients With Respiratory Carriage. *Front Genet*. 2019;10:546.
47. Altman DR, Sullivan MJ, Chacko KI, Balasubramanian D, Pak TR, Sause WE, et al. Genome Plasticity of *agr*-Defective *Staphylococcus aureus* during Clinical Infection. *Infect Immun*. 2018;86(10).
48. 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(7):e1008931.
49. Toft C, Andersson SG. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat Rev Genet*. 2010;11(7):465-75.
50. McEvoy CR, Tsuji B, Gao W, Seemann T, Porter JL, Doig K, et al. Decreased vancomycin susceptibility in *Staphylococcus aureus* caused by IS256 tempering of *WalKR* expression. *Antimicrob Agents Chemother*. 2013;57(7):3240-9.
51. Schreiber F, Szekat C, Josten M, Sahl HG, Bierbaum G. Antibiotic-induced autoactivation of IS256 in *Staphylococcus aureus*. *Antimicrob Agents Chemother*. 2013;57(12):6381-4.
52. Smyth DS, Kafer JM, Wasserman GA, Velickovic L, Mathema B, Holzman RS, et al. Nasal carriage as a source of *agr*-defective *Staphylococcus aureus* bacteremia. *J Infect Dis*. 2012;206(8):1168-77.
53. Wang Y, Bojer MS, George SE, Wang Z, Jensen PR, Wolz C, et al. Inactivation of TCA cycle enhances *Staphylococcus aureus* persister cell formation in stationary phase. *Sci Rep*. 2018;8(1):10849.
54. Lopatkin AJ, Bening SC, Manson AL, Stokes JM, Kohanski MA, Badran AH, et al. Clinically relevant mutations in core metabolic genes confer antibiotic resistance. *Science*. 2021;371(6531):eaba0862.
55. Phaneuf PV, Yurkovich JT, Heckmann D, Wu M, Sandberg TE, King ZA, et al. Causal mutations from adaptive laboratory evolution are outlined by multiple scales of genome annotations and condition-specificity. *BMC Genomics*. 2020;21(1):514.
56. Mäder U, Nicolas P, Depke M, Pané-Farré J, Debarbouille M, van der Kooi-Pol MM, et al. *Staphylococcus aureus* Transcriptome Architecture: From Laboratory to Infection-Mimicking Conditions. *PLoS genetics*. 2016;12(4):e1005962-e.
57. Prados J, Linder P, Redder P. TSS-EMOTE, a refined protocol for a more complete and less biased global mapping of transcription start sites in bacterial pathogens. *BMC Genomics*. 2016;17(1):849.
58. Both A, Huang J, Qi M, Lausmann C, Weißelberg S, Büttner H, et al. Distinct clonal lineages and within-host diversification shape invasive *Staphylococcus epidermidis* populations. *PLOS Pathogens*. 2021;17(2):e1009304.
59. Suligoy CM, Díaz RE, Gehrke AK, Ring N, Yebra G, Alves J, et al. Acapsular *Staphylococcus aureus* with a non-functional *agr* regains capsule expression after

- passage through the bloodstream in a bacteremia mouse model. *Sci Rep.* 2020;10(1):14108.
60. Machado H, Seif Y, Sakoulas G, Olson CA, Hefner Y, Anand A, et al. Environmental conditions dictate differential evolution of vancomycin resistance in *Staphylococcus aureus*. *Communications Biology.* 2021;4(1):793.
61. Guerillot R, Goncalves da Silva A, Monk I, Giulieri S, Tomita T, Alison E, et al. Convergent Evolution Driven by Rifampin Exacerbates the Global Burden of Drug-Resistant *Staphylococcus aureus*. *mSphere.* 2018;3(1).
62. Austin CM, Garabaglu S, Krute CN, Ridder MJ, Seawell NA, Markiewicz MA, et al. Contribution of YjbH to Virulence Factor Expression and Host Colonization in *Staphylococcus aureus*. *Infection and Immunity.* 2019;87(6):e00155-19.
63. Göhring N, Fedtke I, Xia G, Jorge AM, Pinho MG, Bertsche U, et al. New role of the disulfide stress effector YjbH in β -lactam susceptibility of *Staphylococcus aureus*. *Antimicrob Agents Chemother.* 2011;55(12):5452-8.
64. Renzoni A, Andrey DO, Jousselin A, Barras C, Monod A, Vaudaux P, et al. Whole genome sequencing and complete genetic analysis reveals novel pathways to glycopeptide resistance in *Staphylococcus aureus*. *PLoS One.* 2011;6(6):e21577.
65. Engman J, Rogstam A, Frees D, Ingmer H, von Wachenfeldt C. The YjbH adaptor protein enhances proteolysis of the transcriptional regulator Spx in *Staphylococcus aureus*. *J Bacteriol.* 2012;194(5):1186-94.
66. Paudel A, Panthee S, Hamamoto H, Grunert T, Sekimizu K. YjbH regulates virulence genes expression and oxidative stress resistance in *Staphylococcus aureus*. *Virulence.* 2021;12(1):470-80.
67. Sause WE, Balasubramanian D, Irnov I, Copin R, Sullivan MJ, Sommerfield A, et al. The purine biosynthesis regulator PurR moonlights as a virulence regulator in *Staphylococcus aureus*. *Proc Natl Acad Sci U S A.* 2019;116(27):13563-72.
68. Goncheva MI, Flannagan RS, Sterling BE, Laakso HA, Friedrich NC, Kaiser JC, et al. Stress-induced inactivation of the *Staphylococcus aureus* purine biosynthesis repressor leads to hypervirulence. *Nat Commun.* 2019;10(1):775.
69. Goncheva MI, Flannagan RS, Heinrichs DE. De Novo Purine Biosynthesis Is Required for Intracellular Growth of *Staphylococcus aureus* and for the Hypervirulence Phenotype of a purR Mutant. *Infect Immun.* 2020;88(5).
70. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob Agents Chemother.* 2019;63(11).
71. Levin-Reisman I, Brauner A, Ronin I, Balaban NQ. Epistasis between antibiotic tolerance, persistence, and resistance mutations. *Proc Natl Acad Sci U S A.* 2019;116(29):14734-9.
72. Skwark MJ, Croucher NJ, Puranen S, Chewapreecha C, Pesonen M, Xu YY, et al. Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLoS Genet.* 2017;13(2):e1006508.
73. Wadsworth CB, Arnold BJ, Sater MRA, Grad YH. Azithromycin Resistance through Interspecific Acquisition of an Epistasis-Dependent Efflux Pump Component and Transcriptional Regulator in *Neisseria gonorrhoeae*. *mBio.* 2018;9(4).
74. Yokoyama M, Stevens E, Laabei M, Bacon L, Heesom K, Bayliss S, et al. Epistasis analysis uncovers hidden antibiotic resistance-associated fitness costs hampering the evolution of MRSA. *Genome Biology.* 2018;19(1):94.
75. Phillips PC. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics.* 2008;9(11):855-67.

76. Aseev LV, Boni IV. Extraribosomal functions of bacterial ribosomal proteins. *Molecular Biology*. 2011;45(5):739.
77. Monk IR, Shaikh N, Begg SL, Gajdiss M, Sharkey LKR, Lee JYH, et al. Zinc-binding to the cytoplasmic PAS domain regulates the essential Walk histidine kinase of *Staphylococcus aureus*. *Nature Communications*. 2019;10(1):3067.
78. Cameron DR, Ward DV, Kostoulis X, Howden BP, Moellering RC, Jr., Eliopoulos GM, et al. Serine/threonine phosphatase Stp1 contributes to reduced susceptibility to vancomycin and virulence in *Staphylococcus aureus*. *J Infect Dis*. 2012;205(11):1677-87.
79. Trong HNG, Prunier A-L, Leclercq R. Hypermutable and fluoroquinolone-resistant clinical isolates of *Staphylococcus aureus*. *Antimicrobial agents and chemotherapy*. 2005;49(5):2098-101.
80. Wheeler NE, Gardner PP, Barquist L. Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*. *PLoS Genet*. 2018;14(5):e1007333.
81. Bailey SF, Blanquart F, Bataillon T, Kassen R. What drives parallel evolution?: How population size and mutational variation contribute to repeated evolution. *Bioessays*. 2017;39(1):1-9.
82. Bryant JM, Brown KP, Burbau S, Everall I, Belardinelli JM, Rodriguez-Rincon D, et al. Stepwise pathogenic evolution of *Mycobacterium abscessus*. *Science*. 2021;372(6541).
83. Key FM, Khadka VD, Romo-González C, Blake KJ, Deng L, Lynn TC, et al. On-person adaptive evolution of *Staphylococcus aureus* during atopic dermatitis increases disease severity. *bioRxiv*. 2021:2021.03.24.436824.
84. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. *Nat Rev Microbiol*. 2016;14(3):150-62.
85. McVicker G, Prajsnar TK, Williams A, Wagner NL, Boots M, Renshaw SA, et al. Clonal expansion during *Staphylococcus aureus* infection dynamics reveals the effect of antibiotic intervention. *PLoS Pathog*. 2014;10(2):e1003959.
86. Guerillot R, Li L, Baines S, Howden B, Schultz MB, Seemann T, et al. Comprehensive antibiotic-linked mutation assessment by resistance mutation sequencing (RM-seq). *Genome Med*. 2018;10(1):63.
87. Peleg AY, Miyakis S, Ward DV, Earl AM, Rubio A, Cameron DR, et al. Whole genome characterization of the mechanisms of daptomycin resistance in clinical and laboratory derived isolates of *Staphylococcus aureus*. *PLoS One*. 2012;7(1):e28316.
88. Jiang JH, Dexter C, Cameron DR, Monk IR, Baines SL, Abbott IJ, et al. Evolution of Daptomycin Resistance in Coagulase-Negative Staphylococci Involves Mutations of the Essential Two-Component Regulator WalkR. *Antimicrob Agents Chemother*. 2019;63(3).
89. Guerillot R, Kostoulis X, Donovan L, Li L, Carter GP, Hachani A, et al. Unstable chromosome rearrangements in *Staphylococcus aureus* cause phenotype switching associated with persistent infections. *Proc Natl Acad Sci U S A*. 2019;116(40):20135-40.
90. Laabei M, Uhlemann AC, Lowy FD, Austin ED, Yokoyama M, Ouadi K, et al. Evolutionary Trade-Offs Underlie the Multi-faceted Virulence of *Staphylococcus aureus*. *PLoS Biol*. 2015;13(9):e1002229.
91. Stinear TP, Seemann T, Pidot S, Frigui W, Reysset G, Garnier T, et al. Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. *Genome Res*. 2007;17(2):192-200.

92. Batut B, Knibbe C, Marais G, Daubin V. Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat Rev Microbiol.* 2014;12(12):841-50.
93. Thwaites GE, Gant V. Are bloodstream leukocytes Trojan Horses for the metastasis of *Staphylococcus aureus*? *Nat Rev Microbiol.* 2011;9(3):215-22.
94. Surewaard BG, Deniset JF, Zemp FJ, Amrein M, Otto M, Conly J, et al. Identification and treatment of the *Staphylococcus aureus* reservoir in vivo. *J Exp Med.* 2016;213(7):1141-51.
95. Basuino L, Jousselin A, Alexander JAN, Strynadka NCJ, Pinho MG, Chambers HF, et al. PBP4 activity and its overexpression are necessary for PBP4-mediated high-level β -lactam resistance. *J Antimicrob Chemother.* 2018;73(5):1177-80.
96. Gao W, Cameron DR, Davies JK, Kostoulas X, Stepnell J, Tuck KL, et al. The RpoB H(4)(8)(1)Y 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(6):929-39.
97. Novick RP, Geisinger E. Quorum sensing in staphylococci. *Annu Rev Genet.* 2008;42:541-64.
98. Shopsis B, Drlica-Wagner A, Mathema B, Adhikari RP, Kreiswirth BN, Novick RP. Prevalence of agr dysfunction among colonizing *Staphylococcus aureus* strains. *J Infect Dis.* 2008;198(8):1171-4.
99. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature.* 2017;545(7655):446-51.
100. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biology.* 2019;20(1):257.
101. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30(14):2068-9.
102. Lemoine F, Gascuel O. Gotree/Goalign : Toolkit and Go API to facilitate the development of phylogenetic workflows. *bioRxiv.* 2021:2021.06.09.447704.
103. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics.* 2016;2(4).
104. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution.* 2020;37(5):1530-4.
105. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841-2.
106. Brynildsrud O, Snipen LG, Bohlin J. CNOGpro: detection and quantification of CNVs in prokaryotic whole-genome sequencing data. *Bioinformatics.* 2015;31(11):1708-15.
107. Fuchs S, Mehlan H, Bernhardt J, Hennig A, Michalik S, Surmann K, et al. AureoWiki- The repository of the *Staphylococcus aureus* research and annotation community. *International Journal of Medical Microbiology.* 2018;308(6):558-68.
108. Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, et al. MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.* 2010;38(Database issue):D396-400.
109. Stoletzki N, Eyre-Walker A. Estimation of the Neutrality Index. *Molecular Biology and Evolution.* 2010;28(1):63-70.

110. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*. 2017;171(5):1029-41.e21.
111. Törönen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res*. 2018;46(W1):W84-w8.
112. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*. 2012;16(5):284-7.
113. Griffith DM, Veech JA, Marsh CJ. cooccur: Probabilistic Species Co-Occurrence Analysis in R. 2016. 2016;69(Code Snippet 2):17.

ACKNOWLEDGMENTS

The authors declare that they have no competing interests. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. This work was supported by a Research Fellowship to BPH and TPS from the National Health and Medical Research Council, Australia. SGG was supported by a PhD scholarship of the University of Melbourne. We acknowledge the CAMERA2 investigators and trial steering committee for sharing sequences and clinical metadata of trial participants with multiple sequential bacteraemia strains.

A **B** **C** **D**

A

B

C

Number of episode-specific variants

A

B

A

B

C

agrA

Mutation

- Q46*

I75N

T88M

L131fs

Ter133Lext*?

D137fs

Ter141fs

G148D

I150fs

L162fs

E163fs
- E163Y

H174fs

L186F

F197fs

R195C

H208fs

R218*

R218C

E226fs

I238fs

Y238fs

agrC

Mutation

- D2fs

I30fs

L38fs

F117L

F117LI139fs

S28TI297V

Y168fs

IS insertion (intragenic)

Q175E

L193*

K206fs

stp1

Mutation

- D9V

Q70R

N93_K111del

S136fs

N137D

Q149*

A226V

V247L

mprF

Mutation

- L78F

S295A

S295L

P314L

S337L

I348del

W424C

L451_V452delinsF

T472K

E593D

L776S

L826F

sucA

Mutation

- V42I

H242Q

V425D

I500M

P562T

S622G

S691P

P717H

P717T

D897fs

P909L

saeR

Mutation

- V49_M53del

P74S

A165D

E179K

A190T

accB

Mutation

- V53A

A70T

T112I

I134V

SAUSA300_1856

Mutation

- I32T

G85A

R86*

D157Y

xpaC

Mutation

- I31V

Y132fs

D179N

SAUSA300_1613

Mutation

- I92M

L162F

D171N

I196T

Ter356Yext*?

A

Missense mutations

B

Nonsense mutations

C

Indels

Phage gene

- No
- Yes

Independent deletions

- 1
- 2
- 3
- 4

A

B

C

A

B

C

A

B

C

$-\log_{10}(\text{p value})$

type CC

type CI

type II

Enrichment

$-\log_{10}(\text{p value})$

type CC

type CI

type II

Enrichment

A **B**

A

bioRxiv preprint doi: <https://doi.org/10.1101/2022.02.11.480068>; this version posted February 12, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

B

D

E

C

F

Infection episode

Adaption network of type II variants

Adaptation network of type CC variants

Adaptation network of type CI variants