Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease


Whole-genome sequencing offers new insights into the evolution of bacterial pathogens and the etiology of bacterial disease. *Staphylococcus aureus* is a major cause of bacteria-associated mortality and community-acquired infections that has come to prominence through the rise of drug-resistant forms, notably methicillin-resistant *S. aureus* (MRSA) (1, 2). To better understand the biology of *S. aureus* carriage, we surveyed nasal carriage in more than 1,100 adults attending general practices in Oxfordshire, United Kingdom. We recruited 360 carriers for regular screening. One elderly participant (participant P) developed a *S. aureus* bloodstream infection 15 mo after joining the study. Using whole-genome sequencing, we charted the evolution of the bacterial population during carriage and disease and contrasted it to ongoing evolution in two asymptomatic carriers (Q and R). Our results reveal dynamic populations of nasally carried staphylococci that harbor genetic variation that evolves measurably over time. We found that the number of mutations separating disease-causing from asymptptomatically carried bacteria in participant P was very few. However, the clustering of protein-truncating mutations preceding disease progression, including a transcriptional regulator of stress response and pathogenesis, was a unique pattern absent from the asymptomatic carriers and suggests a role for loss-of-function mutations in bacterial pathogenesis.


Freedly available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the European Nucleotide Archive Sequence Read Archive under study accession number ERP001185 (http://www.ebi.ac.uk/ena/data/view/ERP001185).

1 B.C.Y. and T.G. contributed equally to this work.

2 D.W.C., R.B., and D.J.W. contributed equally to this work.

3 To whom correspondence should be addressed. E-mail: daniel.wilson@ndm.ox.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1113219109/-/DCSupplemental.
Results

Invasive Bloodstream Bacteria Emerged from a Nasal Population of Methicillin-Sensitive *S. aureus*. We used Illumina HiSeq 2000 to sequence the genomes of 68 colonies isolated from six nasal swabs and a blood culture from participant P (Fig. 1A and Table S1). The nasal swabs sent for sequencing fell into two groups: early nasal cultures (ENC) comprised five swabs (N1, N2, N4, N6, and N8) representing prolonged stable carriage over 6 mo. In month 10, a nasal swab showed no growth, suggesting that this stable carriage had been disturbed, possibly in connection with drug intervention in the form of amoxicillin prescribed for a cough in month 9 (although subsequent testing demonstrated resistance to penicillin G). The late nasal cultures (LNC) comprised a single swab (N12) collected in month 12. Further medical intervention followed. In month 13, a B-cell neoplasia with cardiac complications was diagnosed and a permanent pacemaker was fitted.

![Fig. 1. Molecular diversity during progression from carriage to disease in participant P. (A) Sampling frame, variants, and the time line of disease progression. Seven groups of colonies sequenced from six nasal swabs and blood culture are shaded from light to dark gray. (B) The location of virulence factors and variants in the chromosome (outer ring) and plasmid (inner ring). Positions are inferred by mapping to MSSA476 and pSAS. From the outermost track inwards: Virulence-associated surface proteins (dark red), toxins (olive green), and regulatory genes (dark blue) are identified in the bacterial chromosome. Variants detected in the bacterial chromosome are the following by type: synonymous (green), nonsynonymous (orange), premature stop codon (red), and intergenic (gray). Solid lines represent SNPs and dashed lines represent indels. Variants detected in the plasmid are the same color scheme. (C) The maximum-likelihood unrooted tree relating all sequences. Nodes represent genotypes, where area is proportional to sample frequency, and small black circles represent hypothetical intermediate genotypes. Shading within the circles indicates the sample, and darker shading corresponds to later samples as in A. Edges represent mutations, color-coded as in B. The ordering of mutations among hypothetical genotypes is arbitrary. Numbers represent genotypes observed more than once for cross-reference with Fig. 2.](image-url)
under single-dose fluoroaxacin and penicillin prophylaxis. The following month the participant began a chemotherapy regimen consisting of a proteasome inhibitor and an alkylating agent, along with prophylactic antimicrobials (co-trimoxazole). Sixteen days later the participant developed fever and was admitted to hospital with features of septic shock, including neutropenia. At this point, the late blood cultures (LBC) comprising sample B15 were taken. No source for the bacteremia was identified: there were no endovascular catheters or evidence of endocarditis or surgical-site infection from the pacemaker, but the patient developed fatal multiorgan failure.

Genomic analysis and standard molecular typing indicated a typical community-carried S. aureus with no obvious predisposition toward disease. All samples were methicillin-sensitive S. aureus (MSSA), multilocus-sequence type (19) (MLST) ST-15 and spa type t4714 (Table S1), a newly described allele closely related to the commonly carried, community-associated t084 (20). Alignment of genome C1285 from sample N1 to reference genomes MRSA252 and MSSA476 (21) confirmed the absence of the methicillin resistance gene meca and the lack of the staphylococcal cassette chromosome that frequently encodes virulence factors (22) (Fig. S1). Sequence similarity to MSSA476 was 99.5%. No putative virulence factors were found within 13 kb of coding sequence that did not align to MSSA476 (Table S2). Two copies of a 2.7-kb plasmid that was 99.9% similar to MSSA476 pSAS plasmid and that contained a region homologous to the staphylococcal transposon Tn552 were detected. MSSA476 prophages φSa3 and φSa4 were absent from C1285, and no other prophages were detected. Pathogenicity islands homologous to Saα and Saβ were detected with partial deletions. No virulence or toxin genes were detected beyond those that were present in MSSA476 (Dataset S1 and Fig. 1B).

Carriage and Invasive Bacteria Formed Distinct Clades Within the Host. Extremely limited microvariation was found among the 68 sequences (Fig. 1), well below that detectable by conventional methods and consistent with a homogeneous population arising from a single acquisition. There was no variation in MLST, in spa type, or in 61 minisatellite repeats (Dataset S2), suggesting that the resident bacterial population was not eradicated by several episodes of antibiotic treatment. There was no evidence for large-scale insertions/deletions (indels) or copy number variants. We discovered a total of 30 single nucleotide polymorphisms (SNPs) and four short indels in the 2.7-megabase genome comprising 5 synonymous, 16 nonsynonymous, 3 nonsense, and 6 intergenic SNPs (one of which occurred in the plasmid), 2 intergenic indels, and 2 frameshift-inducing indels (one of which led to a premature stop codon) (Table S3). There was no homoplasmy or evidence for within-host recombination.

Bloodstream colonies (LBC) and nasal colonies (ENC/LNC) formed distinct clades within a population characterized by extremely limited genetic variation. Nasal colonies clustered further into those sampled before (ENC) and after (LNC) the first of a series of drug interventions beginning in month 9 (Fig. 1C). Bayesian coalescent analysis showed slow but detectable evolution of the bacterial population (Fig. 2), at a rate of 2.72 mutations per megabase per year [95% credible interval (CI): 1.64–4.42], close to other estimates of the short-term mutation rate in S. aureus (2, 23, 24). A similar molecular clock rate was estimated for ENC sequences alone, but between the ENC clade and the LNC/LBC clades there was greater sequence divergence than expected (Fig. S2), suggesting a departure from the neutral evolutionary model. The estimated effective population size was small, corresponding to an average life span of polymorphisms of 4 mo. The most recent common ancestor of all of the sequences was dated to 1 mo before enrollment (Table S4), but there is no evidence to rule out carriage before this time.

Excess of Protein-Truncating Mutations Preceded Disease Progression. Unusual patterns of molecular evolution were observed along the branches separating early nasal sequences from invasive bloodstream sequences. Tests based on neutral coalescent simulations showed that the most recent common ancestor (MRCA) of ENC and LNC/LBC (Coal.i in Fig. 2 and Table S3) was significantly older than expected (P = 0.005). Indeed, five mutations occurred on this branch, whereas none of the derived polymorphisms from ENC sequences were retained through to LNC/LBC. This may indicate (i) cryptic populations of differentiated bacteria within the host, (ii) resequencing by a latent population of ancestral genotypes, (iii) adaptive evolution, or (iv) relaxed functional constraint associated with a population bottleneck. A similar pattern was seen for the branch leading to the LBC sequences, which coincided with periods of anti-neoplastic chemotherapy and antibiotic treatment. The MRCA of LBC and LNC (Coal.ii in Fig. 2 and Table S3) was also significantly older than expected (P = 0.021). Although LNC sequences were diverse, no SNPs were observed within LBC. This represents significantly reduced diversity (Fig. S3) that is unlikely to be due to sampling limitations because LBC was derived from three separate blood culture bottles.

The distribution of premature stop codons among the branches of the tree was uneven, with a significant excess on the two branches leading from ENC to LBC (Fisher’s exact test, P = 0.0015; Table 1). Four SNPs and one indel separated the ENC clade from the LNC/LBC clades, including a premature stop codon in an ArcC family transcriptional regulator (AFTR), which presents the best current candidate for functional mutation. AFTRs are regulators of carbon metabolism, stress response, and virulence that respond to changing environmental conditions such as antibiotic use and oxidative stress (17). In Neisseria meningitidis, a pseudogene induced by a premature stop codon in the AFTR mpeR, is associated with the hypervirulent ST 32 complex (25).

The mutation that we observed maps to MSSA476 SAS2271, radically truncating the sequence from 702 to 77 amino acids. A premature stop codon induced by a frameshift on the same branch occurred in a protein of unknown function, SAS1429. We observed two further premature stop codons, significantly truncating proteins on the branch leading to the LBC clade: SAS0973, an iron-compound binding protein/transporter, and SAS1361, a GNAT family acetyltransferase.

Clusters of Protein-Truncating Mutations Were Not Observed in Noninvasive Carriage Populations. To investigate the evolution of S. aureus during asymptomatic nasal carriage, we used the Illumina GAIIx and HiSeq 2000 platforms to sequence the whole genomes of 101 colonies isolated from two other participants (Table S1). Twenty-two colonies isolated from two swabs taken at different time points were sequenced for participant Q, who had no history of staphylococcal disease or recent antibiotic use. Seventy-nine colonies isolated from eight swabs taken over an 18-mo period were sequenced from participant P, who similarly had no history of staphylococcal disease. However, participant P had completed a treatment of fluoroaxacin shortly before enrollment in the carriage study and took a course of amoxicillin in month 20. No bacterial growth was detected in any of six nasal swabs taken from month 22 to month 32, suggesting that carriage was cleared. The bacterial populations in both carriers were MSSA, and both carried a single spa type (t164 and t012, respectively), indicating significantly similar multilocus-sequence type (ST-20 and ST-30, respectively), consistent with a single founding colonization in each case. The repertoire of virulence and toxin genes was indistinguishable among genomes sequenced within a single participant, and it was more similar between the genomes sequenced from participants P and Q than those from R (Dataset S1).

Limited microvariation was detected in both participant Q and participant P. We discovered a total of 42 SNPs and 4 short indels in participant Q, comprising 10 synonymous, 20 nonsynonymous, 1 nonsense, and 11 intergenic SNPs, 2 intergenic indels, and 2 frameshift-inducing indels, both of which led to premature stop codons (Table S3). Two large deletions were also detected in one of the genomes: an 8-kb deletion partially matching S. aureus pathogenicity island SaPI4 and a 1.6-kb deletion of an integrase. In participant R, we discovered a total of 39 SNPs and 9 short indels.
comprising 14 synonymous, 15 nonsynonymous, 1 nonsense, and 9 intergenic SNPs, 6 intergenic indels, and three frameshift-inducing indels, all of which led to premature stop codons. There was no significant difference in the overall pattern of mutation types across participants P, Q, and R (Fisher’s exact test, \( P = 0.457 \)). As in participant P, there was no homoplasy or evidence for within-host recombination in participants Q and R.

Rather than forming distinct clusters, the colonies isolated 2 mo apart in participant Q were genetically overlapping, with the descendants from multiple lineages detected in the earlier nasal swab present in the latter (Fig. S4). There was a clearer temporal trend in participant R, such that the diversity sampled at one time was usually descended from a single one of the lineages present in the previous sample, leading to a steady accumulation of mutations over time. Bayesian coalescent analysis revealed a molecular clock rate of 1.87 mutations per megabase per year in participant R (95% CI: 1.08–3.06), consistent with the rate estimated in P. There was insufficient power to independently estimate the rate of evolution in participant Q. Assuming the same clock rate as in R implies a large effective population size of 17 mo. The effective population size estimated for participant R was intermediate between P and Q, with an average life span of polymorphisms of 5 mo (Table S4).

The evidence from participants Q and R provided additional support for the view that a significant excess of protein-truncating mutations occurred on the two branches separating the genomes sampled early during asymptomatic nasal carriage (ENC) from those sampled from the invasive bloodstream infection (LBC) in participant P. Three of 48 mutations detected in participant Q, and 4 of 48 mutations detected in participant R, were protein-truncating. Treating the mutations in Q and R as control groups confirmed that the number of premature stop codons occurring on the ENC-LBC branches in participant P was statistically

### Table 1. Evidence for an excess of protein-truncating mutations on the two branches leading from ENC to LBC

<table>
<thead>
<tr>
<th>Branches</th>
<th>No. of mutations</th>
<th>Protein-truncating</th>
<th>Other</th>
<th>( P ) value (Fisher’s exact test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>From ENC to LBC</td>
<td>4</td>
<td>4</td>
<td></td>
<td>0.0015</td>
</tr>
<tr>
<td>All others in participant P</td>
<td>0</td>
<td>26</td>
<td></td>
<td>0.0055</td>
</tr>
<tr>
<td>All others in participant Q</td>
<td>3</td>
<td>45</td>
<td></td>
<td>0.0103</td>
</tr>
<tr>
<td>All others in participant R</td>
<td>4</td>
<td>44</td>
<td></td>
<td>0.0017</td>
</tr>
<tr>
<td>All others in P, Q, and R</td>
<td>7</td>
<td>115</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. Bayesian coalescent tree. The maximum clade credibility tree representing the genealogy of sequences in the study, reconstructed from SNPs using BEAST. Genotypes are enumerated as in Fig. 1C. SNPs (filled circles) and indels (open circles) are superimposed on the tree and color-coded by type: synonymous (green), nonsynonymous (orange), premature stop codon (red), and intergenic (gray). The ordering of mutations within a branch is arbitrary.
significant (Table 1). To maximize statistical power, we combined information across participants P, Q, and R, yielding a highly significant P value of 0.0017. To further investigate the unusual clustering of premature stop codons in participant P, we constructed an empirical distribution for this P value by considering every possible pair of branches occurring in participants P, Q, and R. None of the 589 possible permutations yielded a P value as significant as 0.0017 (Fig. S5), demonstrating that the cluster of protein-truncating mutations on the two branches leading from ENC to LBC was indeed highly unusual.

Discussion

Just eight mutations accompanied the transition of an asymptptomatically carried MSSA population to a fatal bloodstream infection. Half of these mutations were premature stop codons, one of which truncated a putative transcriptional regulator of virulence (17). Two further premature stop codons were detected only among invasive bloodstream bacteria. Loss-of-function mutations that truncate the amino acid sequence may play an important role in pathogenesis because point mutations of this sort can quickly effect radical functional change (18). However, the patient’s general health was also likely to have been important, with the interaction between genome evolution and clinical context likely to have been critical.

Using whole-genome sequencing of 169 bacterial colonies isolated from three nasal MSSA carriers, we have detected limited but measurable cross-sectional diversity and ongoing evolution within singly colonized carriers that would be undetectable by traditional methods and have charted the evolutionary changes associated with the progression to invasive disease in one individual. High-throughput sequencing offers opportunities for understanding bacterial molecular evolution within the host and promises to shed light on the in vivo dynamics of bacterial carriage and the role of circulating variants. Our results suggest that genetics in invasive bacterial disease is yet to be determined, but the exhaustive characterization of bacterial genetic variation within the host is an important step.

Materials and Methods

Isolate Collection and Preliminary Analysis. Ethical approval for the carriage study was obtained from the Oxfordshire B Oxfordshire Research Ethics Committee (reference no. 08/H0605/02). Each nasal swab culture was prepared and incubated. We included the broth culture of the growth from the subsequent blood culture on SASelect agar (BioRad) overnight at 37 °C, then picked 12 colonies, streaked each onto Columbia blood agar, and incubated the colonies overnight at 37 °C. Methicillin sensitivity was determined by the disk diffusion method. Blood cultures were prepared using the BD Bactec system; we used the inoculum of the broth culture of the growth from each bottle for bacterial growth. Blood extracted from the bottles was cultured on SASelect agar, and four colonies were picked from each bottle for sequencing.

Sequencing and Assembly. Blood cultures were prepared using the BD Bactec system; we used a 1/100 dilution of the broth culture of the growth from each bottle for bacterial growth. Blood extracted from the bottles was cultured on SASelect agar (BioRad) agar, and four colonies were picked from each bottle for sequencing. DNA was extracted using a commercial kit (FastDNA by MP Biomedical) employing mechanical disruption of bacteria and column-based purification of DNA. Staphylococcal protein A (spa) type was determined by Sanger sequencing of the variable X region of the 3′ end of the spa gene, using commercially designed primers (S′-AGACGATCCTTCGGTGAGC-3′; SpaR: 5′-GCTTGGCAATCTTTAGTTCACTTGCC-3′; F1458121 5′-ATACGTGTCACCTAGTCCGC-3′ and R1458121 5′-GGCGGCTTGGTATGCATCG-3′).

Variant Calling. We used SAMtools (29) and Picard (http://picard.sourceforge.net) to call variants from mapping, which we then filtered using criteria including base quality, mapping quality, and depth. We used Cortex (30) to detect SNPs and short indels. Visual inspection of every filtered and unfiltered variant call in participant P was used to manually validate the approach. To detect large deletions relative to the host-specific references, unmapped reads were assembled by Velvet with a hash length of 31 bp.

Experimental Validation. We chose the four protein-truncating mutations detected in participant P for validation using PCR and capillary sequencing. The variants detected at positions 1043150 (C→A), 1458121 (G→A), 1555915 (deletion of A), and 2430183 (C→T) relative to the MSSA476 genome were successfully amplified and sequenced in each of two single-colony isolates from N1 and two single-colony isolates from B15 using the following primer pairs: F1043150 5′-CTTGGTGAAACC-3′ and R1043150 5′-ATGTAACGAT-3′; F1458121 5′-ATACGTGTCACCTAGTCCGC-3′ and R1458121 5′-GGCGGCTTGGTATGCATCG-3′; F1555915 5′-GCAATTGAGTACCTCGC-3′ and R1555915 5′-ACATTAGGTGGTGCGC-3′; and F2430183 5′-GTTGTAACACAAGAACGTAAG-3′ and R2430183 5′-GCTGTAAACCCAGGATTGC-3′. For every variant, both the N1 and both the B15 isolates showed the expected sequence, confirming the existence of the variant in our samples.

Mobile Elements. We used BLAST to search for short flanking sequences of six staphylococcal cassette chromosome (SCC)-associated loci (22) in C1285 using a word size of 16. As a control, we repeated the searches in MRSA252 (21) and MSSA476 using a word size of 16. We used Tandem Repeats Finder (31) to search for MSSA476 phages ϕSa4 and pathogenicity islands using BLAST. We used StaphType and Ridom StaphType (32) to search for MSSA476 inserts and pathogenicity islands using BLAST. We used Blastnt to search for MSSA476 inserts and pathogenicity islands using BLAST.}

Population Genetics Analysis. We used a permutation test for recombination that detects any correlation between physical distance and linkage disequilibrium (41). We inferred tree topology and branch lengths using maximum likelihood (ML) under the assumption of no repeat mutation and homogeneous mutation rates. We used the ML tree to reconstruct haplotypes. We performed Bayesian coalescent inference to estimate evolutionary parameters, including the molecular rate using BEAST (42), assuming constant population size and the Hasegawa, Kishino, and Yano mutation model (43). All validated SNPs were included, together with 1% of invariant nucleotides. Separate analyses of participant P, Q, and R genomes were undertaken, along with separate analyses of ENC alone and LNC and LBC together.

Further analyses of each sample (N1, N2, N4, N6, N8, N12, and B15) within participant P were undertaken to estimate diversity (θ = 2Nμg) for each group separately by fixing μ. For the analysis of participant Q sequences, there was insufficient power to estimate μ; instead, μ was fixed at the rate estimated for participant R. In all cases, we assumed an improper uniform prior on Nμg (the product of effective population size and generation length), an improper uniform prior on μ (mutation rate per day, unless fixed), a uniform prior on nucleotide frequencies, and a log-normal prior on
\( \kappa \) (transition:transversion ratio) with mean 1 and SD 1.25 on the logarithmic scale. Pairs of chains of 10 million iterations each were run, which were sampled every 1,000 iterations with a burn-in of 100,000 iterations removed before merging the chains to obtain final results. We quote the posterior median and (2.5%, 97.5%) quantiles as point estimates and credible intervals, respectively. To obtain the maximum clade credibility tree with BEAST, we used an outgroup constructed with 1% of the fixed differences between MSSA476 and the host-specific internal reference, which allowed us to infer the direction of mutation. To remedy the strong leverage that the outgroup sequence has on estimates of the molecular rate, we assumed an uninformative improper uniform prior on the sampling date of the outgroup sequence. A pair of chains of 600 million iterations were run, which were sampled every 10,000 iterations with a burn-in of 100,000 iterations removed before merging the chains to obtain final results.

**Evolution Associated with Progression to Invasive Disease.** To test whether the number of premature stop mutations occurring on the two branches leading from ENC to LBC was unusual, we used Fisher’s exact test, cross-tabulating the number of protein-truncating premature stop mutations versus all other mutations against the branch on which they occurred: those leading from ENC to LBC versus (i) all others in participant P, (ii) all others in participant Q, (iii) all others in participant R, and (iv) all others combined. To test empirically whether the clustering of protein-truncating mutations on the two branches of the tree leading from ENC to LBC was unusual, we considered all pairs of branches within each participant and calculated a \( P \) value using Fisher’s exact test on the total number of premature stop codons seen in those two branches versus all other branches in all participants. We then compared \( P \) value (iv) to this empirically generated distribution. To test whether the coalescence times for the branches leading from ENC to LBC were unusually ancient, we used coalescent simulations based on the output from BEAST to calculate a predictive \( P \) value under the standard neutral model of evolution. For each branch independently, we calculated the prior probability of observing a coalescent time as long or longer, which was conditional on the rest of the inferred tree. The \( P \) value was taken as a mean over the iterations of the Markov chain Monte Carlo.

**ACKNOWLEDGMENTS.** We thank L. O’Connor, A. S. Walker, P. Piazza, and the Oxford Medical Research Council High Throughput Sequencing Hub Team. This study was supported by the National Institute for Health Research Oxford Biomedical Research Centre and the Modernising Medical Microbiology Consortium, the latter funded under the UK Clinical Research Collaboration Translational Infection Research Initiative supported by the Medical Research Council; Biotechnology and Biological Sciences Research Council; National Institute for Health Research on behalf of the UK Department of Health Grant G0800778; and Wellcome Trust Grant 087646/Z/08/Z. We acknowledge the support of Wellcome Trust core funding through Grant 089790/Z/09/Z. E.P. and I.M. are supported by a fellowship from the UK Health Research Oxford Biomedical Research Centre senior investigators.

19. Connor, A. S. Walker, P. Piazza, and the Oxford Medical Research Council High Throughput Sequencing Hub Team. This study was supported by the National Institute for Health Research Oxford Biomedical Research Centre and the Modernising Medical Microbiology Consortium, the latter funded under the UK Clinical Research Collaboration Translational Infection Research Initiative supported by the Medical Research Council; Biotechnology and Biological Sciences Research Council; National Institute for Health Research on behalf of the UK Department of Health Grant G0800778; and Wellcome Trust Grant 087646/Z/08/Z. We acknowledge the support of Wellcome Trust core funding through Grant 089790/Z/09/Z. E.P. and I.M. are supported by a fellowship from the UK Health Research Oxford Biomedical Research Centre senior investigators.