Severe infections emerge from the microbiome by adaptive evolution

Subject Areas: Genomics & Evolutionary Biology, Microbiology & Infectious Disease

Bernadette C Young*[a,b], Chieh-Hsi Wu[a], N Claire Gordon[a], Kevin Cole[c], James R Price[c,d], Elian Liu[a,b], Anna E Sheppard[a,e], Sanuki Perera[a,b], Jane Charlesworth[a], Tanya Golubchik[a], Zamin Iqbal[f], Rory Bowden[f], Ruth C. Massey[g], John Paul[h,i], Derrick W Crook[a,h,i], Timothy E A Peto[a,i], A Sarah Walker[a,i], Martin J Llewelyn[c,d], David H Wyllie[a,j], Daniel J Wilson*[a,f,k]

[a] Nuffield Department of Medicine, Experimental Medicine Division, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK
[b] Microbiology and Infectious Diseases Department, Oxford University Hospitals NHS Trust, John Radcliffe Hospital, Oxford OX3 9DU, UK
[c] Department of Infectious Diseases and Microbiology, Royal Sussex County Hospital, Brighton BN2 5BE, UK
[d] Department of Global Health and Infection, Brighton and Sussex Medical School, University of Sussex, Falmer BN1 9PS, UK
[e] NIHR Health Protection Unit in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England, Oxford, United Kingdom
[f] Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom
[g] Department of Biology and Biochemistry and The Milner Centre for Evolution, University of Bath, Bath BA2 7AY, United Kingdom
[h] National Infection Service, Public Health England, London, UK
[i] National Institute for Health Research, Oxford Biomedical Research Centre, Oxford, UK
[j] Jenner Institute, Centre for Molecular and Cellular Physiology, Oxford OX3 7BN, UK
[k] Institute for Emerging Infections, Oxford Martin School, University of Oxford, Oxford, OX1 3BD, UK

Author contributions
BCY, study design, sample collection, DNA extraction, bioinformatics, analysis, writing
C-HW, bioinformatics, analysis, writing
NCG, JRP, sample collection, DNA extraction
KC, EL, SP, DNA extraction
AS, JC, TG, ZI, bioinformatics
RB, RCM, study design, interpretation
JP, DWC, TEAP, ASW, MJL, study design, sample collection, interpretation
DHW, study design, analysis
DJW, study design, analysis, writing

*Corresponding authors: Bernadette C Young, Daniel J Wilson
Nuffield Department of Medicine, Experimental Medicine Division, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK
bernadette.young@ndm.ox.ac.uk +44 1865 221918
daniel.wilson@ndm.ox.ac.uk

Keywords: *Staphylococcus aureus*; within-host evolution; microbiome; virulence; infection; pathogen genomics; adaptation

Abstract

Bacteria responsible for the greatest global mortality colonize the human microbiome far more frequently than they cause severe infections. Whether mutation and selection within the microbiome precipitate infection is unknown. To address this question, we investigated *de novo* mutation in 1163 *Staphylococcus aureus* genomes from 105 infected patients with nose-colonization. We report that 72% of the infections emerged from the microbiome, with infecting and nose-colonizing bacteria showing systematic adaptive differences. We found 3.6-fold, 2.9-fold and 2.8-fold enrichments of protein-altering variants in genes responding to *rsp*, which regulates surface antigens and toxicity; *agr*, which regulates quorum-sensing, toxicity and abscess formation; and host-derived antimicrobial peptides, respectively. These adaptive signatures were not observed in healthy carriers and differed from prevailing species-level signals of selection, suggesting disease-associated, short-term, within-host selection pressures. Our results show that infection, like a cancer of the microbiome, emerges through spontaneous adaptive evolution, raising new possibilities for diagnosis and treatment.

Introduction

Communicable diseases remain a leading cause of global mortality, with bacterial pathogens among the greatest concern[1]. However, many of the bacteria imposing the greatest burden of mortality, such as *Staphylococcus aureus*, are frequently found as commensal components of the body's microbiome[2]. For them invasive disease is a relatively uncommon event that is unnecessary[3,4], and perhaps disadvantageous[5], for onward transmission. Genomics is shedding light on important bacterial traits such as host-specificity, toxicity and antimicrobial resistance[6-10]. These approaches offer new opportunities to understand the role of genetics and within-host evolution in the outcome of human interactions with major bacterial pathogens[11].

Several lines of evidence support a plausible role for within-host evolution influencing the virulence of bacterial pathogens. Common bacterial infections, including *S. aureus*, are often associated with colonization of the microbiome by a genetically similar strain[12]. Genome sequencing suggests that bacteria mutate much more quickly than previously accepted, and this confers a potent ability to adapt, for example evolving antimicrobial resistance *de novo* within individual patients[13,14]. Opportunistic pathogens infecting cystic fibrosis patients have been found to rapidly adapt to the lung environment, with strong evidence of parallel evolution across patients[15-19]. However, the selection pressures associated with antimicrobial resistance and opportunistic infections of cystic fibrosis patients may not typify within-host adaptation in common commensal pathogens that have co-evolved with humans for thousands or millions of years[20,21].

Candidate gene studies have demonstrated that certain regions, notably quorum-sensing systems such as the *S. aureus* accessory gene regulator (*agr*), mutate particularly quickly *in vivo* and in culture[22]. The *agr* operon encodes a pheromone that coordinates a shift at higher cell densities from production of surface proteins promoting biofilm formation to production of secreted toxins and proteases promoting inflammation and dispersal[23]. Mutants typically produce the pheromone but no longer respond to it[24]. The evolution of *agr* has been variously ascribed to directional selection[25], balancing selection[26], social cheating[27] and life-history trade-off[28]. However, the role of *agr* mutants in disease progression remains unclear, since they are frequently sampled from both asymptomatic carriage and severe infections[24].

Whole-genome sequencing case studies add weight to the idea that within-host evolution could alter disease propensity. In one persistent *S. aureus* infection, a single mutation was sufficient to permanently activate the stringent stress response, reducing growth, colony size and experimentally measured disease severity[29]. In another patient we found that bloodstream bacteria differed from those initially colonizing the nose by several mutations including loss-of-function of the *rsp* regulator[30]. Functional follow-up revealed that the *rsp* mutant expressed reduced toxicity[31], but maintained the ability to cause disseminated infection[32]. Unexpectedly, we found that bloodstream-infecting bacteria exhibit lower toxicity than nose-colonizing bacteria in general[31]. These results raise the question: does *de novo* mutation and selection within the microbiome contribute systematically to severe infection?

We addressed this question by investigating the genetic variants arising from within-patient evolution of *S. aureus* sampled from 105 patients with concurrent nose microbiome colonization and blood or deep tissue infection. We annotated variants to test for systematic differences between colonizing and invading bacteria. We discovered several groups of genes showing significant enrichments of protein-altering variants indicating adaptive evolution. Similar enrichments were not observed in asymptomatic carriers, nor between unrelated bacteria, indicating they reflect disease-associated, within-host selection pressures. Our results reveal that adaptive evolution of genes involved in toxicity, abscess formation, cell-cell communication and bacterial-host interaction is associated with the transformation of commensal constituents of the microbiome into invasive infections, providing new insights into the mechanisms of disease in a major pathogen.

Results

Infecting bacteria are typically descended from the patient's microbiome

We identified 105 patients suffering severe *S. aureus* infections admitted to hospitals in Oxford and Brighton, England, for whom we could recover contemporaneous nose swabs from admission screening. Of the 105 patients, 55 had bloodstream infections, 37 had soft tissue infections and 13 had bone and joint infections (Table 1). The infection was most often sampled on the same day as the nose, with an interquartile range of 1 day earlier to 2 days later (Table S1).

	Relation of colonizing to infecting bacteria		
Infection sites	Unrelated (≥1104 variants)	Closely related (≤66 variants)	
		Zero shared genotypes	One shared genotype
Bloodstream	4	43	8
Soft tissue	4	23	10
Bone & joint	2	8	3
Total	10	74	21

Table 1. Distribution of infection types and relatedness of nose-colonizing and infecting *S. aureus* among 105 patients revealed by genomic comparison.

To discover *de novo* mutations within and between the nose microbiome and infection site, we whole-genome sequenced 1163 bacterial colonies, a median of 5 per site. We detected single nucleotide polymorphisms (SNPs) and short insertions/deletions (indels) using combined reference-based mapping and *de novo* assembly approaches[30,33,34]. We identified 35 distinct strains, defined by multilocus sequence type (ST), across patients (Table S1). As expected[12], colonizing and infecting bacteria were usually extremely closely related (95 patients), sharing the same ST and differing by at most 66 variants. Unrelated colonizing and infecting bacteria (10 patients) differed by at least 1104 variants – usually many more – and typically possessed distinct STs (e.g. Fig. 1a). After excluding variants differentiating unrelated STs, we catalogued 1322 *de novo* mutations within the 105 patients.

In patients with closely related strains, the within-patient population structure was always consistent with a unique migration event from the nose-colonizing microbiome to the infection site, or occasionally, vice versa. Infecting and colonizing bacteria usually formed closely-related but distinct populations with no shared genotypes (74/95 patients, e.g. Fig. 1b), separated by a mean of 5.7 variants. There was never more than one identical genotype between nose-colonizing and infecting bacteria, (21/95 patients, e.g. Fig. 1c), indicating that the migration event from one population to the other involved a small number of founding bacteria[35,36]. In such patients, the shared genotype likely represents the migrating genotype itself. Population structure did not differ significantly between infection types (p = 0.38, Table 1). Genetic diversity in the nose (mean pairwise distance, π = 2.8 variants) was similar to that previously observed in asymptomatic nasal carriers[33] (Reference Panel I, π = 4.1, p = 0.13), but was significantly lower in the infection site (π = 0.6, $p = 10^{-10.0}$), revealing limited diversification post-infection.

In most patients the infection appeared to be descended from the nose microbiome. We used sequences from other patients and carriers (Reference Panel II) to reconstruct the most recent common ancestor (MRCA) for the 95/105 (90%) patients with related nose-colonizing and infecting bacteria. We thereby distinguished wild type from mutant alleles). In 49 such patients, we could determine the ancestral population. The nose microbiome was likely ancestral in 39/49 (80% of patients with related strains, or 72% of all patients) because all infecting bacteria shared *de novo* mutations in common that distinguished them from the MRCA, whereas nose-colonizing bacteria did not. In 16 of those, confidence was high because both mutant and ancestral alleles were observed in the nose, confirming it as the origin of the *de novo* mutation (e.g. Fig. 1d). Conversely, in 10/49 patients, bacteria colonizing the microbiome were likely descended from blood or deep tissue infections (20% of patients with related strains, or 18% of all patients) (e.g. Fig. 1f). Confidence was high for just three of those patients, and they showed unusually high diversity, suggestive of persistent infections (Supplementary data, P063, P072, P093).

Protein-truncating mutants are over-represented within infected patients

To help identify variants that could increase the propensity of bacteria colonizing the nose microbiome to infect the blood and deep tissue, we reconstructed within-patient phylogenies and classified variants by their position in the phylogeny. Sequencing multiple colonies per site enabled us to classify variants into those representing genuine differences *between* nose-colonizing and infection populations (*B*-class), transient variants within the nose-*colonizing* microbiome population (*C*-class) and transient variants within the *disease*-causing infection population (*D*-class). We hypothesized that B-class variants would be most enriched for virulence-altering variants, if such variants occur (Fig. 1g).

We cross-classified variants by their predicted functional effect: synonymous, non-synonymous or truncating within protein-coding sequences, or non-coding (Table 2, Table S2). As expected, the prevailing tendency of selection within patients was to conserve protein sequences, with d_N/d_S ratios indicating rates of non-synonymous change 0.55, 0.68 and 0.63 times that expected under neutral evolution for B, C and D-class variants respectively.

	Number of variants (Neutrality index)				
Phylogenetic position	Synonymous	Non-synonymous	Protein truncating	Non-coding	Total
Patients with severe infections (*n*=105)					
Between colonization and disease (B-class)	93	265 (1.1)	**39 (3.1)**	140 (1.2)	537
Within colonization (C-class)	93	325 (1.3)	**59 (4.7)**	145 (1.3)	622
Within-disease (D-class)	26	82 (1.2)	**15 (4.3)**	40 (1.3)	163
Total	213	672 (1.2)	**113 (3.9)**	325 (1.3)	1322
Asymptomatic carriers[33] (Reference panel I, for comparison, *n*=13)					
Within colonization (C-class)	37	97	5	45	184

Table 2. Cross-classification of variants within patients by phylogenetic position and predicted functional effect, and comparison to asymptomatic carriers. Neutrality indices[37] are defined as the odds ratio of mutation counts relative to synonymous variants in patients versus asymptomatic carriers (Reference Panel I). Those significant at $p < 0.05$ and $p < 0.005$ are emboldened and underlined respectively

In a longitudinal study of one long-term carrier, we previously reported that a burst of protein-truncating variants punctuated the transition from asymptomatic carriage to invasive infection[30]. Here we found a 3.9-fold over-abundance of protein-truncating variants of all phylogenetic classes in infected patients compared to asymptomatic carriers (Reference Panel I, $p = 0.002$, Table 2), supporting the conclusion that loss-of-function mutations are disproportionately associated with evolution within infected patients. This may reflect a reduction within patients in the efficiency with which selection removes deleterious protein-truncating mutations.

Quorum sensing and cell-adhesion proteins exhibit adaptive evolution between colonizing and infecting bacteria

We hypothesized that variants associated with differential propensity to cause or perpetuate invasive infection would be enriched among the protein-altering B-class variants between the nose and infection site (Fig. 1g). Therefore we aggregated mutations by genes in our reference genome (MRSA252) and tested each gene for an excess of non-synonymous and protein-truncating B-class variants, taking into account the length of the gene. Aggregating by gene was necessary because 1318/1322 variants were unique to single patients. The two exceptions involved non-coding variants arising in two patients each, one B-class variant 130 bases upstream of *azlC*, an azaleucine resistance protein (SAR0010), and one D-class variant 88 bases upstream of *eapH*1, a secreted serine protease inhibitor[38] (SAR2295).

We found a significant excess of five protein-altering B-class variants representing a 58.3-fold enrichment in *agrA*, which encodes the response regulator that mediates activation of the quorum sensing system at high cell densities (p=$10^{−7.5}$, Fig. 2a, Table 3). The *clfB* gene encoding clumping factor B, which binds human fibrinogen and loricrin[39], showed an excess of five protein-altering B-class variants, representing a 15.9-fold enrichment that was marginally significant after multiple testing correction (p=$10^{−4.7}$).

Previously we identified a truncating mutation in the transcriptional regulator *rsp* to be the most likely candidate for involvement in the progression to invasive disease in one long-term nasal carrier[30]. Although we observed just one variant in *rsp* among the 105 patients (3.9-fold enrichment, p=0.27), we found it was a non-synonymous B-class variant resulting in an alanine to proline substitution in the regulator's helix-turn-helix DNA binding domain. In separately published experiments[32], we demonstrated that this and the original mutation induce similar loss-of-function phenotypes which, like *agr* loss-of-function mutants, express reduced toxicity, but maintained an ability to persist, disseminate and cause abscesses *in vivo*.

We found no significant enrichments of protein-altering variants among D-class variants, but we observed a significant excess of six protein-altering C-class variants in *pbp2* which encodes a penicillin binding protein involved in cell wall synthesis (19.0-fold enrichment, p=$10^{−6.0}$, Fig. S1a). Pbp2 is an important target of β-lactam antibiotics[40], revealing adaption – potentially in response to antibiotic treatment – in the nose populations of some patients.

Genes modulated by virulence regulators and antimicrobial peptides show adaptive evolution between colonizing and infecting bacteria

To improve the sensitivity to identify adaptive evolution associated with invasive infection, we developed a gene set enrichment analysis (GSEA) approach in which we tested for enrichments of protein-altering B-class variants among groups of genes. GSEA allowed us to detect signatures of adaptive evolution in groups of related genes that were not apparent when interrogating individual genes.

We grouped genes in two different ways: by gene ontology and by expression pathway. First, we obtained a gene ontology for the reference genome from BioCyc[41], which classifies genes into biological processes, cellular components and molecular functions. There were 552 unique gene ontology groupings of two or more genes. We tested for an enrichment among genes belonging to the ontology, compared to the rest that did not.

Second, we obtained 248 unique expression pathways from the SAMMD database of transcriptional studies[42]. For each expression pathway genes were classified as up-regulated, down-regulated or not differentially regulated in response to experimentally manipulated growth conditions or expression of a regulatory gene. For each expression pathway, we tested for an enrichment in genes that were up- or down-regulated compared to genes not differentially regulated.

The most significant enrichment for protein-altering B-class variants between nose and infection sites occurred in the group of genes down-regulated by the cationic antimicrobial peptide (CAMP) ovispirin-1 (p=$10^{-7.8}$), with a similar enrichment in genes down-regulated by temporin L exposure (p=$10^{-6.9}$ Fig. 2c). Like human CAMPs, the animal-derived ovispirin and temporin compounds inhibit epithelial infections by killing phagocytosed bacteria and mediating inflammatory responses[43]. In response to inhibitory levels of ovispirin and temporin, *agr*, surface-expressed adhesins and secreted toxins are all down-regulated. Collectively, down-regulated genes showed 2.7-fold and 2.8-fold enrichments of adaptive evolution, respectively. Conversely, genes up-regulated in response to CAMPs, including the *vraSR* and *vraDE* cell-wall operons and stress response genes[43], exhibited 0.4-fold and 0.7-fold enrichments (i.e. depletions), respectively (Table 3). Thus, genes undergoing adaptive evolution are strongly inhibited by the CAMP-mediated immune response.

Gene group	No. protein-altering B-class variants		Cumulative length of genes (kb)		Enrichment		Significance (-log$_{10}$ p)
Locus							
agrA	5		0.7		58.27		**7.53**
clfB	5		2.6		15.87		4.70
Total	289		2363.8				
BioCyc Gene Ontology							
Cell wall	18		30.9		5.01		**7.03**
Cell adhesion	13		17.2		6.44		**6.47**
Pathogenesis	31		112.5		2.41		4.44
Total	288		2359.3				
SAMMD Expression Pathway	*Down-regulated*	*Up-regulated*	*Down-regulated*	*Up-regulated*	*Down-regulated*	*Up-regulated*	
Ovispirin-1	40	7	121.2	142.9	2.65	0.39	**7.80**
Temporin L	42	14	125.1	156.1	2.78	0.74	**6.85**
rsp	27	1	61.1	13.7	3.61	0.60	**6.35**
agrA (RN27)	9	30	41.0	85.0	1.83	2.94	**5.57**
VISA-vs-VSSA (Mu50 vs N315)	0	17	0	34.4		3.95	**5.23**
VISA-vs-VSSA (Mu50 vs Mu50-P)	0	17	0	36.7		3.70	**4.90**
VISA-vs-VSSA (isolate pair 2)	14	3	26.9	59.7	4.06	0.39	4.71
Total	275		2093.5				

Table 3. Genes, gene ontologies and expression pathways exhibiting the most significant enrichments or depletions of protein-altering B-class variants separating nose microbiome and infection site bacteria. Enrichments below one represent depletions. The total number of variants and genes available for analysis differed by database. A -log$_{10}$ p-value above 4.8 was considered genome-wide significant (in bold).

Genes belonging to the cell wall ontology showed the second most significant enrichment for adaptive evolution (p=$10^{-7.0}$). Genes contributing to this 5.0-fold enrichment included the immunoglobulin-binding *S. aureus* Protein A (*spa*), the serine rich adhesin for platelets (*sasA*), clumping factors A and B (*clfA, clfB*), fibronectin binding protein A (*fnbA*) and bone sialic acid binding protein (*bbp*). The latter four genes contributed to another statistically significant 6.4-fold enrichment of adaptive protein evolution in the cell adhesion ontology (p=$10^{-6.5}$, Fig. 3). Therefore, there is a general enrichment of surface-expressed antigens undergoing adaptive evolution.

The *rsp* regulon showed the most significant enrichment among gene sets defined by response to individual bacterial regulators (p=$10^{-6.4}$). Genes down-regulated by *rsp* in exponential phase[44], including surface antigens and the urease operon, exhibited a 3.6-fold enrichment for adaptive evolution, while up-regulated genes showed 0.6-fold enrichment. So whereas *rsp* loss-of-function mutants were rare *per se*, genes up-regulated in such mutants were hotspots of within-patient adaptation during invasion. Since expression is a prerequisite for adaptive protein evolution, this implies there are alternative routes by which genes down-regulated by intact *rsp* can be expressed and thereby play an important role within patients other than direct inactivation of *rsp*.

Loss-of-function in *agr* mutants represent one alternative route, since they exhibit similar phenotypes to *rsp* mutants, with reduced toxicity and increased surface antigen expression, albeit reduced ability to form abscesses[32]. We found significant 2.9-fold and 1.8-fold enrichments respectively of genes both up- and down-regulated by *agrA* during stationary phase[45], underlining the influence of adaptive evolution on both secreted and surface-expressed proteins during infection (p=$10^{-5.6}$). We further found 3.7 to 4.0-fold enrichment among genes – including agrA – up-regulated in expression changes induced by mutations conferring vancomycin-intermediate *S. aureus* (VISA) (p=$10^{-5.6}$ and p=$10^{-5.2}$).

Several genes contributed to multiple evolutionary signals, particularly cell-wall anchored proteins involved in adhesion, invasion and immune evasion[39], including *fnbA*, *clfA*, *clfB*, *sasA* and *spa*. These multifactorial, partially overlapping signals suggest a large target for selection in adapting to the within-patient environment (Fig. 3). The fact that we observed no comparable significant enrichments in C-class and D-class protein-altering variants (Fig. S1) indicates that these evolutionary patterns are associated specifically with invasion.

Adaptive evolution is evident in both the nose and infection site during severe infections

Having identified adaptive evolution differentiating nose-colonizing and disease-causing bacteria, we next asked whether the mutant alleles were preferentially found in the nose or infection site. We used sequences from other patients or carriers (Reference Panel II) to reconstruct the genotype of the MRCA of colonizing and infecting bacteria respectively in each patient. This allowed us to sub-classify B-class variants by whether the mutant allele was found in the nose-colonizing bacteria (B_C-class) or the disease-causing bacteria (B_D-class). *A priori*, we had expected the enrichments of adaptive evolution to be driven primarily by mutants occurring in the disease-causing bacteria (B_D-class). But instead we found that the most significantly enriched gene sets were driven by mutant alleles occurring both in colonizing and infecting bacteria (Fig. S2). This indicates there are common selection pressures in the nose and infection site in severely infected patients, leading to convergent evolution across body sites.

The group of genes showing the strongest disparity in signal of enrichment among B_D-class vs B_C-class variants was the pathogenesis ontology. Genes involved in pathogenesis were near genome-wide significance in B_D-class variants, showing a 3.1-fold enrichment ($p = 10^{-4.6}$) and a statistically insignificant 1.7-fold enrichment in B_C-class variants (p=0.13). B_D-class mutants driving this differential signal arose in toxins including gamma haemolysin and several regulatory loci implicated in toxicity and virulence regulation (*rot*, *sarS* and *saeR*). Therefore most, but not all, drivers of adaptive evolution within severely infected patients are as likely to favour mutants in nose-colonizing bacteria as infecting bacteria.

Signals of invasion-associated evolution are specific to infected patients and differ from prevailing signatures of selection

Two lines of evidence show that the newly discovered signatures of within-host adaptive evolution are unique to evolution in infected patients. To test the robustness of our conclusions against the alternative explanation that our approach merely detects the most rapidly evolving proteins, we searched for similar signals in alternative settings: evolution within asymptomatic carriers, and species-level evolution between unrelated bacteria.

There was no significant enrichment of protein-altering variants in any gene, ontology or pathway among 235 variants identified from 10 longitudinally sampled asymptomatic nasal carriers (Reference Panel III, Fig. S3, Table S3). To address the modest sample size, we performed goodness-of-fit tests, focusing on the signals most significantly enriched in patients. We found significant depletions of protein-altering variants in carriers relative to patients in the *rsp*, *agr* and *sarA* regulons (p=$10^{-4.0}$) and the pathogenesis ontology (p=$10^{-3.2}$, Table S4).

Nor were the relative rates of non-synonymous to synonymous substitution (d_N/d_S) higher between unrelated *S. aureus* (Reference Panel IV) in the genes that contributed most to the signals associated with invasion within patients: *agrA*, *agrC clfA*, *clfB*, *fnbA* and *sasA*. Although synonymous diversity was somewhat higher than typical in these genes, the d_N/d_S ratios showed no evidence for excess protein-altering change in these compared to other genes (Fig. S4). Accordingly, incorporating this locus-specific variability of d_N/d_S into the GSEA did not affect the results (Fig. S5). Taken together these lines of evidence show that the ontologies, pathways and genes significantly differentiated between colonizing and infecting bacteria are a signature specific to evolution within infected patients, and are not repeated in asymptomatic carriers or species-level evolution.

a

Discussion

We have discovered that common, life-threatening infections of *S. aureus* are frequently descended from bacteria colonizing the human microbiome. These infections are associated with repeatable patterns of bacterial evolution driven by within-patient mutation and selection. The strongest signatures of adaptation occurred in genes responding to cationic antimicrobial peptides and the virulence regulators *rsp* and *agr*. Such genes mediate toxicity, abscess formation, immune evasion and bacterial-host binding. Adaptation within both regulator and effector genes reveals that multiple, alternative evolutionary paths are associated with the transition from microbiome colonization to invasive infection.

The signatures of within-patient adaptation that we found differed from prevailing signals of selection at the species level. This discordance means that infection-associated adaptive mutations within patients are rarely transmitted, and argues against a straightforward host-pathogen arms race as the predominant evolutionary force acting within and between patients. Instead, it supports the notion of a life-history trade-off between adaptations favouring colonization and infection distinct from those favouring dissemination and onward transmission. As such, invasive disease may be analogous to cancer in multicellular organisms, representing an ever-present risk of mutations in the microbiome favoured by short-term selection but ultimately incidental or damaging to the bacterial reproductive life cycle.

The existence of signatures of adaptive substitutions associated with invasive disease raises the possibility of developing new diagnostic techniques and personalizing treatment to the individual patient's microbiome. The ability of genomics to characterize the selective forces driving adaption within the human body in unprecedented detail provides new opportunities to improve experimental models of disease. Ultimately, it may be possible to develop therapies that utilize our new understanding of within-patient evolution to target the root causes of invasive disease from the bacterial perspective.

Materials and methods

Patient sample collection. 105 patients with severe *S. aureus* infections for whom the organism could be cultured from both admission screening nasal swab and clinical sample were identified from the microbiological laboratories of hospitals in Oxford and Brighton, England. This study design builds in robustness to potential confounders by matching disease-causing and nose-colonizing bacteria within the same patients. Clinical samples comprised blood cultures (*n* = 55) and pus, soft tissue, bone or joint samples (*n* = 50). The bacteria sampled and sequenced from one patient ('patient S', P005 in this study) have been previously described[32]. Five individuals had both blood and another culture-positive clinical sample; we focus analysis on the blood sample. Nasal swabs were incubated in 5% NaCl broth overnight at 37C, then streaked onto SASelect agar (BioRad) and incubated overnight at 37C. We picked five colonies per sample (twelve during the pilot phase involving nine patients), streaked each onto Columbia blood agar and incubated overnight at 37C for DNA extraction. Clinical samples were handled according to the local laboratory standard operating procedure for pus, sterile site and blood cultures. When bacterial growth was confirmed as *S. aureus*, the primary culture plate was retrieved, and multiple colonies were picked. These were streaked onto Columbia blood agar and incubated overnight at 37C for DNA extraction. Sequencing multiple colonies per site allowed us to distinguish genuine genetic differences between nose-colonizing and disease-causing bacteria from transient variants.

Reference Panels. For comparison to the patient-derived isolates, we collated previously described samples from other sources to construct four Reference Panels: I. A collection of 131 genomes capturing cross-sectional diversity in the noses of 13 asymptomatic carriers[33], arising from the same Oxfordshire carriage study (BioProject PRJEB2881). II. A compilation of 95 unrelated samples from a carriage study of *S. aureus* in Oxfordshire[48] (BioProject accession number PRJEB255), 145 sequences from a study of within-host evolution of *S. aureus* in 3 individuals[30] (BioProject PRJEB2892) and 909 sequences from nasal carriage and bloodstream infection used in a study of whole genome sequencing to predict antimicrobial resistance[49] (BioProject PRJEB6251). We used these samples to improve our reconstruction of ancestral genotypes in each patient. III. A collection of 237 genomes from longitudinal samples from 10 patients[33,50], (BioProject PRJEB2862) arising from the same Oxfordshire carriage study. We used these to compare evolution within patients and asymptomatic carriers. IV. A collection of 16 previously-published high-quality closed reference genomes, comprising unrelated isolates mainly of clinical and animal origin: MRSA252 (Genbank accession number BX571856.1), MSSA476 (BX571857.1), COL (CP000046.1), NCTC 8325 (CP000253.1), Mu50 (BA000017.4), N315 (BA000018.3), USA300_FPR3757 (CP000255.1), JH1 (CP000736.1), Newman (AP009351.1), TW20 (FN433596.1), S0385 (AM990992.1), JKD6159 (CP002114.2), RF122 (AJ938182.1), ED133 (CP001996.1), ED98 (CP001781.1), EMRSA15 (HE681097.1)[51-63]. We used these to contrast species-level evolution to within-patient evolution.

Whole genome sequencing. For each bacterial colony, DNA was extracted from the subcultured plate using a mechanical lysis step (FastPrep; MPBiomedicals, Santa Ana, CA) followed by a commercial kit (QuickGene; Fujifilm, Tokyo, Japan), and sequenced at the Wellcome Trust Centre for Human Genetics, Oxford on the Illumina (San Diego, California, USA) HiSeq 2000 platform, with paired-end reads 101 base pairs for 9 patients in the pilot phase, and 150 bases in the remainder.

Variant calling. We used Velvet[64] to assemble reads into contigs *de novo*, and Stampy[65] to map reads against two reference genomes: MRSA252[51] and a patient-specific reference comprising the contigs assembled for one colony sampled from each patient's nose. Repetitive regions, defined by BLASTing[66] the reference genome against itself, were masked prior to variant calling. To obtain multilocus sequence types[67] we used BLAST to find the relevant loci, and looked up the nucleotide sequences in the online database at http://saureus.mlst.net/.

Bases called at each position in the reference and those passing previously described[30,33,68] quality filters were used to identify single nucleotide polymorphisms (SNPs) from Stampy-based mapping to MRSA252 and the patient-specific reference genomes. We used Cortex[34] to identify SNPs and short indels. Variants found by Cortex were excluded if they had fewer than ten supporting reads or if the base call was heterozygous at more than 5% of reads.

Where physically clustered variants with the same pattern of presence/absence across genomes were found, these were considered likely to represent a single evolutionary event: tandem repeat mutation or recombination. These were de-duplicated to a single variant to avoid inflating evidence of evolutionary events in these regions.

Variant annotation and phylogenetic classification. Maximum likelihood trees were built to infer bacterial relationships within patients[69]. To prioritize variants for further analysis, they were classified according to their phylogenetic position in the tree: B-class (between colonization and disease), C-class (within colonizing population) and D-class (within disease population). Variants were cross-classified by their predicted functional effect based on mapping to the reference genome or BLASTing to a reference allele: synonymous, non-synonymous or truncating for protein-coding sequences, or non-coding.

Where variation was found using a patient-specific reference, these variants were annotated by first aligning to MRSA252 using Mauve[70]. If no aligned position in MRSA252 could be found, additional annotated references were used. Where variation was found using Cortex only, the variant was annotated by first locating it by comparing the flanking sequence to MRSA252 and other annotated references using BLAST. MRSA252 orthologs were identified using geneDB[71] and KEGG[72].

Reconstructing ancestral genotypes per patient. We constructed a species-level phylogeny for all bacteria sampled from the 105 patients together with Reference Panel

II (unrelated asymptomatic carriage isolates and bacteraemia isolates) using a two-step neighbour-joining and maximum likelihood approach, based on a whole-genome alignment derived from mapping all genomes to MRSA252. We first clustered individuals into seven groups using neighbour-joining[73], before resolving the relationships within each cluster by building a maximum likelihood tree using RAxML[74], assuming a general time reversible (GTR) model. To overcome a limitation in the presence of divergent sequences whereby RAxML fixes a minimum branch length that may be longer than a single substitution event, we fine-tuned the estimates of branch lengths using ClonalFrameML[75]. We used these subtrees to identify, for each patient, the most closely related 'nearest neighbour' sampled from another patient or carrier. We employed this nearest neighbour as an outgroup, and used the tree to reconstruct the sequence of the MRCA of colonizing and infecting bacteria for each patient using a maximum likelihood method[76] in ClonalFrameML[75]. This in turn allowed us to identify the ancestral (wild type) and derived (mutant) allele for all variants mapping to MRSA252. For variants not mapping to MRSA252, we repeated the Cortex variant calling analysis, including the nearest neighbour, and identified the ancestral allele as the one possessed by the nearest neighbour. This approach allowed us to identify ancestral versus derived alleles for 97% of within-patient variants. We used the reconstructions of the within-patient MRCA sequences and identity of ancestral vs derived alleles to sub-categorize B-class variants into those in which the mutant allele was found in the colonizing population (B_C-class) versus the disease-causing population (B_D-class). 521 (97%) of B-class variants were typeable, and in 281 (54%) of these, the mutant allele was found in the disease population. This allowed us to test for differential enrichments in these two sub-classes.

Mean pairwise genetic diversity. Separately for the nose site and infection site of each patient, we calculated the mean pairwise diversity π as the mean number of variants differing between each pair of genomes. We compared the distributions of π between patients and Reference Panel II (13 cross-sectionally sampled asymptomatic carriers) using a Mann-Whitney-Wilcoxon test.

Calculating d_N/d_S ratio. For assessing the d_N/d_S ratio within patients, we adjusted the ratio of raw counts of total numbers of non-synonymous and synonymous SNPs by the ratio expected under strict neutrality. We estimated that the rate of non-synonymous mutation was 4.9 times higher than that of synonymous mutation in *S. aureus* based on codon usage in MRSA252 and the observed transition:transversion ratio in non-coding SNPs.

The Neutrality Index. To compare the relative d_N/d_S ratios between two groups of variants we computed a Neutrality Index as R_1/R_2 where R_1 and R_2 were the ratio of counts of non-synonymous to synonymous variants in each group respectively[37]. We compared B, C and D-class variants within patients to C-class patients within Reference Panel I (13 cross-sectionally sampled asymptomatic carriers). A Neutrality Index in excess of one indicates a higher d_N/d_S ratio in the former group. We used Fisher's exact test to evaluate the significance of the differences between the groups.

Gene enrichment analysis. To test for significant enrichment of variants in a particular gene, we employed a Poisson regression in which we modelled the expected numbers of *de novo* variants across patients in any gene j as $\lambda_0 L_j$ under the null hypothesis of no enrichment, where λ_0 gives the expected number of variants per kilobase and L_j is the length of gene j in kilobases. We compared this to the alternative hypothesis in which the expected number of variants was $\lambda_i L_i$ for gene i, the gene of interest, and $\lambda_1 L_j$ for any other gene j. Using R[77], we estimated the parameters λ_0, λ_1 and λ_i from the data by maximum likelihood and tested for significance via a likelihood ratio test with one degree of freedom. This procedure assumes no recombination within patients, which was reasonable since we found little evidence of recombination in this study or previously[33], including no within-host genetic incompatibilities, and we removed physically clustered variants associated with possible recombination events. We analysed all protein-coding genes in MRSA252, testing for an enrichment of variants expected to alter the transcribed protein (both non-synonymous and truncating mutations). These tests were also applied to synonymous mutations and no enrichments were found.

Gene set enrichment analysis. Since the number of genes outweighed the number of variants detected, we had limited power to detect weak to modest enrichments at the individual gene level. Instead we pooled genes using ontologies from the BioCyc MRSA252 database[41] and expression pathways from the SAMMD database of transcriptional studies[42]. The BioCyc database comprises ontologies describing biological processes, cellular components and molecular functions. The SAMMD database groups genes up-regulated, down-regulated or not differentially regulated in response to experimentally manipulated growth conditions or isogenic mutations, usually of a regulatory gene. After excluding ontologies or pathways with two groups, one involving a single gene, and combining ontologies or pathways with identical groupings of genes, we conducted 800 GSEAs in addition to the 2650 ontologies comprised of individual loci. The number of groupings of genes was always two for BioCyc (included/excluded from the ontology) and two or three for SAMMD (up-/down-/un-differentially regulated in the experiment). Again we employed a Poisson regression in which we modelled the expected numbers of variants in any gene j as $\lambda_0 L_j$ under the null hypothesis of no enrichment where λ_0 gives the expected number of variants per kilobase and L_j is the length of gene j in kilobases. We compared this to the alternative hypothesis in which the expected number of variants was $\lambda_1 L_j$, $\lambda_2 L_j$ or $\lambda_3 L_j$ for gene j depending on the grouping in the ontology/pathway. Using R, we estimated the parameters λ_0, λ_1, λ_2 and λ_3 from the data by maximum likelihood and tested for significance via a likelihood ratio test with one or two degrees of freedom, depending on the number of groupings in the ontology/pathway. To account for the multiplicity of testing, we adjusted the *p-value* significance thresholds from a nominal $\alpha = 0.05$ using the Bonferroni method. This gave an adjusted threshold $10^{-4.8}$.

Longitudinal evolution in asymptomatic carriers. To test whether the patterns of evolution we observed between colonizing and invading bacteria in severely infected patients were typical or unusual, we analysed Reference Panel III (a collection of 10 longitudinally sampled asymptomatic carriers). Since natural selection is more efficacious over longer periods of time, the longitudinal sampling of these individuals gave us greater opportunity to detect subtle evolutionary patterns in asymptomatic carriers. We characterized variation in these carriers as in the patients. Given the modest sample size and smaller number of variants detected in these individuals (235), we performed GSEA to test for enrichments only in particular genes, ontologies and pathways that were significantly enriched within patients, requiring less stringent multiple testing correction.

omegaMap analysis. We estimated d_N/d_S ratios between unrelated *S. aureus* to characterize the prevailing patterns of selection at the species level. We used Mauve[70] to pairwise align 15 reference genomes against MRSA252, i.e. Reference Panel IV. This allowed us to distinguish orthologs from paralogs in the next step in which we multiply aligned all coding sequences overlapping those in MRSA252 using PAGAN[78]. After removing sequences with premature stop codons, we analysed each alignment of between two and 16 genes using a modification of omegaMap[79], assuming all sites were unlinked. We previously showed this assumption, which confers substantial computational efficiency savings, does not adversely affect estimates of selection coefficients[80]. We estimated variation in d_N/d_S within genes using Monte Carlo Markov chain, running each chain for 10,000 iterations. We assumed exponential prior distributions on the population scaled mutation rate (θ), the transition:transversion ratio (κ) and the d_N/d_S ratio (ω) with means 0.05, 3 and 0.2 respectively. We assumed equal codon frequencies and a mean of 30 contiguous codons sharing the same d_N/d_S ratio. For each gene, we computed the posterior mean d_N/d_S ratio across sites. This allowed us to rank the relative strength of selection across genes in MRSA252, and to account for differences in d_N/d_S, as well as gene length, in the GSEA. We achieved this by modifying the expected number of variants in gene j to be $\lambda_0\omega_jL_j$ under the null hypothesis of no enrichment versus $\lambda_1\omega_jL_j$, $\lambda_2\omega_jL_j$ or $\lambda_3\omega_jL_j$ under the alternative hypothesis depending on the ontology or pathway, where ω_j is the posterior mean d_N/d_S in gene j.

Ethical framework. Ethical approval for linking genetic sequences of *S. aureus* isolates to patient data without individual patient consent in Oxford and Brighton in the U.K. was obtained from Berkshire Ethics Committee (10/H0505/83) and the U.K. Health Research Agency [8-05(e)/2010].

Accession numbers. (data to be uploaded). RNA-Seq data relating to isolate from P005 (aka 'patient S') previously submitted under BioProject PRJNA279958.

Acknowledgements

We would like to thank Ed Feil, Stephen Leslie, Gil McVean and Richard Moxon for helpful insights and useful discussions. The views expressed in this publication are those of the authors and not necessarily those of the funders.

Funding information

This study was supported by the Oxford NIHR Biomedical Research Centre, a Mérieux Research Grant, the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with Public Health England (PHE) (grant HPRU-2012-10041), and the Health Innovation Challenge Fund (a parallel funding partnership between the Wellcome Trust (grant WT098615/Z/12/Z) and the Department of Health (grant HICF-T5-358)).

T.E.P. and D.W.C. are NIHR Senior Investigators. D.J.W. and Z.I. are Sir Henry Dale Fellows, jointly funded by the Wellcome Trust and the Royal Society (Grants 101237/Z/13/Z and 102541/Z/13/Z). B.C.Y is a Research Training Fellow funded by the Wellcome Trust (Grant 101611/Z/13/Z). We acknowledge the support of Wellcome Trust Centre for Human Genetics core funding (Grant 090532/Z/09/Z).

References

1. GBD 2015 Mortality and Causes of Death Collaborators. 2016. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: A systematic analysis for the global burden of disease study 2015 Lancet 388(10053):1459-544
2. Turnbaugh PJ, et al. 2007. The human microbiome project Nature 449(7164):804-810
3. Casadevall A, Fang FC, Pirofski LA. 2011. Microbial virulence as an emergent property: Consequences and opportunities PLoS Pathog 7(7):e1002136
4. Methot PO and Alizon S. 2014. What is a pathogen? Toward a process view of host-parasite interactions Virulence 5(8):775-85
5. Brown SP, Cornforth DM, Mideo N. 2012. Evolution of virulence in opportunistic pathogens: Generalism, plasticity, and control Trends Microbiol 20(7):336-42
6. Sheppard SK, et al. 2013. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter Proc Natl Acad Sci U S A 110(29):11923-7
7. Laabei M, et al. 2014. Predicting the virulence of MRSA from its genome sequence Genome Res 24(5):839-49
8. Chewapreecha C, et al. 2014. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes PLoS Genet 10(8):e1004547
9. Chen PE and Shapiro BJ. 2015. The advent of genome-wide association studies for bacteria Curr Opin Microbiol 25:17-24
10. Earle SG, et al. 2016. Identifying lineage effects when controlling for population structure improves power in bacterial association studies Nat Microbiol 1:16041
11. Didelot X, et al. 2016. Within-host evolution of bacterial pathogens Nat Rev Microbiol 14(3):150-62
12. von Eiff C, et al. 2001. Nasal carriage as a source of Staphylococcus aureus bacteremia. N Engl J Med 344(1):11-6
13. Howden BP, et al. 2011. Evolution of multidrug resistance during Staphylococcus aureus infection involves mutation of the essential two component regulator WalKR PLoS Pathog 7(11):e1002359
14. Eldholm V, et al. 2014. Evolution of extensively drug-resistant mycobacterium tuberculosis from a susceptible ancestor in a single patient Genome Biol 15(11):490,014-0490-3
15. Lieberman TD, et al. 2011. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes Nat Genet 43(12):1275-80
16. Marvig RL, Johansen HK, Molin S, Jelsbak L. 2013. Genome analysis of a transmissible lineage of Pseudomonas aeruginosa reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators PLoS Genet 9(9):e1003741
17. Markussen T, et al. 2014. Environmental heterogeneity drives within-host diversification and evolution of Pseudomonas aeruginosa Mbio 5(5):e01592-14
18. Lieberman TD, et al. 2014. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures Nat Genet 46(1):82-7
19. Marvig RL, Sommer LM, Molin S, Johansen HK. 2015. Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis Nat Genet 47(1):57-64
20. Moeller AH, et al. 2016. Cospeciation of gut microbiota with hominids Science 353(6297):380-2
21. Lees JA, et al 2016. Large scale genomic analsyis shows no evidence for pathogen adatption between the blood and cerebrospinal fluid niches during bacterial meningitis Mgen
22. Traber KE, et al. 2008. Agr function in clinical Staphylococcus aureus isolates. Microbiology 154(8):2265-74

23. Novick RP and Geisinger E. 2008. Quorum sensing in staphylococci Annu Rev Genet 42:541-64
24. Painter KL, Krishna A, Wigneshweraraj S, Edwards AM. 2014. What role does the quorum-sensing accessory gene regulator system play during Staphylococcus aureus bacteremia? Trends Microbiol 22(12):676-85
25. Sakoulas G, Moise PA, Rybak MJ. 2009. Accessory gene regulator dysfunction: An advantage for Staphylococcus aureus in health-care settings? J Infect Dis 199(10):1558-9
26. Robinson DA, et al. 2005. Evolutionary genetics of the accessory gene regulator (agr) locus in Staphylococcus aureus J Bacteriol 187(24):8312-21
27. Pollitt EJ, et al. 2014. Cooperation, quorum sensing, and evolution of virulence in Staphylococcus aureus Infect Immun 82(3):1045-51.
28. Shopsin B, et al. 2010. Mutations in agr do not persist in natural populations of methicillin-resistant Staphylococcus aureus. J Infect Dis 202(10):1593
29. Gao W, et al. 2010. Two novel point mutations in clinical Staphylococcus aureus reduce linezolid susceptibility and switch on the stringent response to promote persistent infection PLoS Pathog 6(6):e1000944
30. Young BC, et al. 2012. Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. Proc Natl Acad Sci U S A 109(12):4550
31. Laabei M, et al. 2015. Evolutionary trade-offs underlie the multi-faceted virulence of Staphylococcus aureus PLoS Biol 13(9):e1002229
32. Das S, et al. 2016. Natural mutations in a Staphylococcus aureus virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation Proc Natl Acad Sci U S A 113(22):E3101-10
33. Golubchik T, et al. 2013. Within- host evolution of Staphylococcus aureus during asymptomatic carriage. PloS One 8(5):e61319
34. Iqbal Z, et al. 2012. De novo assembly and genotyping of variants using colored de bruijn graphs Nat Genet 44(2):226-32
35. Moxon ER and Murphy PA. 1978. Haemophilus influenzae bacteremia and meningitis resulting from survival of a single organism Proc Natl Acad Sci U S A 75(3):1534-6
36. Margolis E and Levin BR. 2007. Within-host evolution for the invasiveness of commensal bacteria: An experimental study of bacteremias resulting from Haemophilus influenzae nasal carriage J Infect Dis 196(7):1068-1075
37. Rand DM and Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Mol Biol Evol 13(6):735-48
38. Stapels DA, et al. 2014 Staphylococcus aureus secretes a unique class of neutrophil serine protease inhibitors. Proc Natl Acad Sci U S A. 111(36):13187-92
39. Foster TJ, Geoghegan JA, Ganesh VK, Höök M. 2013. Adhesion, invasion and evasion: The many functions of the surface proteins of Staphylococcus aureus Nature Reviews Microbiology 12(1):49-62
40. Leski TA and Tomasz A. 2005. Role of penicillin-binding protein 2 (PBP2) in the antibiotic susceptibility and cell wall cross-linking of Staphylococcus aureus: Evidence for the cooperative functioning of PBP2, PBP4, and PBP2A J Bacteriol 187(5):1815-1824
41. Caspi R, et al. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases Nucleic Acids Res 44(D1):D471-80
42. Nagarajan V and Elasri M. 2007. SAMMD: Staphylococcus aureus microarray meta-database. BMC Genomics 8(1):351
43. Pietiainen M, et al. 2009. Transcriptome analysis of the responses of Staphylococcus aureus to antimicrobial peptides and characterization of the roles of vraDE and vraSR in antimicrobial resistance BMC Genomics 10:429,2164-10-429
44. Lei MG, et al. 2011. Rsp inhibits attachment and biofilm formation by repressing fnbA in Staphylococcus aureus MW2. J Bacteriol 193(19):5231
45. Dunman PM, et al. 2001. Transcription profiling-based identification of Staphylococcus aureus genes regulated by the agr and/or sarA loci J Bacteriol 183(24):7341-53

46. Cui L, et al. 2005. DNA microarray-based identification of genes associated with glycopeptide resistance in Staphylococcus aureus. Antimicrob Agents Chemother. 49(8):3404-13
47. Mongodin E, et al. 2003. Microarray transcription analysis of clinical Staphylococcus aureus isolates resistant to vancomycin. J Bacteriol. 185(15):4638-43
48. Everitt RG, et al. 2014. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus Nat Commun 5:3956
49. Gordon NC, et al. 2014. Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing J Clin Microbiol 52(4):1182-91
50. Gordon NC, et al. 2016. Whole genome sequencing reveals the contribution of long-term carriers in Staphylococcus aureus outbreak investigation. (Under review, submitted as accompanying manuscript)
51. Holden MTG, et al. 2004. Complete genomes of two clinical Staphylococcus aureus strains: Evidence for the rapid evolution of virulence and drug resistance Proceedings of the National Academy of Sciences 101(26):9786-9791
52. Gill SR, et al. 2005. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant Staphylococcus aureus strain and a biofilm-producing methicillin-resistant Staphylococcus epidermidis strain J Bacteriol 187(7):2426-2438.
53. Gillaspy AF, et al. 2006. The staphylococcus aureus NCTC8325 genome. In: Gram positive pathogens. Fischetti V, Novick R, Ferretti J, et al, editors. 1st ed. Washington, DC: ASM Press. 381-412
54. Kuroda M, et al. 2001. Whole genome sequencing of meticillin-resistant Staphylococcus aureus Lancet 357(9264):1225-40
55. Diep BA, et al. 2006. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus Lancet 367(9512):731-9
56. Baba T, et al. 2008. Genome sequence of Staphylococcus aureus strain newman and comparative analysis of staphylococcal genomes: Polymorphism and evolution of two major pathogenicity islands J Bacteriol 190(1):300-10
57. Holden MT, et al. 2010. Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant Staphylococcus aureus, sequence type 239 (TW) J Bacteriol 192(3):888-92
58. Schijffelen MJ, Boel CH, van Strijp JA, Fluit AC. 2010. Whole genome analysis of a livestock-associated methicillin-resistant Staphylococcus aureus ST398 isolate from a case of human endocarditis BMC Genomics 11:376,2164-11-376
59. Chua K, et al. 2010. Complete genome sequence of Staphylococcus aureus strain JKD6159, a unique australian clone of ST93-IV community methicillin-resistant Staphylococcus aureus J Bacteriol 192(20):5556-7)
60. Herron-Olson L, Fitzgerald JR, Musser JM, Kapur V. 2007. Molecular correlates of host specialization in Staphylococcus aureus PLoS One 2(10):e1120
61. Guinane CM, et al. 2010. Evolutionary genomics of Staphylococcus aureus reveals insights into the origin and molecular basis of ruminant host adaptation Genome Biol Evol 2:454-66
62. Lowder BV, et al. 2009. Recent human-to-poultry host jump, adaptation, and pandemic spread of Staphylococcus aureus Proc Natl Acad Sci U S A 106(46):19545-50.
63. Holden MT, et al. 2013. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic Genome Res 23(4):653-64
64. Zerbino DR and Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de bruijn graphs Genome Res 18(5):821-9.
65. Lunter G and Goodson M. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of illumina sequence reads Genome Res 21(6):936-9
66. Altschul SF, et al. 1990. Basic local alignment search tool J Mol Biol 215(3):403-10

67. Enright MC, et al. 2000. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus J Clin Microbiol 38(3):1008-15
68. Didelot X, et al. 2012. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission Genome Biol 13(12):R118,2012-13-12-r118.
69. Gusfield, D. 1991. Efficient algorithms for inferring evolutionary trees, Networks 21:19-28
70. Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements Genome Res 14(7):1394-403
71. Logan-Klumpler FJ, et al. 2012. GeneDB--an annotation database for pathogens Nucleic Acids Res 40(Database issue):D98-108
72. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation Nucleic Acids Res 44(D1):D457-62
73. Saitou N and Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees Mol Biol Evol 4(4):406-25.
74. Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies Bioinformatics 30(9):1312-3.
75. Didelot X and Wilson DJ. 2015. ClonalFrameML: Efficient inference of recombination in whole bacterial genomes PLoS Comput Biol 11(2):e1004041
76. Pupko T, Pe'er I, Shamir R, Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences Mol Biol Evol 17(6):890-6
77. R Core Team. 2015. R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/
78. Loytynoja A, Vilella AJ, Goldman N. 2012. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm Bioinformatics 28(13):1684-91
79. Wilson DJ and McVean G. 2006. Estimating diversifying selection and functional constraint in the presence of recombination Genetics 172(3):1411-25
80. Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M. 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences PLoS Genet 7(12):e1002395

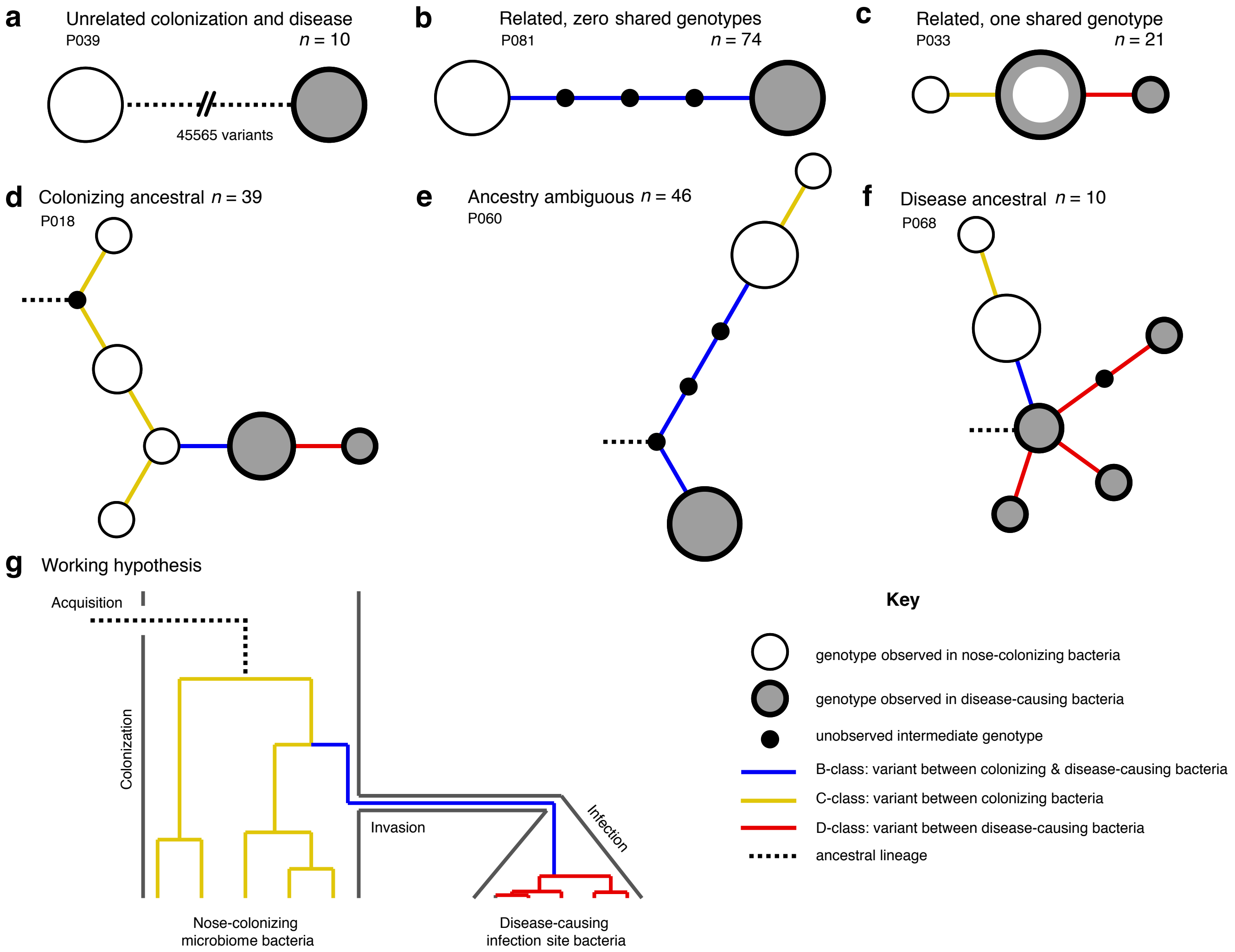
a
Unrelated colonization and disease
P039
n = 10
45565 variants
b
Related, zero shared genotypes
P081
n = 74
c
Related, one shared genotype
P033
n = 21
d
Colonizing ancestral n = 39
P018
e
Ancestry ambiguous n = 46
P060
f
Disease ancestral n = 10
P068
g
Working hypothesis
Acquisition
Colonization
Invasion
Infection
Nose-colonizing microbiome bacteria
Disease-causing infection site bacteria
Key
genotype observed in nose-colonizing bacteria
genotype observed in disease-causing bacteria
unobserved intermediate genotype
B-class: variant between colonizing & disease-causing bacteria
C-class: variant between colonizing bacteria
D-class: variant between disease-causing bacteria
ancestral lineage

Figure 1. Disease-causing *S. aureus* form closely related but distinct populations descended from microbiome-colonizing bacteria in the majority of infections. Bacteria sampled from the nose and infection site of 105 patients formed one of three population structures, illustrated with example haplotrees: **a)** Unrelated populations differentiated by many variants. **b)** Highly related populations separated by few variants. **c)** Highly related populations with one genotype in common. Reconstructing the ancestral genotype in each patient helped identify the ancestral population: **d)** Nose-colonizing bacteria ancestral. **e)** Ambiguous ancestral population. **f)** Disease-causing bacteria ancestral. **g)** Phylogeny illustrating the working hypothesis that variants differentiating highly related nose-colonizing and disease-causing bacteria would be enriched for variants that promote infection. In **a-f**, haplotree nodes represent observed genotypes sampled from the nose (white) or infection site (grey), with area proportional to genotype frequency, or unobserved intermediate genotypes (black). Edges represent mutations. Patient identifiers and sample sizes (*n*) are given. In **a-g**, edge colour indicates that mutations occurring on those branches correspond to B-class variants between nose-colonizing and disease-causing bacteria (blue), C-class variants among nose-colonizing bacteria (gold) or D-class variants among disease-causing bacteria (red). Black dashed edges indicate ancestral lineages.

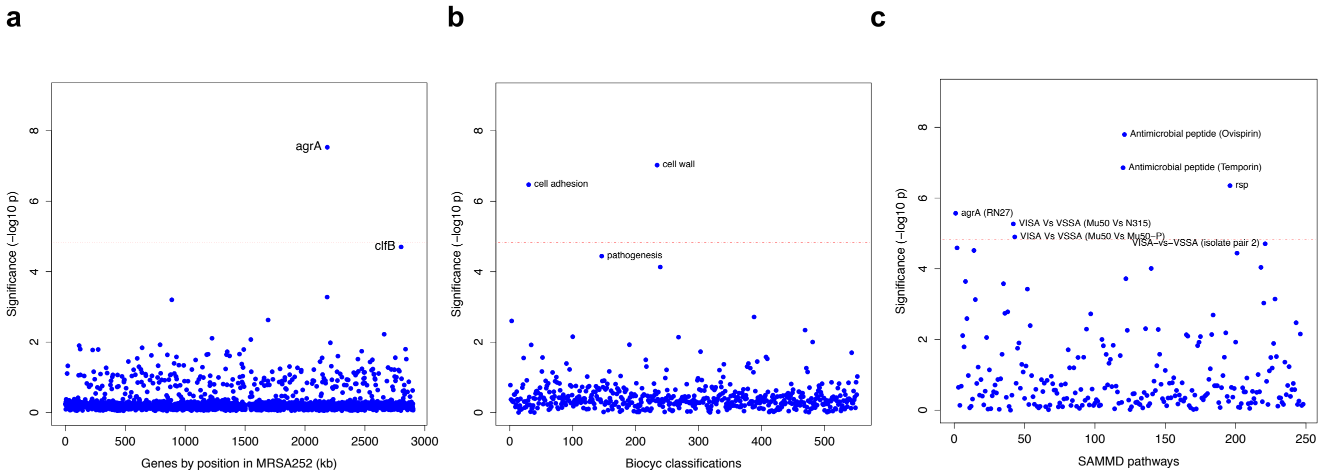

Figure 2. Genes, ontologies and pathways enriched for protein-altering substitutions between nose-colonizing and disease-causing bacteria within infected patients. a) Significance of enrichment of 2650 individual genes. **b)** Significance of enrichment of 552 gene sets defined by BioCyc gene ontologies. **c)** Significance of enrichment of 248 gene sets defined by SAMMD expression pathways. Genes, pathways and ontologies that approach or exceed a Bonferroni-corrected significance threshold of $\alpha = 0.05$ (red lines) are named.

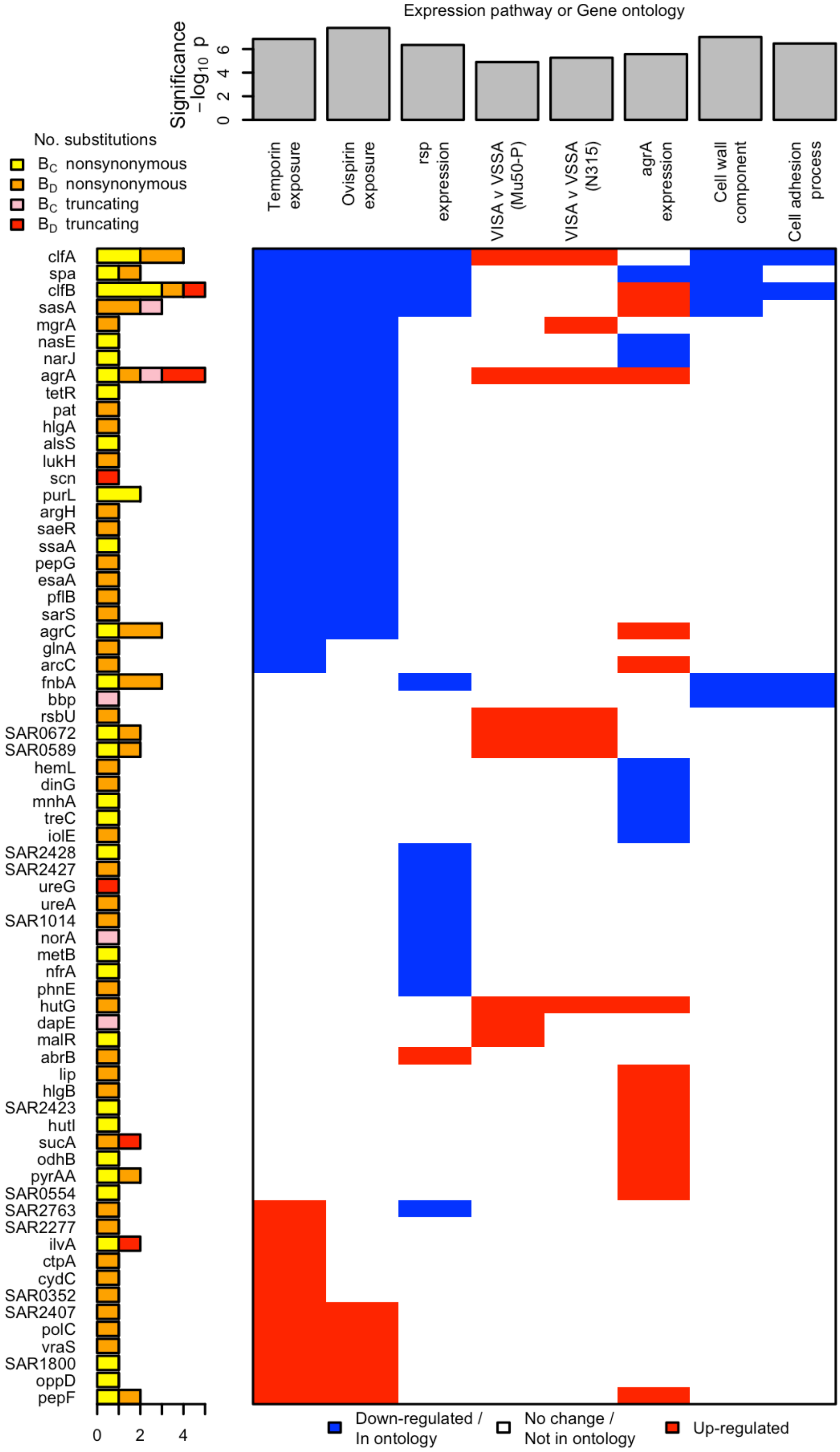

Figure 3. All genes contributing to the pathways and ontologies most significantly enriched for protein-altering substitutions between nose-colonizing and disease-causing bacteria. Every gene with at least one substitution between nose-colonizing and disease-causing bacteria and which was up- (red) or down- regulated (blue) in a significantly enriched pathway or a member of a significantly enriched ontology (blue) is shown. Above, the significance (-$\log_{10}$ *p*-value) of the enrichment is shown. To the left, the number of altering (yellow/orange) and truncating (pink/red) B-class variants is shown, broken down by the population in which the mutant allele was found: nose (B_C; yellow/pink) or infection site (B_D; orange/red).

Table S1. List of all cultures included in the site, the site of infection (and any known source if bloodstream), number of isolates sequenced from each site, ST or CC by in silico MLST, number of variants found at each site and the mean pair-wise difference comparing isolates.

Table S2. List of all variants found within patients with *S. aureus* disease, location on shared reference (MRSA252), or position and reference genome name and accession number if variant could not be localised on MRSA252. Each variant is described by the alleles found, its location in gene, the predicted effect on gene product and the location of the variant on the phylogenetic tree.

Table S3. List of all variants found within long term asymptomatic carriers, location on shared reference (MRSA252), or position and reference genome name and accession number if variant was not localised on MRSA252. Each variant is described by the alleles found, its location in gene and the predicted effect on gene product.

Gene Ontology or Expression Pathway (Loci with protein-altering B_D-class variants within patients)	Number of variants*		*p*-value
	Within patients	Within carriers	
AgrA locus (SAR2126)	3/156	0/115	n.s.
Rsp transcriptional pathway (spa, SAR0143, clfA, SAR1014, SAR1745, ureA, ureG, SAR2427, fnbA, clfB, sasA, SAR2763)	16/147	0/109	0.0001 ***
SarA transcriptional pathway (SAR0109, spa, SAR0211, pyrAA, SAR1397, agrC, agrA, SAR2245, SAR2420, SAR2430, hlgB, fnbA, arcC, sasA, lip)	20/147	1/109 (agrC)	0.0001 ***
AgrA transcriptional pathway (spa, SAR0211, pyrAA, SAR1397, sucA, SAR1466, hemL, agrC, agrA, SAR2430, hlgB, hlgC, clfB, arcC, sasA, lip)	21/147	1/109 (agrC)	<0.0001 ***
Cell wall (spa, clfA, fnbA, clfB, sasA)	9/156	0/115	0.01 *
Cell adhesion (clfA, fnbA, clfB)	6/156	0/115	0.04 *
Pathogenesis (spa, SAR0115, SAR280, SAR0464, SAR0739, saeR, clfA, ebh, rot, SAR2035, SAR2448, hlgA, hlgB, fnbA, clfB, sasA)	21/156	2/115 (ebh)	0.0006**

Table S4. For all ontologies showing enrichment in within-patient B_D-class variants, we identified the genes with variants contributing to the signal. We counted the number of protein-altering variants in these genes within patients, and compared to the number in long-term asymptomatic carriers. P values calculated using Fisher's exact test. *Variant totals are different for SAMMD pathways (*rsp, agrA, sarA*) and Biocyc ontologies (cell wall, cell adhesion, pathogenesis) because pathway information is available for a different number of loci in each database.

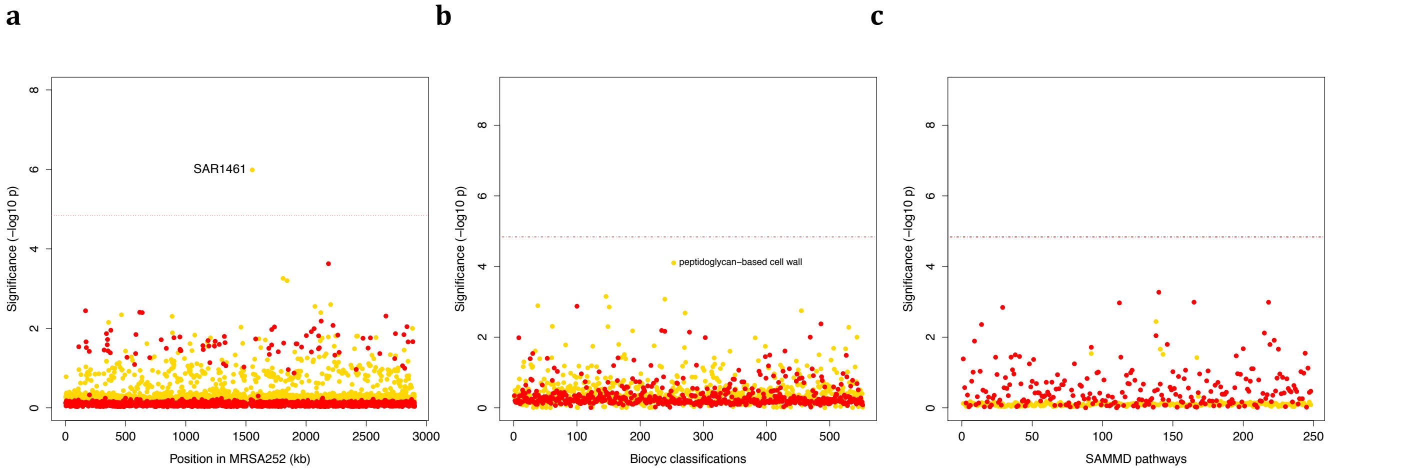

Fig. S1. Genes, ontologies and pathways enriched for protein-altering transient variants within nose-colonizing and disease-causing bacteria. a) Significance of enrichment of 2650 individual genes. **b)** Significance of enrichment of 552 gene sets defined by BioCyc gene ontologies. **c)** Significance of enrichment of 248 gene sets defined by SAMMD expression pathways. C-class variants among nose-colonizing bacteria are coloured gold, D-class variants among disease-causing bacteria are coloured red. Genes, pathways and ontologies that approach or exceed a Bonferroni-corrected significance threshold of $\alpha = 0.05$ (red lines) are named.

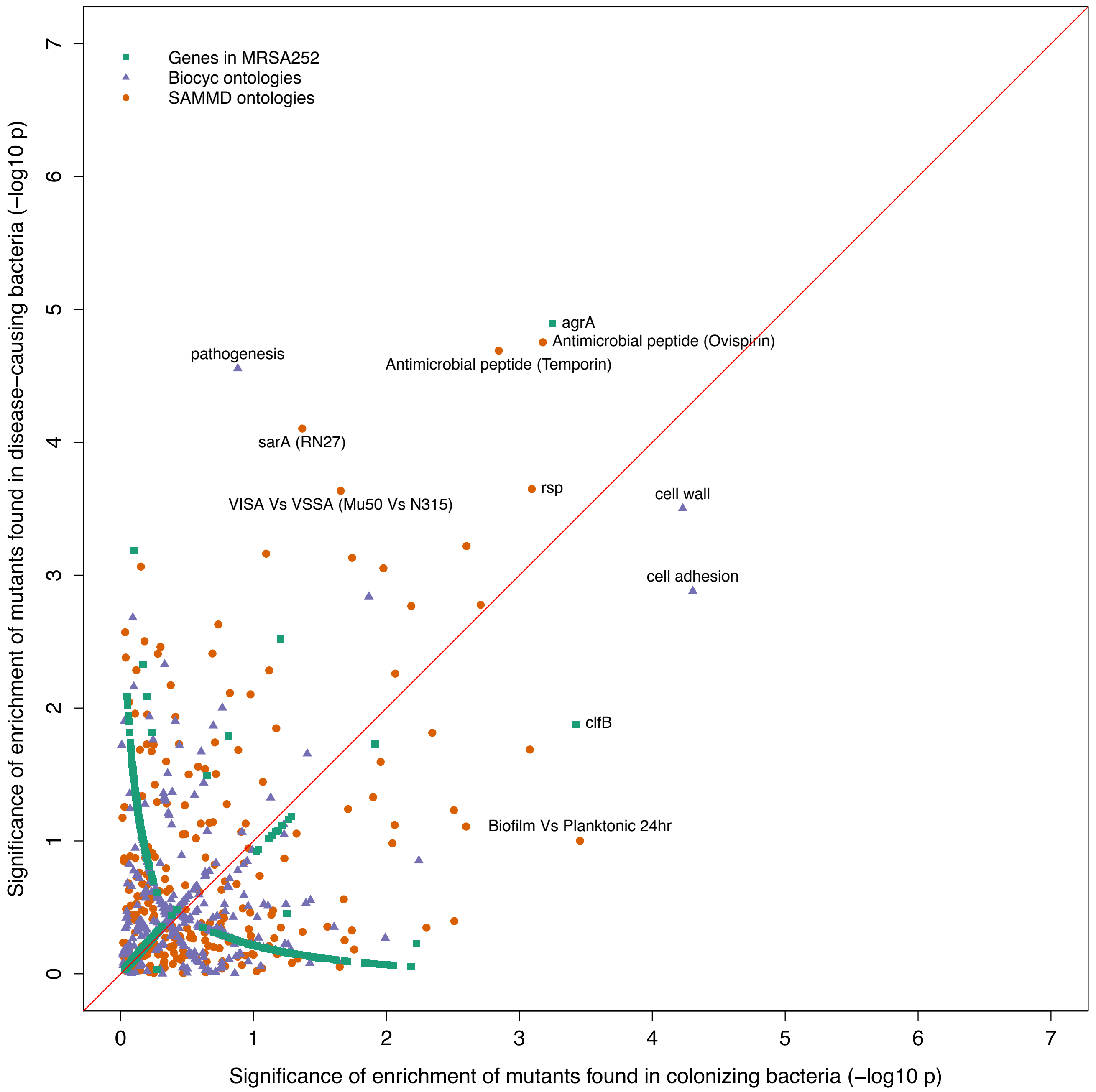

Fig. S2. Gene set enrichment analysis of B-class mutants occurring in the nose or the infection site. Each point indicates the –log10 *p*-values of two tests for enrichment of protein-altering variants found among mutants in nose-colonizing bacteria vs disease-causing bacteria. The shape of each point represents the type of enrichment tested (squares: within 2650 genes in MRSA252, triangles: 552 Biocyc gene ontologies, circles: 248 SAMMD expression pathways). A line of 1:1 correspondence is plotted in red.

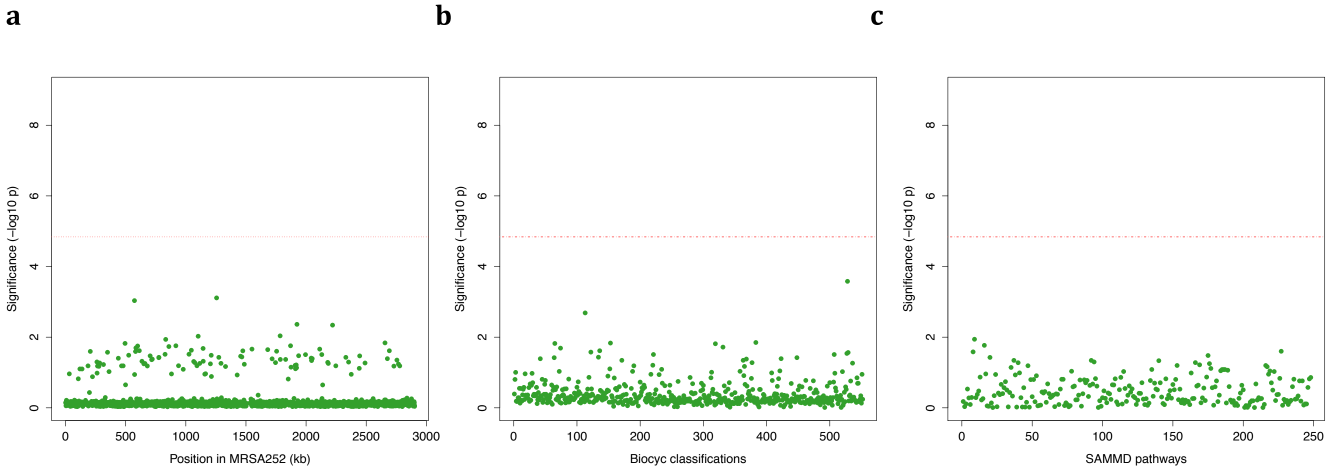

Fig. S3. Genes, ontologies and pathways enriched for protein-altering variants among longitudinally sampled asymptomatic nasal carriers. a) Significance of enrichment of 2650 individual genes. **b)** Significance of enrichment of 552 gene sets defined by BioCyc gene ontologies. **c)** Significance of enrichment of 248 gene sets defined by SAMMD expression pathways. Genes, pathways and ontologies that exceed a Bonferroni-corrected significance threshold of $\alpha = 0.05$ (red lines) are named.

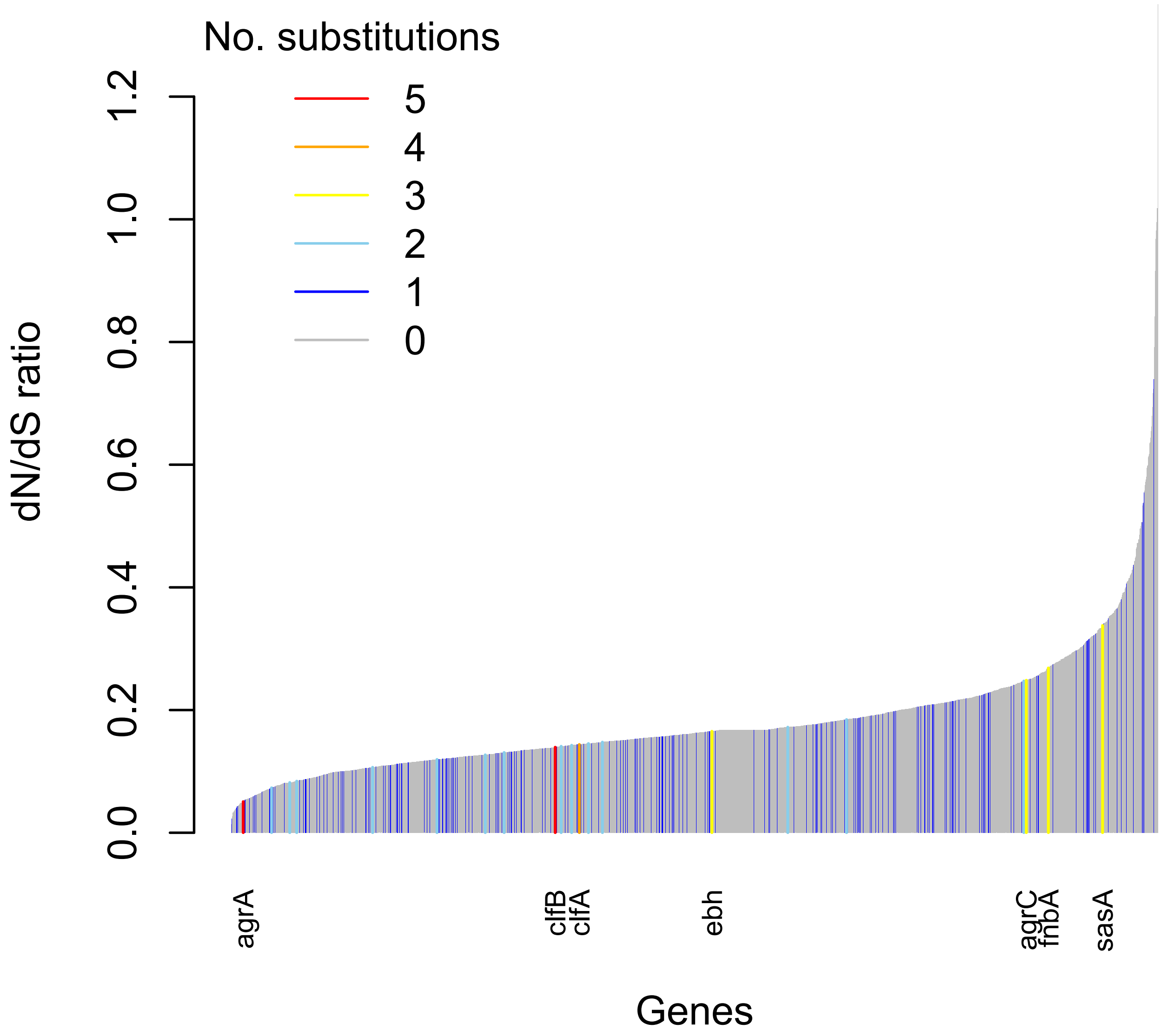

Fig. S4. Genes enriched for substitutions between nose-colonizing and disease-causing bacteria within patients are not the most rapidly evolving at the species level. An estimate of the d_N/d_S ratio between unrelated bacteria is shown for each gene, colour-coded by the number of protein-altering substitutions between nose-colonizing and disease-causing bacteria within patients. There was a negative Spearman rank correlation between d_N/d_S ratio and substitutions within patients (ρ = –0.04, p = 0.02).

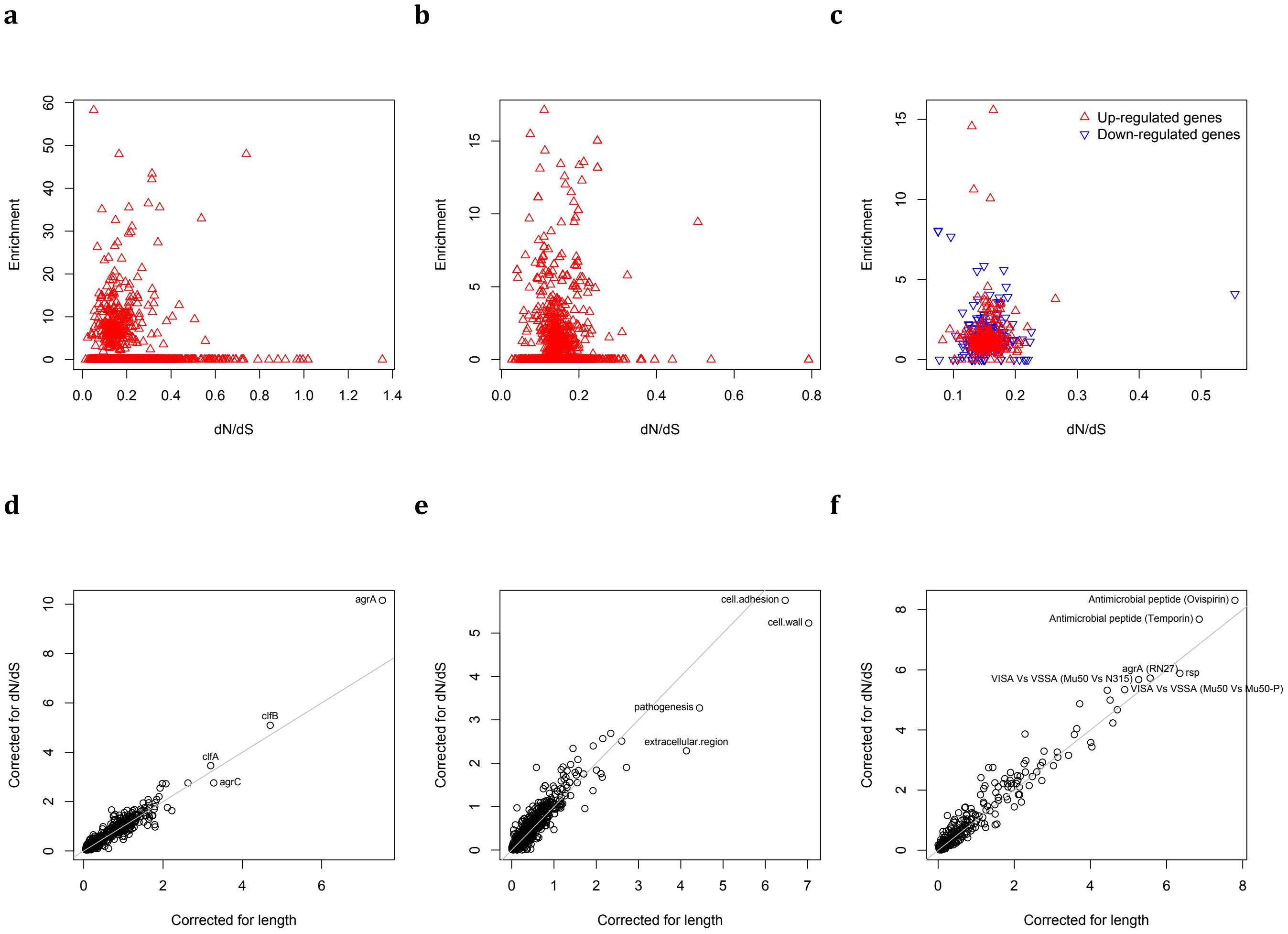

Figure S5. Gene set enrichment analysis is robust to species-level differences in d_N/d_S between genes. For every locus, expression pathway and gene ontology, we estimated d_N/d_S between unrelated *S. aureus*. There was no relationship between d_N/d_S and enrichment of protein-altering substitutions between nose-colonizing and disease-causing bacteria in **a)** loci, **b)** ontologies nor **c)** pathways (non-significant correlations, $p > 0.05$). When we incorporated variability in d_N/d_S between genes in the gene set enrichment analyses, the results were robust for **d)** loci, **e)** ontologies and **f)** pathways, showing only small differences in significance (-$\log_{10}$ p-value) between the analyses that correct for locus length only (horizontal axes) and those that correct for locus length and d_N/d_S (vertical axes).