Title: Severe infections emerge from the microbiome by adaptive evolution

Authors: Bernadette C. Young*[1,2], Chieh-Hsi Wu[1], N. Claire Gordon[1], Kevin Cole[3], James R. Price[3,4], Elian Liu[1,2], Anna E. Sheppard[1,5], Sanuki Perera[1,2], Jane Charlesworth[1], Tanya Golubchik[1], Zamin Iqbal[6], Rory Bowden[6], Ruth C. Massey[7], John Paul[8,9], Derrick W. Crook[1,8,9], Timothy E. A. Peto[1,9], A. Sarah Walker[1,9], Martin J. Llewelyn[3,4], David H. Wyllie[1,10], Daniel J. Wilson*[1,6,11]

Affiliations:

[1]Nuffield Department of Medicine, Experimental Medicine Division, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK.

[2]Microbiology and Infectious Diseases Department, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford OX3 9DU, UK.

[3]Department of Infectious Diseases and Microbiology, Royal Sussex County Hospital, Brighton BN2 5BE, UK.

[4]Department of Global Health and Infection, Brighton and Sussex Medical School, University of Sussex, Falmer BN1 9PS, UK.

[5]NIHR Health Protection Unit in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England, Oxford, UK.

[6]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.

[7]School of Cellular and Molecular Medicine, University of Bristol, UK.

[8]National Infection Service, Public Health England, London, UK.

[9]National Institute for Health Research, Oxford Biomedical Research Centre, Oxford, UK.

[10]Jenner Institute, Centre for Molecular and Cellular Physiology, Oxford OX3 7BN, UK.

[11]Institute for Emerging Infections, Oxford Martin School, University of Oxford, Oxford, OX1 3BD, UK.

*Correspondence to: bernadette.young@ndm.ox.ac.uk, daniel.wilson@ndm.ox.ac.uk.

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 accompany infection is unknown. We investigated *de novo* mutation in 1163 *Staphylococcus aureus* genomes from 105 infected patients with nose-colonization. We report that 72% of infections emerged from the microbiome, with infecting and nose-colonizing bacteria showing parallel adaptive differences. We found 2.8-to-3.6-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. Adaptive mutations in pathogenesis-associated genes were 3.1-fold enriched in infecting but not nose-colonizing bacteria. None of these signatures were observed in healthy carriers nor at the species-level, 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.

One Sentence Summary: Life-threatening *S. aureus* infections emerge from nose microbiome bacteria in association with repeatable adaptive evolution.

Main Text: 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 often 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 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 plays an important role in infection. 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 more generally (*31*). These results raise the question: are unique hallmarks of *de novo* mutation and selection associated with bacterial evolution in severely infected patients?

We addressed this question by investigating the genetic variants arising from within-patient evolution of *S. aureus* sampled from 105 patients with concurrent nose colonization and blood or deep tissue infection. We annotated variants to test for systematic differences between colonizing and infecting bacteria. We discovered several groups of genes showing significant enrichments of protein-altering variants indicating adaptive evolution. For genes implicated in pathogenesis, adaptive mutants were limited to infecting bacteria, while other pathways showed

adaptation in the nose and infection site. Adaptive enrichments were not observed in asymptomatic carriers, nor between unrelated bacteria, indicating evolution in response to disease-associated, within-host selection pressures. Our results reveal that adaptive evolution of genes involved in regulation, toxicity, abscess formation, cell-cell communication and bacterial-host interaction drives parallel differentiation between commensal constituents of the nose microbiome and invasive infections, providing new insights into the evolution 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).

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 previously developed, 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 0-66 variants. Unrelated colonizing and infecting bacteria (10 patients) differed by 1104-50573 variants and typically possessed distinct STs (e.g. Fig. 1a, Fig. S1). 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 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, $\pi = 2.8$ variants) was similar to that previously observed in asymptomatic nasal carriers (*33*) (Reference Panel I, $\pi = 4.1$, $p = 0.13$), but was significantly lower in the infection site ($\pi = 0.6$, $p = 10^{-10.0}$), revealing limited diversification post-infection.

In most patients the infection appeared to be descended from the nose. We used 1149 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

130 bacteria did not. In 16 of those, confidence was high because both mutant and ancestral alleles
131 were observed in the nose, confirming it as the origin of the *de novo* mutation (e.g. Fig. 1d).
132 Conversely, in 10/49 patients, bacteria colonizing the microbiome were likely descended from
133 blood or deep tissue infections (20% of patients with related strains, or 18% of all patients) (e.g.
134 Fig. 1f). Confidence was high for just three of those patients, and they showed unusually high
135 diversity (Supplementary data, P063, P072, P093), suggesting that in persistent infections,
136 infecting bacteria can recolonize the nose.

137 **Protein-truncating mutants are over-represented within infected patients**

138 To help identify variants that could promote, or be promoted by, infection of the blood
139 and deep tissue by bacteria colonizing the nose, we reconstructed within-patient phylogenies and
140 classified variants by their position in the phylogeny. Sequencing multiple colonies per site
141 enabled us to classify variants into those representing genuine differences *between* nose-
142 colonizing and infection populations (*B*-class), variants specific to the nose-*colonizing*
143 microbiome population (*C*-class) and variants specific to the *disease*-causing infection
144 population (*D*-class). We hypothesized that B-class variants would be most enriched for variants
145 promoting, or promoted by, infection, if such variants occur (Fig. 1g).

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

151 In a longitudinal study of one long-term carrier, we previously reported that a burst of
152 protein-truncating variants punctuated the transition from asymptomatic carriage to invasive
153 infection (*30*). Here we found a 3.9-fold over-abundance of protein-truncating variants of all
154 phylogenetic classes in infected patients compared to asymptomatic carriers (Reference Panel I,
155 $p = 0.002$, Table 2), supporting the conclusion that loss-of-function mutations are
156 disproportionately associated with evolution within infected patients. This may reflect a
157 reduction in the efficiency with which selection removes deleterious protein-truncating
158 mutations, and provides evidence of a systematic difference in selection within severely infected
159 patients.

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

162 We hypothesized that variants associated with invasive infection would be enriched
163 among the protein-altering B-class variants between the nose and infection site (Fig. 1g).
164 Therefore we aggregated mutations by genes in a well-annotated reference genome, MRSA252,
165 and tested each gene for an excess of non-synonymous and protein-truncating B-class variants,
166 taking into account the length of the gene. Aggregating by gene was necessary because
167 1318/1322 variants were unique to single patients. The two exceptions involved non-coding
168 variants arising in two patients each, one B-class variant 130 bases upstream of *azlC*, an
169 azaleucine resistance protein (SAR0010), and one D-class variant 88 bases upstream of *eapH*1, a
170 secreted serine protease inhibitor (SAR2295) (*38*).

171 We found a significant excess of five protein-altering B-class variants representing a
172 58.3-fold enrichment in *agrA*, which encodes the response regulator that mediates activation of

173 the quorum sensing system at high cell densities (p=$10^{-7.5}$, Fig. 2a, Table 3). The *clfB* gene
174 encoding clumping factor B, which binds human fibrinogen and loricrin (*39*), showed an excess
175 of five protein-altering B-class variants, representing a 15.9-fold enrichment that was near
176 genome-wide significance after multiple testing correction (p=$10^{-4.7}$).

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

185 We found no significant enrichments of protein-altering variants among D-class variants,
186 but we observed a significant excess of six protein-altering C-class variants in *pbp2* which
187 encodes a penicillin binding protein involved in cell wall synthesis (19.0-fold enrichment, p=10^{-}
188 $^{6.0}$, Fig. S2a). Pbp2 is an important target of β-lactam antibiotics (*40*), revealing adaption –
189 potentially in response to antibiotic treatment – in the nose populations of some patients.

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

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

197 We grouped genes in two different ways: by gene ontology and by expression pathway.
198 First, we obtained a gene ontology for the reference genome from BioCyc (*41*), which classifies
199 genes into biological processes, cellular components and molecular functions. There were 552
200 unique gene ontology groupings of two or more genes. We tested for an enrichment among genes
201 belonging to the ontology, compared to the rest of the genes.

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

208 The most significant enrichment for protein-altering B-class variants between nose and
209 infection sites occurred in the group of genes down-regulated by the cationic antimicrobial
210 peptide (CAMP) ovispirin-1 (p=$10^{-7.8}$), with a similar enrichment in genes down-regulated by
211 temporin L exposure (p=$10^{-6.9}$, Fig. 2c). Like human CAMPs, the animal-derived ovispirin and
212 temporin compounds inhibit epithelial infections by killing phagocytosed bacteria and mediating
213 inflammatory responses (*43*). In response to inhibitory levels of ovispirin and temporin, *agr*,
214 surface-expressed adhesins and secreted toxins are all down-regulated. Collectively, down-
215 regulated genes showed 2.7-fold and 2.8-fold enrichments of adaptive evolution, respectively.

216 Conversely, genes up-regulated in response to CAMPs, including the *vraSR* and *vraDE* cell-wall
217 operons and stress response genes (*43*), exhibited 0.4-fold and 0.7-fold enrichments (i.e.
218 depletions), respectively (Table 3). Thus, genes undergoing adaptive evolution are strongly
219 inhibited by the CAMP-mediated immune response.

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

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

237 Loss-of-function in *agr* mutants represent one alternative route, since they exhibit similar
238 phenotypes to *rsp* mutants, with reduced toxicity and increased surface antigen expression, albeit
239 reduced ability to form abscesses (*32*). We found significant enrichments of genes regulated by
240 *agrA* in two different backgrounds ($p<10^{-4.5}$) and by *sarA* ($p=10^{-4.6}$), underlining the influence of
241 adaptive evolution on both secreted and surface-expressed proteins during infection. We found
242 that expression of genes enriched for protein-altering substitutions was also altered in strains
243 possessing reduced susceptibility to vancomycin, although not in a consistent direction across
244 strains ($p<10^{-4.7}$), and to pine-oil disinfectant ($p=10^{-4.4}$), suggesting such genes may be generally
245 involved in response to harsh environments.

246 Several genes contributed to multiple evolutionary signals, particularly cell-wall
247 anchored proteins involved in adhesion, invasion and immune evasion (*39*), including *fnbA*, *clfA*,
248 *clfB*, *sasA* and *spa*. These multifactorial, partially overlapping signals suggest a large target for
249 selection in adapting to the within-patient environment (Fig. 3). The fact that we observed no
250 comparable significant enrichments in C-class and D-class protein-altering variants (Fig. S2)
251 indicates that these evolutionary patterns are associated specifically with the infection process.

252 **Adaptive evolution in pathogenesis genes is found only in infecting bacteria**

253 Having identified adaptive evolution differentiating nose-colonizing and disease-causing
254 bacteria, we next asked whether the mutant alleles were preferentially found in the nose or
255 infection site. We used 1149 sequences from other patients or carriers (Reference Panel II) to
256 reconstruct the genotype of the MRCA of colonizing and infecting bacteria respectively in each
257 patient. This allowed us to sub-classify B-class variants by whether the mutant allele was found
258 in the nose-colonizing bacteria (B_C-class) or the disease-causing bacteria (B_D-class).

259 *A priori*, we had expected the enrichments of adaptive evolution to be driven primarily by
260 mutants occurring in the disease-causing bacteria (B_D-class). One group of genes showed a
261 signal of such an enrichment among B_D-class variants specifically. Genes belonging to the
262 BioCyc pathogenesis ontology were marginally genome-wide significant in B_D-class variants,
263 showing a 3.1-fold enrichment ($p=10^{-4.6}$) and a statistically insignificant 1.7-fold enrichment in
264 B_C-class variants ($p=0.13$). B_D-class mutants driving this differential signal arose in toxins
265 including gamma hemolysin and several regulatory loci implicated in toxicity and virulence
266 regulation: *rot*, *sarS* and *saeR*.

267 Surprisingly however, we found that all other significantly enriched gene sets were
268 driven by mutant alleles occurring both in colonizing and infecting bacteria (Fig. S3). This
269 indicates there are common selection pressures in the nose and infection site during the process
270 of infection within patients, leading to convergent evolution across body sites. So while
271 adaptation in pathogenesis genes appears specifically invasion-associated, other signals of
272 adaptation in severely infected patients are driven by selection pressures, which might
273 compensate for an altered within-host environment during infection, that are as likely to favor
274 mutants in nose-colonizing bacteria as infecting bacteria.

275 **Signals of adaptation are specific to infected patients and differ from prevailing signatures**
276 **of selection**

277 Two lines of evidence show that the newly discovered signatures of within-host adaptive
278 evolution, both in infecting and nose-colonizing bacteria, are unique to evolution in infected
279 patients. To test this theory against the alternative explanation that our approach merely detects
280 the most rapidly evolving proteins, we searched for similar signals in alternative settings:
281 evolution within asymptomatic carriers, and species-level evolution between unrelated bacteria.

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

288 Nor were the relative rates of non-synonymous to synonymous substitution (d_N/d_S) higher
289 between unrelated *S. aureus* (Reference Panel IV) in the genes that contributed most to the
290 signals associated with adaptation within patients: *agrA*, *agrC clfA*, *clfB*, *fnbA* and *sasA*.
291 Although synonymous diversity was somewhat higher than typical in these genes, the d_N/d_S
292 ratios showed no evidence for excess protein-altering change in these compared to other genes
293 (Fig. S5). Accordingly, incorporating this locus-specific variability of d_N/d_S into the GSEA did
294 not affect the results (Fig. S6). Taken together these lines of evidence show that the ontologies,
295 pathways and genes significantly differentiated between colonizing and infecting bacteria arise
296 in response to selection pressures specifically associated with infected patients, and are not
297 repeated in asymptomatic carriers or species-level evolution.

298 **Discussion**

299 We have discovered that common, life-threatening infections of *S. aureus* are frequently
300 descended from bacteria colonizing the human microbiome. These infections are associated with
301 repeatable patterns of bacterial evolution driven by within-patient mutation and selection. Genes

involved in pathogenesis, notably toxins and regulators, showed evidence for adaptation in infecting but not nose-colonizing bacteria. Surprisingly, other signatures of adaptation occurred in parallel in nose-colonizing and infecting bacteria, affecting 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 targeted by selection in infected patients.

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 favoring colonization and infection distinct from those favoring 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 favored by short-term selection but ultimately incidental or damaging to the bacterial reproductive life cycle.

Nor did we see these signatures of bacterial adaptation and excess loss-of-function mutations in healthy nose carriers, indicating that risk factors for invasive infections, such as a weakened or over-activated immunological response, comorbidities or medical interventions, may create distinctive selection pressures in infected patients. As in cancer, the effects of such risk factors may be mediated, at least in part, through the selection pressure they exert on the microbiome.

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.

References and Notes:

1. GBD 2015 Mortality and Causes of Death Collaborators. 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 (2016).
2. PJ Turnbaugh et al. The human microbiome project. *Nature.* ***449(7164)***, 804-810 (2007).
3. A Casadevall, FC Fang, LA Pirofski. Microbial virulence as an emergent property: Consequences and opportunities. *PLoS Pathog.* ***7(7)***, e1002136 (2011).
4. PO Methot, S Alizon. What is a pathogen? Toward a process view of host-parasite interactions. *Virulence.* ***5(8)***, 775-85 (2014).
5. SP Brown, DM Cornforth, N Mideo. Evolution of virulence in opportunistic pathogens: Generalism, plasticity, and control. *Trends Microbiol.* ***20(7)***, 336-42 (2012).

343 6. SK Sheppard et al. Genome-wide association study identifies vitamin B5 biosynthesis as a
344 host specificity factor in Campylobacter. *Proc Natl Acad Sci U S A.* ***110(29)***, 11923-7 (2013).

345 7. M Laabei et al. Predicting the virulence of MRSA from its genome sequence. *Genome Res*
346 ***24(5)***, 839-49 (2014).

347 8. C Chewapreecha et al. Comprehensive identification of single nucleotide polymorphisms
348 associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet.*
349 ***10(8),*** e1004547 (2014).

350 9. PE Chen, BJ Shapiro. The advent of genome-wide association studies for bacteria. *Curr Opin*
351 *Microbiol.* ***25***, 17-24 (2015).

352 10. SG Earle et al. Identifying lineage effects when controlling for population structure improves
353 power in bacterial association studies. *Nat Microbiol.* ***1***, 16041 (2016).

354 11. X Didelot et al. Within-host evolution of bacterial pathogens. *Nat Rev Microbiol.* ***14(3)***, 150-
355 62 (2016).

356 12. C von Eiff et al. Nasal carriage as a source of *Staphylococcus aureus* bacteremia. *N Engl J*
357 *Med*. ***344(1)***, 11-16 (2001).

358 13. BP Howden et al. Evolution of multidrug resistance during *Staphylococcus aureus* infection
359 involves mutation of the essential two component regulator WalKR. *PLoS Pathog.* ***7(11)***,
360 e1002359 (2011).

361 14. V Eldholm et al. Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a
362 susceptible ancestor in a single patient. *Genome Biol.* ***15(11)***, 490 (2014).

363 15. TD Lieberman et al. Parallel bacterial evolution within multiple patients identifies candidate
364 pathogenicity genes. *Nat Genet.* ***43(12)***, 1275-80 (2011).

365 16. RL Marvig, HK Johansen, S Molin, L Jelsbak. Genome analysis of a transmissible lineage of
366 *Pseudomonas aeruginosa* reveals pathoadaptive mutations and distinct evolutionary paths of
367 hypermutators. *PLoS Genet*. ***9(9)***, e1003741 (2013).

368 17. T Markussen et al. Environmental heterogeneity drives within-host diversification and
369 evolution of *Pseudomonas aeruginosa. Mbio*. ***5(5)***, e01592-14 (2014).

370 18. TD Lieberman et al. Genetic variation of a bacterial pathogen within individuals with cystic
371 fibrosis provides a record of selective pressures. *Nat Genet*. ***46(1)***, 82-7 (2014).

372 19. RL Marvig, LM Sommer, S Molin, HK Johansen. Convergent evolution and adaptation of
373 *Pseudomonas aeruginosa* within patients with cystic fibrosis. *Nat Genet.* ***47(1)***, 57-64
374 (2015).

375 20. AH Moeller et al. Cospeciation of gut microbiota with hominids. *Science*. ***353(6297)***, 380-2
376 (2016).

377 21. JA Lees et al. Large scale genomic analsyis shows no evidence for pathogen adatption
378 between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microb Genom*.
379 ***3(1)***, e000103 (2017).

380 22. KE Traber et al. Agr function in clinical *Staphylococcus aureus* isolates. *Microbiology.*
381 ***154(8)***, 2265-74 (2008).

382 23. RP Novick and E Geisinger. Quorum sensing in staphylococci. *Annu Rev Genet*. ***42***, 541-64
383 (2008).

384 24. KL Painter, A Krishna, S Wigneshweraraj, AM Edwards. What role does the quorum-sensing
385 accessory gene regulator system play during *Staphylococcus aureus* bacteremia? *Trends*
386 *Microbiol.* ***22(12)***, 676-85 (2014).

387 25. G Sakoulas, PA Moise, MJ Rybak. Accessory gene regulator dysfunction: An advantage for
388 *Staphylococcus aureus* in health-care settings? *J Infect Dis.* ***199(10)***, 1558-9 (2009).

389 26. DA Robinson, et al. Evolutionary genetics of the accessory gene regulator (agr) locus in
390 *Staphylococcus aureus. J Bacteriol.* ***187(24)***, 8312-21 (2005).

391 27. EJ Pollitt, et al. Cooperation, quorum sensing, and evolution of virulence in S*taphylococcus*
392 *aureus. Infect Immun.* ***82(3)***, 1045-51. (2014).

393 28. B Shopsin et al. Mutations in agr do not persist in natural populations of methicillin- resistant
394 *Staphylococcus aureus*. *J Infect Dis*. ***202(10)***, 1593 (2010).

395 29. W Gao et al. Two novel point mutations in clinical Staphylococcus aureus reduce linezolid
396 susceptibility and switch on the stringent response to promote persistent infection. *PLoS*
397 *Pathog*. ***6(6)***, e1000944 (2010).

398 30. BC Young et al. Evolutionary dynamics of *Staphylococcus aureus* during progression from
399 carriage to disease. *Proc Natl Acad Sci U S A*. ***109(12)***, 4550-5 (2012).

400 31. M Laabei et al. Evolutionary trade-offs underlie the multi-faceted virulence of
401 *Staphylococcus aureus. PLoS Biol*. ***13(9)***, e1002229 (2015).

402 32. S Das et al. Natural mutations in a *Staphylococcus aureus* virulence regulator attenuate
403 cytotoxicity but permit bacteremia and abscess formation. *Proc Natl Acad Sci U S A*.
404 ***113(22),*** E3101-10 (2016).

405 33. T Golubchik et al. Within- host evolution of *Staphylococcus aureus* during asymptomatic
406 carriage. *PloS One* ***8(5)***, e61319 (2013).

407 34. Z Iqbal et al. De novo assembly and genotyping of variants using colored de bruijn graphs.
408 *Nat Genet.* ***44(2)***, 226-32 (2012).

409 35. ER Moxon, PA Murphy. *Haemophilus influenzae* bacteremia and meningitis resulting from
410 survival of a single organism. *Proc Natl Acad Sci U S A.* ***75(3)***, 1534-6 (1978).

411 36. E Margolis, BR Levin. Within-host evolution for the invasiveness of commensal bacteria: An
412 experimental study of bacteremias resulting from *Haemophilus influenzae* nasal carriage. *J*
413 *Infect Dis*. ***196(7)***, 1068-1075 (2007).

414 37. DM Rand, LM Kann. Excess amino acid polymorphism in mitochondrial DNA: contrasts
415 among genes from Drosophila, mice, and humans. *Mol Biol Evol.* ***13(6)***, 735-48 (1996).

416 38. DA Stapels et al. *Staphylococcus aureus* secretes a unique class of neutrophil serine protease
417 inhibitors. *Proc Natl Acad Sci U S A*. ***111(36)***, 13187-92 (2014).

418 39. TJ Foster, JA Geoghegan, VK Ganesh, M Höök. Adhesion, invasion and evasion: The many
419 functions of the surface proteins of *Staphylococcus aureus*. *Nat Rev Microbiol*. ***12(1)***, 49-62
420 (2013).

421 40. TA Leski, A Tomasz. Role of penicillin-binding protein 2 (PBP2) in the antibiotic
422 susceptibility and cell wall cross-linking of *Staphylococcus aureus*: Evidence for the
423 cooperative functioning of PBP2, PBP4, and PBP2A. *J Bacteriol*. ***187(5),*** 1815-1824 (2005).

424 41. R Caspi et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc
425 collection of pathway/genome databases. *Nucleic Acids Res*. ***44(D1)***, D471-80 (2016).

426 42. V Nagarajan, M Elasri. SAMMD: *Staphylococcus aureus* microarray meta- database. *BMC*
427 *Genomics.* ***8(1),*** 351 (2007).

428 43. M Pietiainen et al. Transcriptome analysis of the responses of *Staphylococcus aureus* to
429 antimicrobial peptides and characterization of the roles of vraDE and vraSR in antimicrobial
430 resistance. *BMC Genomics*. ***10***, 429 (2009).

431 44. MG Lei et al. Rsp inhibits attachment and biofilm formation by repressing fnbA in
432 *Staphylococcus aureus* MW2. *J Bacteriol.* ***193(19)***, 5231 (2011).

433 45. PM Dunman et al. Transcription profiling-based identification of *Staphylococcus aureus*
434 genes regulated by the agr and/or sarA loci. *J Bacteriol*. ***183(24)***, 7341-53 (2001).

435 46. L Cui et al. DNA microarray-based identification of genes associated with glycopeptide
436 resistance in *Staphylococcus aureus*. *Antimicrob Agents Chemother*. ***49(8)***, 3404-13 (2005).

437 47. BP Howden et al. Different bacterial gene expression patterns and attenuated host immune
438 responses are associated with the evolution of low-level vancomycin resistance during
439 persistent methicillin-resistant *Staphylococcus aureus* bacteraemia. *BMC Microbiol.* ***8,*** 39
440 (2008).

441 48. J Cassat et al. Transcriptional profiling of a *Staphylococcus aureus* clinical isolate and its
442 isogenic agr and sarA mutants reveals global differences in comparison to the laboratory
443 strain RN6390. *Microbiology* ***152,*** 3075-90 (2006).

444 49. R Lamichhane-Khadka et al. Genetic changes that correlate with the pine-oil disinfectant-
445 reduced susceptibility mechanism of *Staphylococcus aureus*. *J. Appl. Microbiol.* ***105(6),***
446 1973-81 (2008).

447 50. RG Everitt et al. Mobile elements drive recombination hotspots in the core genome of
448 *Staphylococcus aureus. Nat Commun*. ***5***, 3956 (2014).

449 51. NC Gordon et al. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-
450 genome sequencing. *J Clin Microbiol*. ***52(4)***, 1182-91 (2014).

451 52. NC Gordon et al. Whole genome sequencing reveals the contribution of long-term carriers in
452 *Staphylococcus aureus* outbreak investigation. (2017)(Under review, submitted as
453 accompanying manuscript)

454 53. MTG Holden et al. Complete genomes of two clinical *Staphylococcus aureus* strains:
455 Evidence for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci U S A.*
456 ***101(26)***, 9786-9791 (2004).

457 54. SR Gill et al. Insights on evolution of virulence and resistance from the complete genome
458 analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-
459 producing methicillin-resistant *Staphylococcus epidermidis* strain. *J Bacteriol*. ***187(7)***, 2426-
460 2438 (2005).

461 55. AF Gillaspy, et al. "The *Staphylococcus aureus* NCTC8325 genome" in *Gram positive*
462 *pathogens*. V Fischetti, R Novick, J Ferretti, et al, Eds. (ASM Press, 2006), chap. 32.

463 56. M Kuroda et al. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus.*
464 *Lancet*. ***357(9264)***, 1225-40 (2001).

465 57. BA Diep et al. Complete genome sequence of USA300, an epidemic clone of community-
466 acquired meticillin-resistant *Staphylococcus aureus. Lancet*. ***367(9512)***, 731-9 (2006).

467 58. T Baba et al. Genome sequence of *Staphylococcus aureus* strain newman and comparative
468 analysis of staphylococcal genomes: Polymorphism and evolution of two major
469 pathogenicity islands. *J Bacteriol*. ***190(1)***, 300-10 (2008).

470 59. MT Holden et al. Genome sequence of a recently emerged, highly transmissible, multi-
471 antibiotic- and antiseptic-resistant variant of methicillin-resistant *Staphylococcus aureus*,
472 sequence type 239 (TW). *J Bacteriol.* ***192(3)***, 888-92 (2010).

473 60. MJ Schijffelen, CH Boel, JA van Strijp, AC Fluit. Whole genome analysis of a livestock-
474 associated methicillin-resistant *Staphylococcus aureus* ST398 isolate from a case of human
475 endocarditis. *BMC Genomics*. ***11***, 376 (2010).

476 61. K Chua et al. Complete genome sequence of *Staphylococcus aureus* strain JKD6159, a
477 unique Australian clone of ST93-IV community methicillin-resistant *Staphylococcus aureus*.
478 *J Bacteriol.* ***192(20)***, 5556-7 (2010).

479 62. L Herron-Olson, JR Fitzgerald, JM Musser, V Kapur. Molecular correlates of host
480 specialization in *Staphylococcus aureus*. *PLoS One*. ***2(10)***, e1120 (2007).

481 63. CM Guinane et al. Evolutionary genomics of *Staphylococcus aureus* reveals insights into the
482 origin and molecular basis of ruminant host adaptation. *Genome Biol Evol*. ***2***, 454-66 (2010).

483 64. BV Lowder et al. Recent human-to-poultry host jump, adaptation, and pandemic spread of
484 *Staphylococcus aureus*. *Proc Natl Acad Sci U S A.* ***106(46)***, 19545-50 (2009).

485 65. MT Holden et al. A genomic portrait of the emergence, evolution, and global spread of a
486 methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Res* ***23(4)***, 653-64 (2013).

487 66. DR Zerbino, E Birney. Velvet: Algorithms for de novo short read assembly using de bruijn
488 graphs. *Genome Res*. ***18(5)***, 821-9 (2008).

489 67. G Lunter, M Goodson. Stampy: A statistical algorithm for sensitive and fast mapping of
490 illumina sequence reads. *Genome Res*. ***21(6)***, 936-9 (2011).

491 68. SF Altschul et al. Basic local alignment search tool. *J Mol Biol*. ***215(3)***, 403-10 (1990).

492 69. MC Enright et al. Multilocus sequence typing for characterization of methicillin-resistant and
493 methicillin-susceptible clones of *Staphylococcus aureus. J Clin Microbiol.* ***38(3)***, 1008-15
494 (2000).

495 70. X Didelot et al. Microevolutionary analysis of *Clostridium difficile* genomes to investigate
496 transmission. *Genome Biol*. ***13(12)***, R118 (2012).

497 71. D Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*. ***21***, 19-28 (1991).

498 72. AC Darling, B Mau, FR Blattner, NT Perna. Mauve: Multiple alignment of conserved
499 genomic sequence with rearrangements. *Genome Res.* ***14(7)***, 1394-403 (2004).

500 73. FJ Logan-Klumpler, et al. GeneDB--an annotation database for pathogens. *Nucleic Acids Res*
501 ***40***, D98-108 (2012).

502 74. M Kanehisa, Y Sato, M Kawashima, M Furumichi, M Tanabe. KEGG as a reference
503 resource for gene and protein annotation. *Nucleic Acids Res* ***44(D1)***, D457-62 (2016).

504 75. N Saitou, M Nei. The neighbor-joining method: A new method for reconstructing
505 phylogenetic trees. *Mol Biol Evol*. ***4(4)***, 406-25 (1987).

506 76. A Stamatakis. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
507 phylogenies. *Bioinformatics.* ***30(9)***, 1312-3 (2014).

508 77. X Didelot, DJ Wilson. ClonalFrameML: Efficient inference of recombination in whole
509 bacterial genomes. *PLoS Comput Biol*. ***11(2)***, e1004041 (2015).

510 78. T Pupko, I Pe'er, R Shamir, D Graur. A fast algorithm for joint reconstruction of ancestral
511 amino acid sequences. *Mol Biol Evol*. ***17(6)***, 890-6 (2000).

512 79. R Core Team. R: A Language and Environment for Statistical Computing, Vienna, Austria:
513 R Foundation for Statistical Computing. URL https://www.R-project.org/ (2015).

514 80. K Roeder, L Wasserman. Genome-Wide Significance Levels and Weighted Hypothesis
515 Testing. *Stat Sci*. ***24(4)***, 398-413 (2009).

516 81. A Loytynoja, AJ Vilella, N Goldman. Accurate extension of multiple sequence alignments
517 using a phylogeny-aware graph algorithm. *Bioinformatics.* ***28(13)***, 1684-91(2012).

518 82. DJ Wilson, G McVean. Estimating diversifying selection and functional constraint in the
519 presence of recombination. *Genetics*. ***172(3)***, 1411-25 (2006).

520 83. DJ Wilson, RD Hernandez, P Andolfatto, M Przeworski. A population genetics-
521 phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet*.
522 ***7(12)***, e1002395 (2011).

523 **Acknowledgments:** We would like to thank Ed Feil, Stephen Leslie, Gil McVean and Richard
524 Moxon for helpful insights and useful discussions. Sequencing reads uploaded to short
525 read archive (SRA) under BioProject PRJNA369475. RNA-Seq data relating to isolate
526 from P005 (aka 'patient S') previously submitted under BioProject PRJNA279958.

527 The views expressed in this publication are those of the authors and not necessarily those of the
528 funders. This study was supported by the Oxford NIHR Biomedical Research Centre, a
529 Mérieux Research Grant, the National Institute for Health Research Health Protection
530 Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial
531 Resistance at Oxford University in partnership with Public Health England (PHE) (grant
532 HPRU-2012-10041), and the Health Innovation Challenge Fund (a parallel funding
533 partnership between the Wellcome Trust (grant WT098615/Z/12/Z) and the Department
534 of Health (grant HICF-T5-358)). T.E.P. and D.W.C. are NIHR Senior Investigators.
535 D.J.W. and Z.I. are Sir Henry Dale Fellows, jointly funded by the Wellcome Trust and
536 the Royal Society (Grants 101237/Z/13/Z and 102541/Z/13/Z). B.C.Y is a Research
537 Training Fellow funded by the Wellcome Trust (Grant 101611/Z/13/Z). We acknowledge
538 the support of Wellcome Trust Centre for Human Genetics core funding (Grant
539 090532/Z/09/Z).

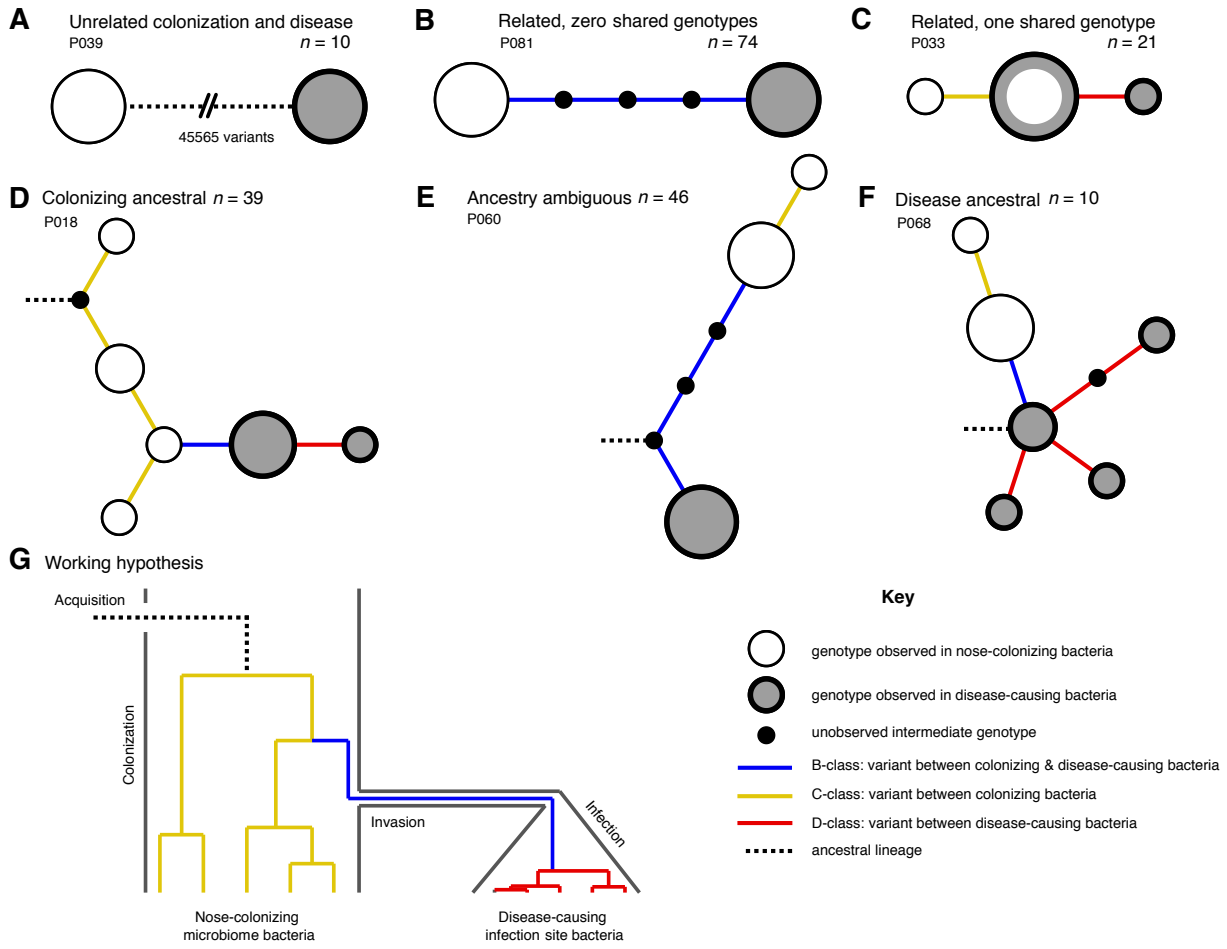
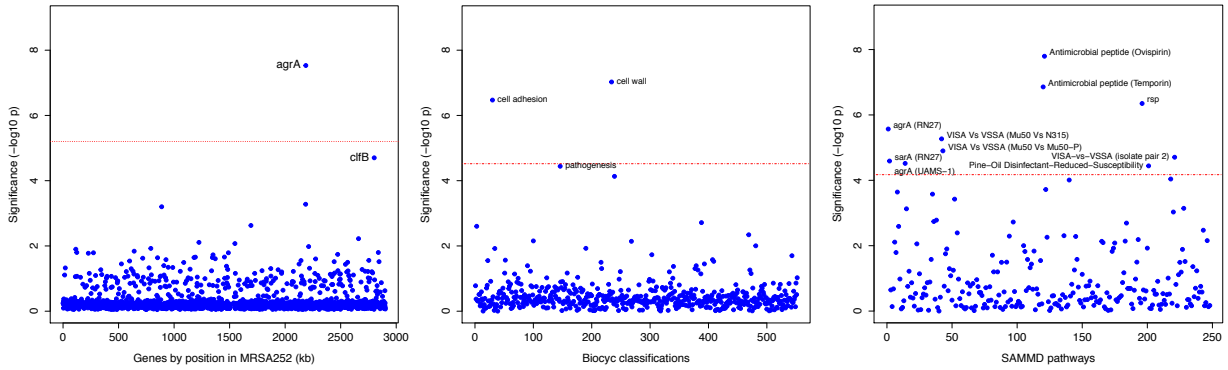


Fig. 1. Disease-causing *S. aureus* form closely related but distinct populations descended from microbiome-colonizing bacteria in the majority of infections. Bacteria sampled from the nose and infection site of 105 patients formed one of three population structures, illustrated with example haplotrees: (**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, or are promoted by, 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 color 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.

558 **A** **B** **C**



559

560 **Fig. 2. Genes, ontologies and pathways enriched for protein-altering substitutions between**
561 **nose-colonizing and disease-causing bacteria within infected patients.** (**A**) Significance of
562 enrichment of 2650 individual genes. (**B**) Significance of enrichment of 552 gene sets defined by
563 BioCyc gene ontologies. (**C**) Significance of enrichment of 248 gene sets defined by SAMMD
564 expression pathways. Genes, pathways and ontologies that approach or exceed a Bonferroni-
565 corrected significance threshold of $\alpha = 0.05$, weighted for the number of tests per category, (red
566 lines) are named.

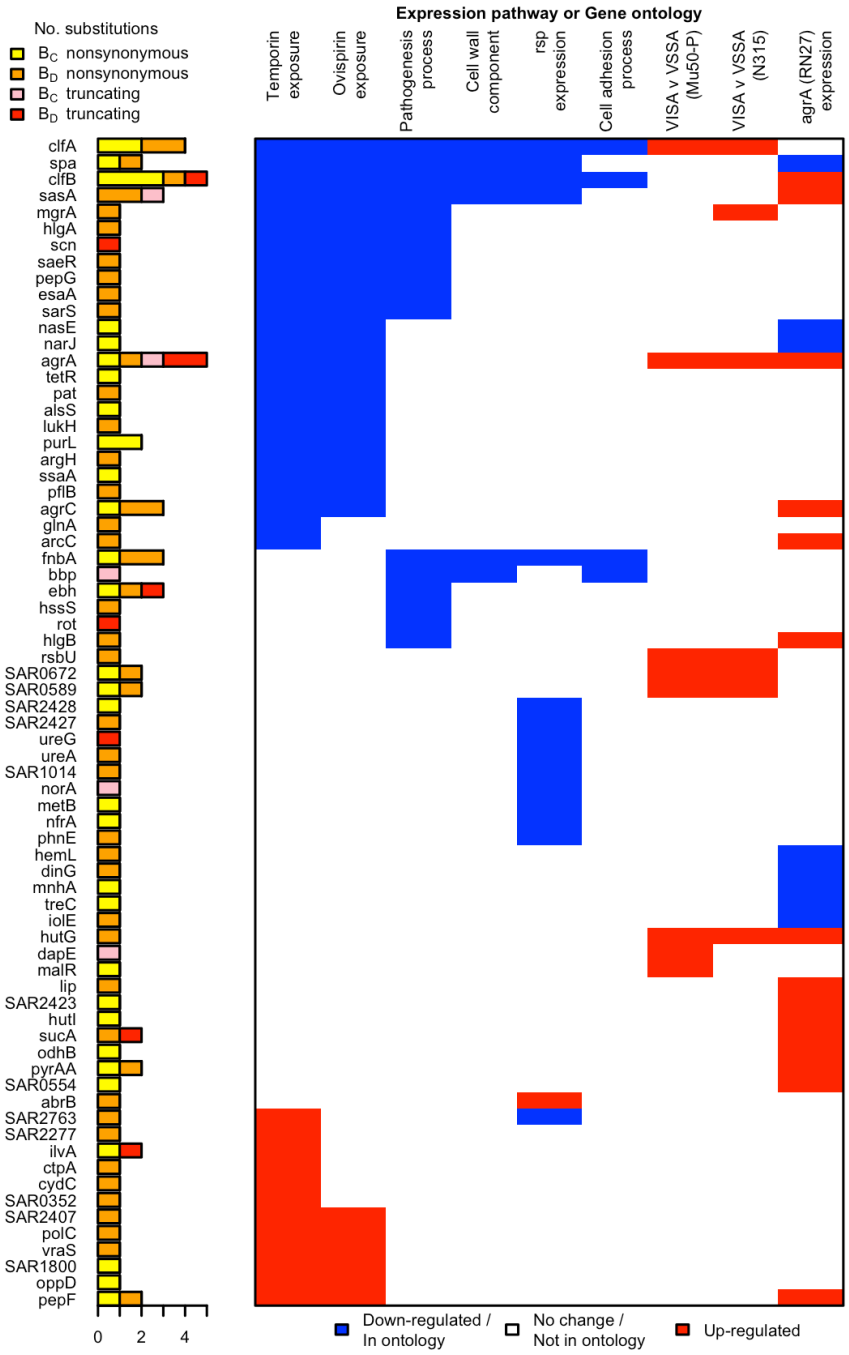


Fig. 3. All genes contributing to the pathways and ontologies most significantly enriched for protein-altering substitutions between nose-colonizing and disease-causing bacteria. The pathogenesis ontology, in which significant enrichments were observed in disease-causing but not nose-colonizing bacteria, is shown for comparison. Every gene with at least one substitution between nose-colonizing and disease-causing bacteria and which was up- (red) or down-regulated (blue) in one of the pathways or a member of one of the ontologies (blue) 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).

Infection sites	Relation of colonizing to infecting bacteria		
	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

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

Phylogenetic position	Number of variants (Neutrality index)				
	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	212	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

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

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 (*41*)							
Cell wall	18		30.9		5.02		**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 (*43*)	40	7	121.2	142.9	2.65	0.39	**7.80**
Temporin L (*43*)	42	14	125.1	156.1	2.78	0.74	**6.86**
rsp (*44*)	27	1	61.1	13.7	3.61	0.60	**6.35**
agrA (RN27) (*45*)	9	30	41.0	85.0	1.83	2.94	**5.57**
VISA-vs-VSSA (Mu50 vs N315) (*46*)	0	17	0	34.4	0	3.95	**5.27**
VISA-vs-VSSA (Mu50 vs Mu50-P) (*46*)	0	17	0	36.7	0	3.70	**4.90**
VISA-vs-VSSA (isolate pair 2) (*47*)	14	3	26.9	59.7	4.06	0.39	**4.71**
sarA (RN27) (*45*)	6	23	49.9	57.7	0.97	3.22	**4.59**
agrA (UAMS-1 OD 1.0) (*48*)	0	5	0	2.7	0	14.57	**4.52**
Pine-Oil Disinfectant-Reduced-Susceptibility (*49*)	17	5	36.4	23.6	3.76	1.70	**4.44**
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 5.2, 4.5 or 4.2 was considered genome-wide significant for loci, gene ontologies or expression pathways respectively (in bold).

Supplementary Materials:

Materials and Methods

Figures S1-S6

Tables S1-S4

References (*50-83*)

Supplementary Materials:

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 55 blood cultures and 50 pus, soft tissue, bone or joint samples. 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 a carriage study of *S. aureus* in Oxfordshire (*50*) (BioProject PRJEB2881). II. A compilation of 95 unrelated samples from the same Oxfordshire carriage study (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 (*51*) (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 (*33,52*), (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) (*53-65*). 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 (*66*) to assemble reads into contigs *de novo*, and Stampy (*67*) to map reads against two reference genomes: MRSA252 (*53*) and a patient-specific reference comprising the contigs assembled for one colony sampled from each patient's nose. Repetitive regions, defined by BLASTing (*68*) the reference genome against itself, were masked prior to variant calling. To obtain multilocus sequence types (*69*) 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 (*30,33,70*) 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 (*71*). 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 (*72*). 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 (*73*) and KEGG (*74*).

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 bacteremia isolates) using a two-step neighbor-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 (*75*), before resolving the relationships within each cluster by building a maximum likelihood tree using RAxML (*76*), 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 (*77*). We used these subtrees to identify, for each patient, the most closely related 'nearest neighbor' sampled from another patient or carrier. We employed this nearest neighbor 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 (*78*) in ClonalFrameML (*77*). 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 neighbor, and identified the ancestral allele as the one possessed by the nearest neighbor. 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 (*79*), 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*)

133 and expression pathways from the SAMMD database of transcriptional studies (*42*). The BioCyc
134 database comprises ontologies describing biological processes, cellular components and
135 molecular functions. The SAMMD database groups genes up-regulated, down-regulated or not
136 differentially regulated in response to experimentally manipulated growth conditions or isogenic
137 mutations, usually of a regulatory gene. After excluding ontologies or pathways with two groups,
138 one involving a single gene, and combining ontologies or pathways with identical groupings of
139 genes, we conducted 800 GSEAs in addition to the 2650 ontologies comprised of individual loci.
140 The number of groupings of genes was always two for BioCyc (included/excluded from the
141 ontology) and two or three for SAMMD (up-/down-/un-differentially regulated in the
142 experiment). Again we employed a Poisson regression in which we modelled the expected
143 numbers of variants in any gene j as $\lambda_0 L_j$ under the null hypothesis of no enrichment where λ_0
144 gives the expected number of variants per kilobase and L_j is the length of gene j in kilobases. We
145 compared this to the alternative hypothesis in which the expected number of variants was $\lambda_1 L_j$,
146 $\lambda_2 L_j$ or $\lambda_3 L_j$ for gene j depending on the grouping in the ontology/pathway. Using R, we
147 estimated the parameters λ_0, λ_1, λ_2 and λ_3 from the data by maximum likelihood and tested for
148 significance via a likelihood ratio test with one or two degrees of freedom, depending on the
149 number of groupings in the ontology/pathway.

150 ***GSEA multiple testing correction***. To account for the multiplicity of testing, we adjusted
151 the *p-value* significance thresholds from a nominal $\alpha = 0.05$ using the weighted Bonferroni
152 method. We weighted the significance thresholds by the relative number of tests in each
153 category: 2650 genes, 552 BioCyc ontologies and 248 SAMMD expression pathways. This
154 avoids overly stringent multiple testing correction in categories with fewer tests (*80*), e.g. the 248
155 SAMMD expression pathways, owing to other categories with very large numbers of tests, e.g.
156 the 2650 genes. This gave adjusted significance thresholds of $10^{-5.2}$ for genes, $10^{-4.5}$ for BioCyc
157 ontologies and $10^{-4.2}$ for SAMMD expression pathways.

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

168 ***omegaMap analysis.*** We estimated d_N/d_S ratios between unrelated *S. aureus* to
169 characterize the prevailing patterns of selection at the species level. We used Mauve (*72*) to
170 pairwise align 15 reference genomes against MRSA252, i.e. Reference Panel IV. This allowed us
171 to distinguish orthologs from paralogs in the next step in which we multiply aligned all coding
172 sequences overlapping those in MRSA252 using PAGAN (*81*). After removing sequences with
173 premature stop codons, we analysed each alignment of between two and 16 genes using a
174 modification of omegaMap (*82*), assuming all sites were unlinked. We previously showed this
175 assumption, which confers substantial computational efficiency savings, does not adversely
176 affect estimates of selection coefficients (*83*). We estimated variation in d_N/d_S within genes using
177 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. Sequencing reads uploaded to short read archive (SRA) under BioProject PRJNA369475. RNA-Seq data relating to isolate from P005 (aka 'patient S') previously submitted under BioProject PRJNA279958.

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.

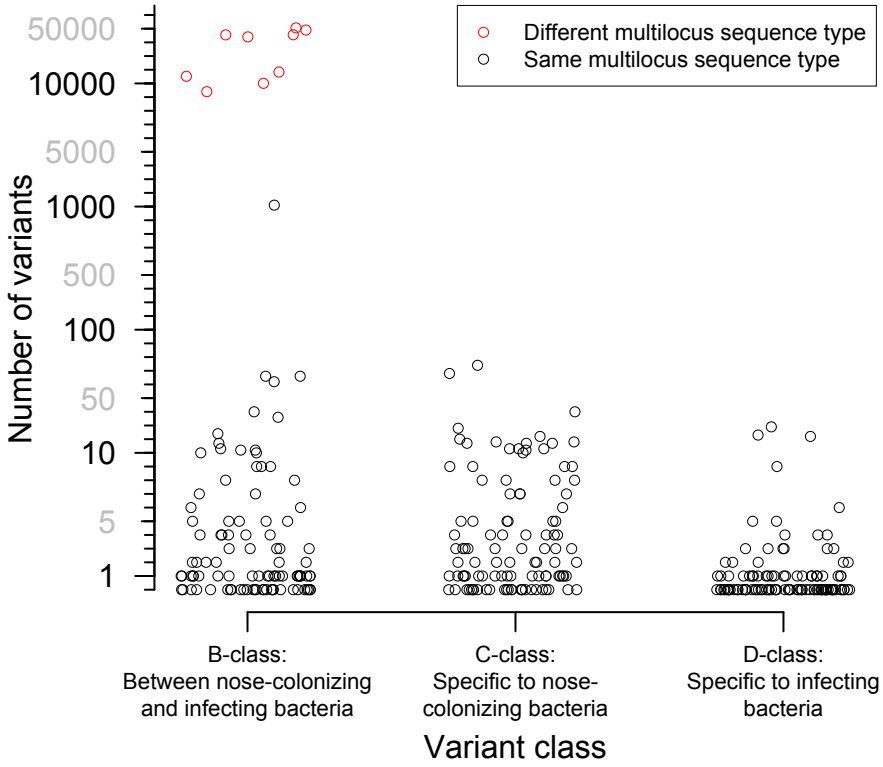
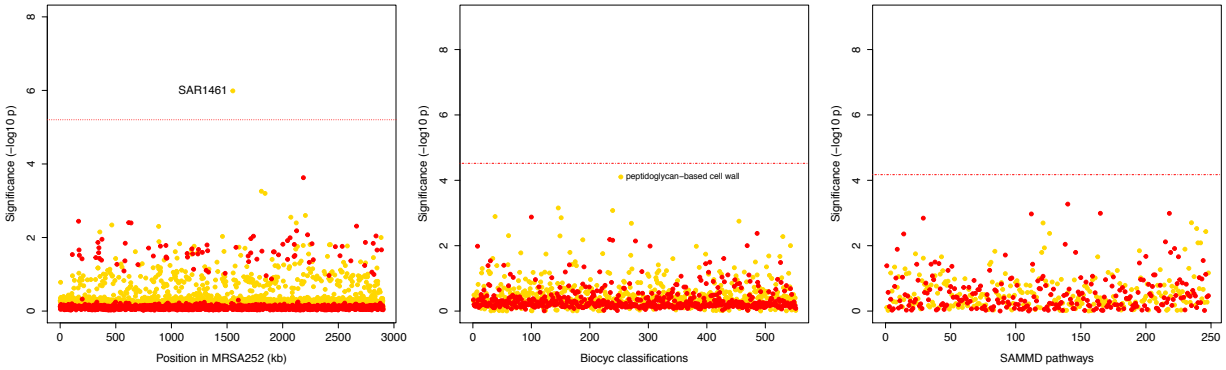


Fig. S1. Distribution of the number of variants identified within 105 severely infected patients, by class. Three classes of variants were identified: those representing genuine differences *between* nose-colonizing and infection populations (*B*-class), variants specific to the nose-*colonizing* microbiome population (*C*-class) and variants specific to the *disease*-causing infection population (*D*-class). The number of variants is shown on a piecewise-linear axis, with horizontal positioning permuted to assist visualization. Where nose-colonizing and infecting bacteria possessed different multilocus sequence types, the number of variants between those populations is colored red. When the number of B-class variants was 66 or less, nose-colonizing and infecting bacteria were considered related, since a similar range of (C-class) diversity was observed within the nose-colonizing populations of bacteria with the same multilocus sequence type. When the number of B-class variants was 1104 or more, nose-colonizing and infecting bacteria were considered unrelated.

212 **A** **B** **C**



213

214 **Fig. S2. Genes, ontologies and pathways enriched for protein-altering transient variants**
215 **within nose-colonizing and disease-causing bacteria.** (**A**) Significance of enrichment of 2650
216 individual genes. SAR1461 encodes Pbp2, penicillin-binding protein 2. (**B**) Significance of
217 enrichment of 552 gene sets defined by BioCyc gene ontologies. (**C**) Significance of enrichment
218 of 248 gene sets defined by SAMMD expression pathways. C-class variants among nose-
219 colonizing bacteria are colored gold, D-class variants among disease-causing bacteria are colored
220 red. Genes, pathways and ontologies that approach or exceed a Bonferroni-corrected significance
221 threshold of $\alpha = 0.05$, weighted for the number of tests per category, (red lines) are named.

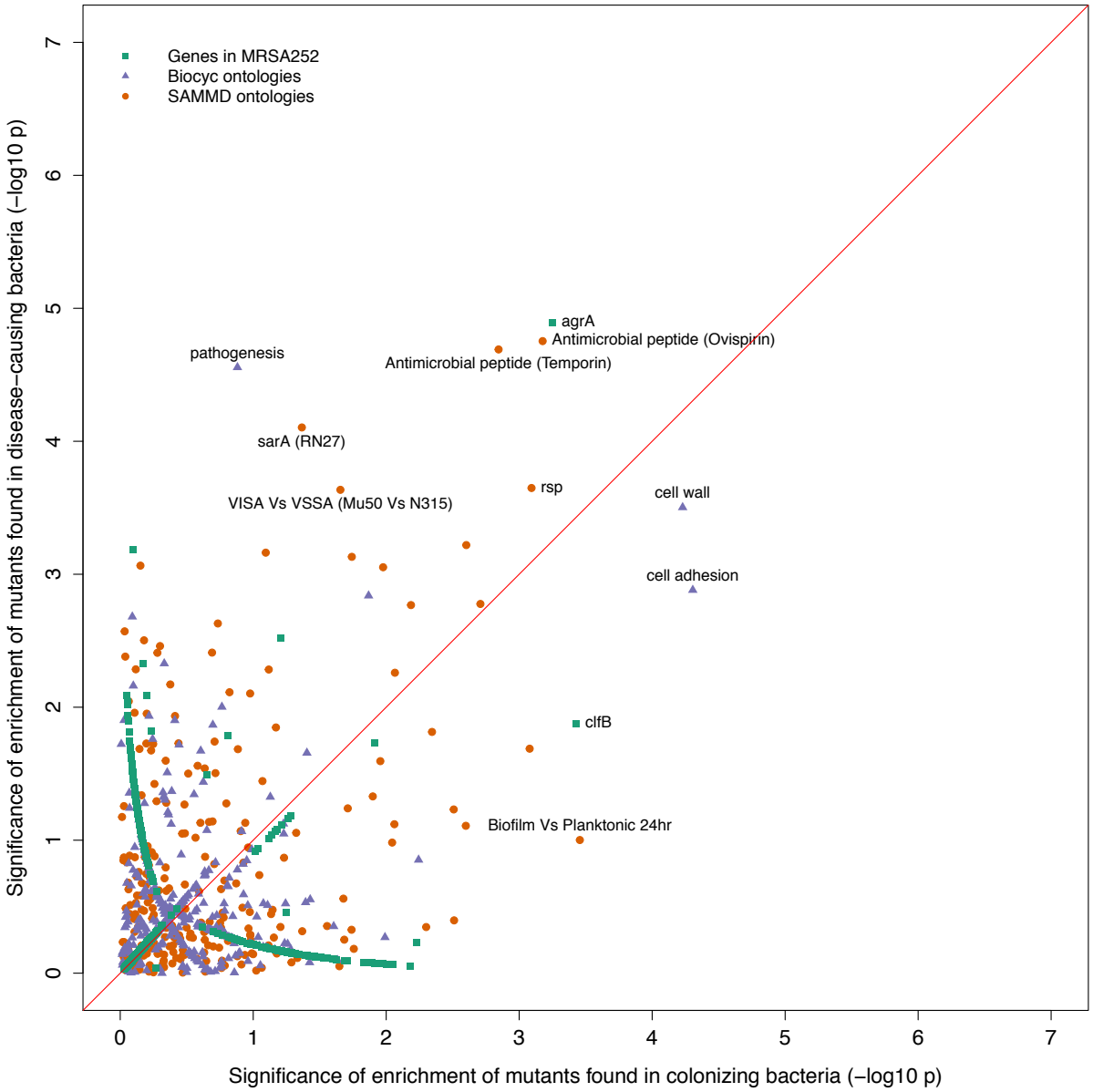


Fig. S3. **Gene set enrichment analysis of B-class mutants occurring in the nose or the infection site**. Each point indicates the –log10 *p*-values of two tests for enrichment of protein-altering variants found among mutants in nose-colonizing bacteria vs disease-causing bacteria. The shape of each 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. A $-\log_{10}$ *p*-value above 5.2, 4.5 or 4.2 was considered genome-wide significant for loci, gene ontologies or expression pathways respectively.

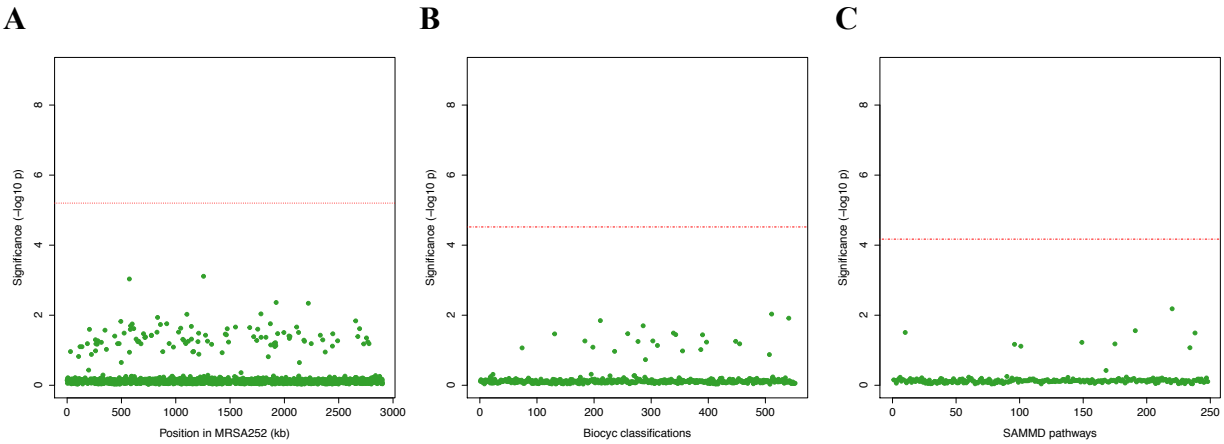


Fig. S4. Genes, ontologies and pathways enriched for protein-altering variants among longitudinally sampled asymptomatic nasal carriers. (**A**) Significance of enrichment of 2650 individual genes. (**B**) Significance of enrichment of 552 gene sets defined by BioCyc gene ontologies. (**C**) Significance of enrichment of 248 gene sets defined by SAMMD expression pathways. Genes, pathways and ontologies that approach or exceed a Bonferroni-corrected significance threshold of $\alpha = 0.05$, weighted for the number of tests per category, (red lines) are named.

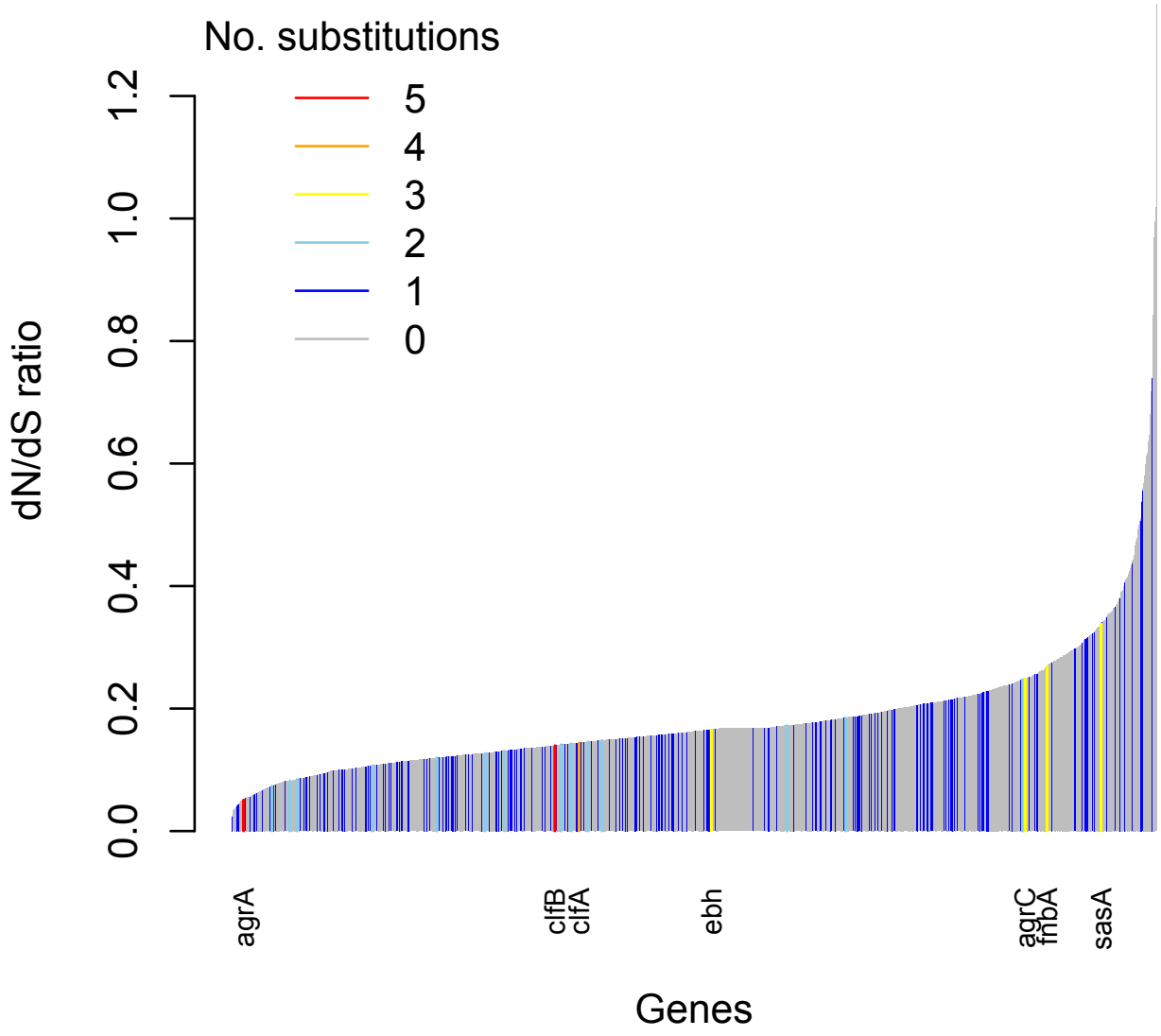


Fig. S5. Genes enriched for substitutions between nose-colonizing and disease-causing bacteria within patients are not the most rapidly evolving at the species level. An estimate of the d_N/d_S ratio between unrelated bacteria is shown for each gene, color-coded by the number of protein-altering substitutions 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 ($\rho = -0.04$, $p = 0.02$).

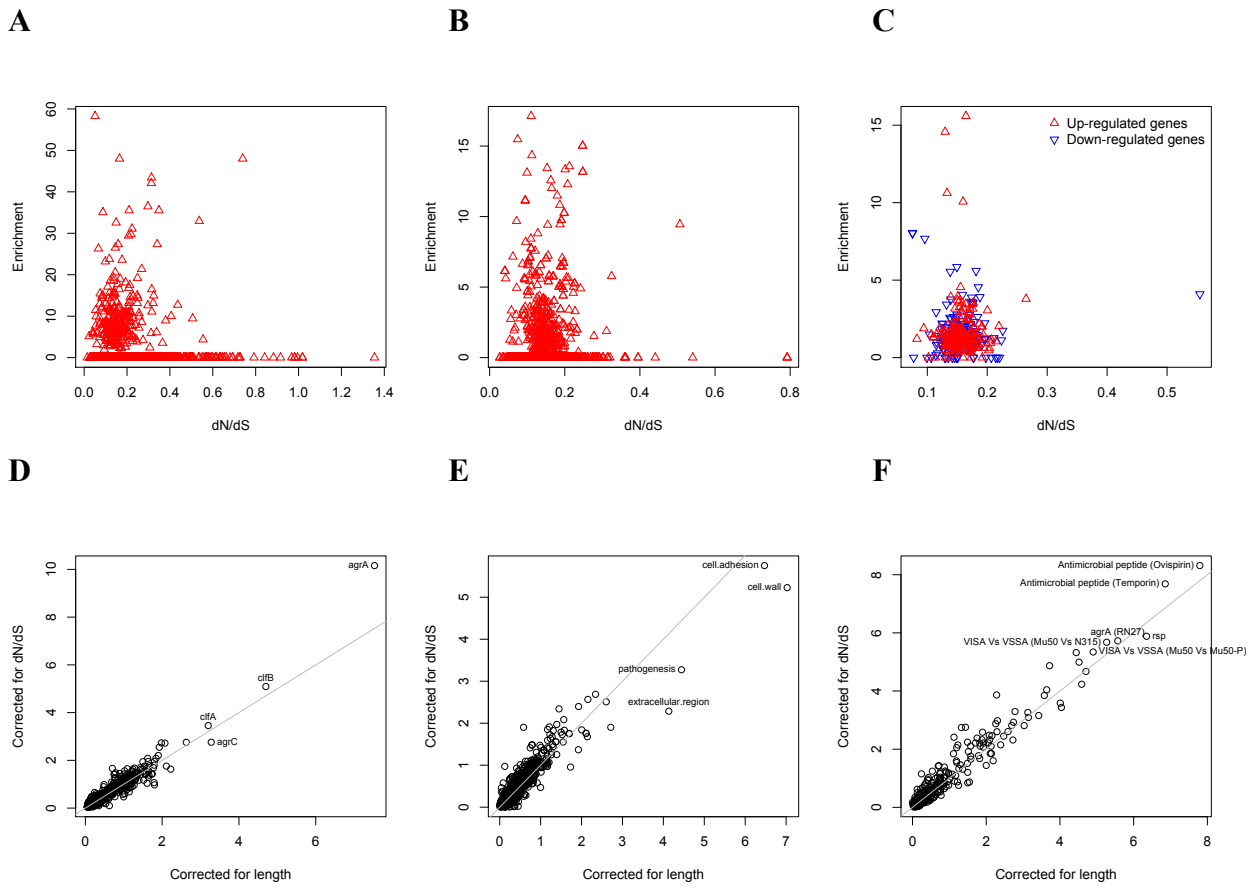


Fig. S6. **Gene set enrichment analysis is robust to species-level differences in d_N/d_S between genes**. For every locus, expression pathway and gene ontology, we estimated d_N/d_S between unrelated *S. aureus*. There was no relationship between d_N/d_S and enrichment of protein-altering substitutions 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).

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

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

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

272

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**

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