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

Major category: BIOLOGICAL SCIENCES Minor category: Evolution

Bernadette C. Young[a,1], Tanya Golubchik[b,1], Elizabeth M. Batty[b], Rowena Fung[a,c], Hanna Larner-Svensson[d], Antonina Votintseva[a], Ruth R. Miller[a], Heather Godwin[e], Kyle Knox[f], Richard G. Everitt[a], Zamin Iqbal[d], Andrew J. Rimmer[d], Madeleine Cule[b], Camilla L. C. Ip[b], Xavier Didelot[b], Rosalind M. Harding[g], Peter J. Donnelly[b,d], Tim E. Peto[a,c], Derrick W. Crook[a,c,2], Rory Bowden[b,c,d,2], Daniel J. Wilson[a,d,2,3]

a. Nuffield Department of Clinical Medicine, Experimental Medicine Division, John Radcliffe Hospital, Oxford, OX3 9DU, United Kingdom

b. Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom

c. NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Headley Way Oxford, OX3 9DU, United Kingdom

d. Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom

e. Oxford University Hospitals NHS Trust, Headley Way, Oxford, OX3 9DU, United Kingdom

f. Department of Primary Care Health Sciences, University of Oxford, 23-38 Hythe Bridge Street, Oxford, OX1 2ET, United Kingdom

g. Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, United Kingdom

1. These authors contributed equally to this work.

2. These authors jointly supervised this work.

3. To whom correspondence should be addressed. Email daniel.wilson@ndm.ox.ac.uk.
Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom. Telephone +44 1865 617 894. Fax +44 1865 287 501

Whole genome sequencing offers new insights into the evolution of bacterial pathogens and the etiology of bacterial disease. *Staphylococcus aureus* is a major cause of bacteria-associated mortality and invasive disease, and is carried asymptomatically by 27% of adults. 80% of bacteremias match the carried strain. However, the role of evolutionary change in the pathogen during the progression from carriage to disease is incompletely understood. Here we use high-throughput genome sequencing to discover the genetic changes that accompany the transition from nasal carriage to fatal bloodstream infection in an individual colonized with methicillin-sensitive *S. aureus* (MSSA). We found a single, cohesive population exhibiting a repertoire of 30 single nucleotide polymorphisms (SNPs) and four insertion/deletion (indel) variants. Mutations accumulated at a steady rate over a 13 month period, except for a cluster of mutations preceding the transition to disease. Although bloodstream bacteria differed by just eight mutations from the original nasally carried bacteria, half of those mutations caused truncation of proteins, including a premature stop codon in an *AraC*-family transcriptional regulator that has been implicated in pathogenicity. Comparison to evolution in two asymptomatic carriers supported the conclusion that clusters of protein-truncating mutations are highly unusual. Our results demonstrate that bacterial diversity *in vivo* is limited but nonetheless detectable by whole genome sequencing, enabling the study of evolutionary dynamics within the host. Regulatory or structural changes that occur during carriage may be functionally important for pathogenesis, so identifying those changes is a crucial step in understanding the biological causes of invasive bacterial disease.

\body

The last 15 years have seen great advances in understanding within-host evolutionary dynamics of important viral pathogens, providing insights that high-throughput whole genome sequencing now promises for the study of evolution in bacterial pathogens and their mechanisms of disease (1-3). Viral evolutionary dynamics within the host can be used to predict the stage of disease, while adaptation is known to act as a trigger for disease progression (4-5). Bacteria have larger genomes than viruses that mutate at a slower rate owing to more faithful DNA replication and mismatch repair systems, so variation that is informative about within-host population dynamics and functional adaptation has proved much harder to capture (6). While within-host evolution has been reported in bacterial housekeeping genes (7), and is well established in hypervariable loci encoding surface antigens or regulating phase variation (8, 9), obtaining an exhaustive catalogue of change in the whole genome is a prerequisite to understanding the genetic basis of pathogenesis.

In vivo evolution of the pathogen population is an important avenue for research into the etiology of bacterial disease. *Staphylococcus aureus* is a major bacterial cause of life-threatening hospital- and community-acquired infections that has come to prominence through the rise of drug-resistant forms, notably methicillin-resistant *S. aureus* (MRSA) (10). *S. aureus*, in common with other important bacterial pathogens, is a predominantly commensal organism and a common constituent of the nasal flora of healthy adults (11, 12), although it is undoubtedly adapted to facultative pathogenicity (13). Carriage predisposes to invasive disease; in a study of bacteremic patients, staphylococci isolated from the blood were undifferentiable by pulsed-field gel

electrophoresis from concomitantly carried nasal bacteria in the majority of cases (12). However, the events leading to invasive disease are incompletely understood (14, 15). While disease may follow ingress of commensal flora to the bloodstream through compromised epithelia, another possibility is that subtle changes in the host-pathogen interaction precipitate the onset of disease. Indeed, spontaneous mutation in the pathogen has been shown to radically alter virulence in mouse models of *S. aureus* infection (16). Regulatory protein dysfunction has been demonstrated to increase bacterial pathogenicity (17), including increased mortality in *S. aureus* bacteraemia (18).

In order to better understand the biology of *S. aureus* carriage, we surveyed nasal carriage in more than 1,100 adults attending general practices in Oxfordshire, UK. We recruited 360 carriers for regular screening. One elderly participant (participant P) developed a *S. aureus* bloodstream infection 15 months after joining the study. Using whole genome sequencing, we charted the evolution of the bacterial population during carriage and disease, and contrasted it to on-going evolution in two asymptomatic carriers (Q and R). Our results reveal dynamic populations of nasally carried staphylococci, harboring genetic variation that evolves measurably over time. We found that the number of mutations separating disease-causing from asymptomatically carried bacteria in participant P was very few. However, the clustering of protein-truncating mutations preceding disease progression, including a transcriptional regulator of stress response and pathogenesis, was a unique pattern absent from the asymptomatic carriers, and suggests a role for loss-of-function mutations in bacterial pathogenesis.

Results

Invasive bloodstream bacteria emerged from a nasal population of methicillin-sensitive *Staphylococcus aureus*. We used Illumina HiSeq 2000 to sequence the genomes of 68 colonies isolated from six nasal swabs and a blood culture from participant P (Fig. 1a & Table S1). The nasal swabs sent for sequencing fell into two groups: early nasal cultures (ENC) comprised five swabs (N1, N2, N4, N6 and N8) representing prolonged stable carriage over six months. In month 10 a nasal swab showed no growth, suggesting this stable carriage had been disturbed, possibly in connection to drug intervention in the form of amoxicillin prescribed for a cough in month 9 (although subsequent testing demonstrated resistance to penicillin G). The late nasal cultures (LNC) comprised a single swab (N12) collected in month 12. Further medical intervention followed. In month 13 a B cell neoplasia with cardiac complications was diagnosed and a permanent pacemaker was fitted under single dose flucloxacillin and penicillin prophylaxis. The following month the participant began a chemotherapy regimen consisting of a proteasome inhibitor and an alkylating agent, along with prophylactic antimicrobials (co-trimoxazole). Sixteen days later the participant developed fever and was admitted to hospital with features of septic shock including neutrophilia. At this point the late blood cultures (LBC) were taken comprising sample B15. No source for the bacteremia was identified: there were no endovascular catheters or evidence of endocarditis or surgical site infection from the pacemaker, but the patient developed fatal multi-organ failure.

Genomic analysis and standard molecular typing indicated a typical community-carried *S. aureus* with no obvious predisposition towards disease. All *S. aureus* samples

were methicillin-sensitive, multilocus sequence type (19) (MLST) ST-15 and *spa* type t4714 (Table S1), a newly-described allele closely related to the commonly carried, community-associated t084 (20). Alignment of genome C1285 from sample N1 to reference genomes MRSA252 and MSSA476 (21) confirmed the absence of the methicillin resistance gene *mecA* and the lack of the staphylococcal cassette chromosome that frequently encodes virulence factors (22) (Fig. S1). Sequence similarity to MSSA476 was 99.5%. No putative virulence factors were found within 13kb of coding sequence that did not align to MSSA476 (Table S2). Two copies of a 20.7kb plasmid 99.9% similar to the MSSA476 pSAS plasmid and containing a region homologous to the staphylococcal transposon Tn552 were detected. MSSA476 prophages φSa3 and φSa4 were absent from C1285, and no other prophages were detected. Pathogenicity islands homologous to νSaα and νSaβ were detected with partial deletions. No virulence or toxin genes were detected beyond those that were present in MSSA476 (Table S3 & Fig. 1b).

Carriage and invasive bacteria formed distinct clades within the host. Extremely limited microvariation was found among the 68 sequenced genomes (Fig. 1), well below that detectable by conventional methods, and consistent with a homogeneous population arising from a single acquisition. There was no variation in MLST, *spa* type or in 61 minisatellite repeats (Table S4), suggesting the resident bacterial population was not eradicated by several episodes of antibiotic treatment. There was no evidence for large-scale indels or copy number variants. We discovered a total of 30 SNPs and four short indels in the 2.7 megabase genome comprising five synonymous, 16 non-synonymous, three premature stop codons and six intergenic SNPs, one of which occurred in the

plasmid, two intergenic indels and two frameshift-inducing indels, one of which led to a premature stop codon (Table S5). There was no homoplasy or evidence for within-host recombination.

Bloodstream colonies (LBC) and nasal colonies (ENC/LNC) formed distinct clades within a population characterized by extremely limited genetic variation. Nasal colonies clustered further into those sampled before (ENC) and after (LNC) the first of a series of drug interventions beginning in month 9 (Fig. 1c). Bayesian coalescent analysis showed slow but detectable evolution of the bacterial population (Fig. 2), at a rate of 2.72 mutations per megabase per year (95% CI 1.64-4.42), close to other estimates of the short-term mutation rate in *S. aureus* (2, 23, 24). A similar molecular clock rate was estimated for ENC sequences alone, but between the ENC clade and the LNC/LBC clades there was greater sequence divergence than expected (Fig. S2), suggesting a departure from the neutral evolutionary model. The estimated effective population size was small, corresponding to an average lifespan of polymorphisms of four months. The most recent common ancestor of all the sequences was dated to one month prior to enrolment (Table S6), but there is no evidence to rule out carriage before this time.

An excess of protein-truncating mutations preceded disease progression. Unusual patterns of molecular evolution were observed along the branches separating early nasal sequences from invasive bloodstream sequences. Tests based on neutral coalescent simulations showed that the most recent common ancestor (MRCA) of ENC and LNC/LBC (Coal.i in Fig. 2 & Table S7) was significantly older than expected ($p = 0.005$). Indeed, five mutations occurred on this branch, while none of the derived

polymorphisms from ENC sequences were retained through to LNC/LBC. This may indicate (i) cryptic populations of differentiated bacteria within the host, (ii) reseeding by a latent population of ancestral genotypes, (iii) adaptive evolution, or (iv) relaxed functional constraint associated with a population bottleneck. A similar pattern was seen for the branch leading to the LBC sequences, which coincided with periods of anti-neoplastic chemotherapy and antibiotic treatment. The MRCA of LBC and LNC (Coal.ii in Fig. 2 & Table S7) was also significantly older than expected ($p = 0.021$). Whereas LNC sequences were diverse, no SNPs were observed within LBC. This represents significantly reduced diversity (Fig. S3) that is unlikely to be due to sampling limitations since LBC was derived from three separate blood culture bottles.

The distribution of premature stop codons among the branches of the tree was uneven, with a significant excess on the two branches leading from ENC to LBC (Fisher's exact test, $p = 0.0015$; Table 1). Four SNPs and one indel separated the ENC clade from the LNC/LBC clades, including a premature stop codon in an AraC family transcriptional regulator (AFTR), which presents the best current candidate for functional mutation. AFTRs are regulators of carbon metabolism, stress response and virulence that respond to changing environmental conditions such as antibiotic usage and oxidative stress (17). In *Neisseria meningitidis* a pseudogene induced by a premature stop codon in the AFTR *mpeR* is associated with the hypervirulent ST 32 complex (25). The mutation we observed maps to MSSA476 SAS2271, radically truncating the sequence from 702 to 77 amino acids. A premature stop codon induced by a frameshift on the same branch occurred in a protein of unknown function, SAS1429. We observed two further premature stop codons predicting significantly truncated proteins on the branch leading to

the LBC clade: SAS0973, an iron-compound binding protein/transporter, and SAS1361, a GNAT family acetyltransferase.

Clusters of protein-truncating mutations were not observed in non-invasive carriage populations. To investigate the evolution of *S. aureus* during asymptomatic nasal carriage, we used the Illumina GAIIx and HiSeq 2000 platforms to sequence the whole genomes of 101 colonies isolated from two other participants (Table S1). Twenty-two colonies isolated from two swabs taken two months apart were sequenced from participant Q, who had no history of staphylococcal disease or recent antibiotic usage. Seventy-nine colonies isolated from eight swabs taken over an 18-month period were sequenced from participant R, who similarly had no history of staphylococcal disease. However, participant R had completed a treatment of flucloxacillin shortly prior to enrolment in the carriage study, and took a course of amoxicillin in month 20. No bacterial growth was detected in any of six nasal swabs taken from months 22 to 32, suggesting that carriage was cleared. The bacterial populations in both carriers were MSSA, and both exhibited a single *spa* type (t164 and t012 respectively) and multilocus sequence type (ST-20 and ST-30 respectively), consistent with a single founding colonization in each case. The repertoire of virulence and toxin genes was indistinguishable among genomes sequenced within a single participant, and more similar between the genomes sequenced from participants P and Q than those from R (Table S3).

Limited microvariation was detected in both participants Q and R. We discovered a total of 42 SNPs and four short indels in participant Q comprising 10 synonymous, 20 non-synonymous, one premature stop codon and 11 intergenic SNPs, two intergenic

indels and two frameshift-inducing indels both of which led to premature stop codons (Table S5). Two large deletions were also detected in one of the genomes: an 8 kb deletion partially matching *S. aureus* pathogenicity island SaPI4, and a 1.6 kb deletion of an integrase. In participant R we discovered a total of 39 SNPs and nine short indels comprising 14 synonymous, 15 non-synonymous, one premature stop codon and nine intergenic SNPs, six intergenic indels and three frameshift-inducing indels, all of which led to premature stop codons. There was no significant difference in the overall pattern of mutation types across participants P, Q and R (Fisher's exact test, $p = 0.457$). As in participant P, there was no homoplasy or evidence for within-host recombination in participants Q and R.

Rather than forming distinct clusters, the colonies isolated two months apart in participant Q were genetically overlapping, with the descendants from multiple lineages detected in the earlier nasal swab present in the latter (Fig. S4). There was a clearer temporal trend in participant R, such that the diversity sampled at one time was usually descended from a single one of the lineages present in the previous sample, leading to a steady accumulation of mutations over time. Bayesian coalescent analysis revealed a molecular clock rate of 1.87 mutations per megabase per year in participant R (95% CI 1.08-3.06), consistent with the rate estimated in P. There was insufficient power to independently estimate the rate of evolution in participant Q. Assuming the same clock rate as in R implies a large effective population size in Q, corresponding to an average lifespan of polymorphisms of 17 months. The effective population size estimated for participant R was intermediate between P and Q, with an average lifespan of polymorphisms of five months (Table S6).

The evidence from participants Q and R provided additional support for the view that a significant excess of protein-truncating mutations occurred on the two branches separating the genomes sampled early during asymptomatic nasal carriage (ENC) from those sampled from the invasive bloodstream infection (LBC) in participant P. Three out of 48 mutations detected in participant Q, and four of 48 mutations detected in participant R were protein-truncating. Treating the mutations in Q and R as control groups confirmed that the number of premature stop codons occurring on the ENC-LBC branches in participant P was statistically significant (Table 1). To maximize statistical power, we combined information across participants P, Q and R, yielding a highly significant *p*-value of 0.0017. To further investigate the unusual clustering of premature stop codons in participant P, we constructed an empirical distribution for this *p*-value by considering every possible pair of branches occurring in participant P, Q or R. None of the 589 possible permutations yielded a *p*-value as significant as 0.0017 (Figure S5), demonstrating that the cluster of protein-truncating mutations on the two branches leading from ENC to LBC was indeed highly unusual.

Discussion

Just eight mutations accompanied the transition of an asymptomatically carried MSSA population to a fatal bloodstream infection. Half of those mutations were premature stop codons, one of which truncated a putative transcriptional regulator of virulence (17). Two further premature stop codons were detected only among invasive bloodstream bacteria. Loss of function mutations that truncate the amino acid sequence may play an important role in pathogenesis, as point mutations of this sort can quickly effect radical functional change (18). However, the patient's general health was also

likely to have been important, with the interaction between genome evolution and clinical context likely to be critical.

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

Materials and Methods

Isolate collection and preliminary analysis. Each nasal swab culture had been prepared and stored in glycerol. We incubated an inoculum of the glycerol stock on SASelect agar (BioRad) overnight at 37C, then picked twelve colonies, streaked each onto Columbia blood agar and incubated overnight at 37C. Methicillin sensitivity was determined by the disc diffusion method. Blood cultures were prepared using the BD Bactec system; blood was drawn from the patient at two times six hours apart and each sample was inoculated separately into two bottles. Both bottles from the first sample and one of two from the second flagged positive for bacterial growth. Blood extracted from the bottles was cultured on SAselect (BioRad) agar and four colonies were picked from each bottle for sequencing. DNA was extracted using a commercial kit (FastDNA by MP Biomedicals)

using mechanical disruption of bacteria and column based purification of DNA. Staphylococcal protein A (*spa*) type was determined by Sanger sequencing of the variable X region of the 3′ end of the *spa* gene, using commercially designed primers (spaF 5′-AGACGATCCTTCGGTGAGC-3′ spaR 5′-GCTTTTGCAATGTCATTTACTG-3′). The software Ridom StaphType (26) was used for *spa* sequence analysis.

Sequencing and assembly. For samples Q2, Q4, R2 and R4 we used the Illumina GAIIx platform with 12-fold multiplexing, read lengths of 51 base pairs (bp), insert sizes of 200bp and mean depth 62.9 reads. For the remaining samples we used the Illumina HiSeq 2000 platform with 96-fold multiplexing, read lengths of 99bp, insert sizes of 200bp and mean depth 214 reads. In 68 out of 84 colonies (participant P), 22/24 colonies (participant Q) and 79/96 colonies (participant R) we successfully performed DNA extraction, library preparation and sequencing to a standard that passed stringent quality control measures. We used Velvet (27) to assemble reads into contigs *de novo* for each genome. We used Stampy (28) with no BWA pre-mapping and an expected substitution rate of 0.01 to map each genome against a host-specific internal reference genome comprising the contigs assembled for genomes C1285 (2.68Mb), C0965 (2.59Mb) or C0764 (2.76Mb). These represent the earliest sequenced sample group in participants P, Q and R respectively, comprising 137, 948 and 167 contigs with an N50 of 71,927, 22,136 and 148,695 bp respectively. 99.5%, 96.9% and 96.6% of reads in C1285, C0965 and C0764 mapped to the respective *de novo* assemblies. Participant P genomes were also mapped to the MRSA252 (2.90Mb) and MSSA476 (2.80Mb) references (21). Repetitive regions, defined by BLASTing the reference genome against itself, were masked prior to variant

calling. This masked 4.0%, 2.2% and 2.7% of the host-specific reference genomes respectively, 5.9% of MRSA252 and 4.5% of MSSA476. The average proportion of the reference genome that we called by mapping was 92%, 92%, 93%, 85% and 81% respectively.

Variant calling. We used SAMtools (29) and Picard (http://picard.sourceforge.net) to call variants from mapping, which we then filtered using criteria including base quality, mapping quality and depth. We used Cortex (30) to detect SNPs and short indels. Visual inspection of every filtered and unfiltered variant call in participant P was used to manually validate the approach. To detect large deletions relative to the host-specific references, the mapped read depth was scanned for regions of at least 1kb in which 500bp or more exhibited zero coverage. To detect large insertions relative to the host-specific references, unmapped reads were assembled by Velvet with a hash length 31bp. We validated our ability to identify large indels by comparing our sequences with MSSA476. We used Tandem Repeats Finder (31) to identify minisatellite-like repeats in the MRSA252 reference, and searched for their flanking sequences in each genome using BLAST. Indeterminate results were obtained when just one of the flanking sequences was found, or when the two were located on different contigs. We likewise used BLAST to search for the presence of known or putative virulence factors, including toxins, adhesins and regulators (32-34).

Experimental validation. We chose the four protein-truncating mutations detected in participant P for validation using PCR and capillary sequencing. The variants detected at positions 1043150 (C→A), 1458121 (G→A), 1555915 (deletion of A) and 2430183 (C→T) relative to the MSSA476 genome were successfully amplified and sequenced in

each of two single-colony isolates from N1 and two single-colony isolates from B15 using the following primer pairs: F1043150 5′-GATTTTAGCCACTGACGGGA-3′ and R1043150 5′-ATGTAACGATGCGCCAATTC-3′, F1458121 5′-ATACGTGTCCAACTGTTCCC-3′ and R1458121 5′-GGCGCCTTTGTTATTCATCG-3′, F1555915 5′-GCAATCGAATCTCCTGTCCA-3′ and R1555915 5′-ACATTAGTGATGGTGTGCCC-3′, F2430183 5′-TGGTGAAACCAAAGACGTAAG-3′ and R2430183 5′-GTCTATGAACACCGGATTGCT-3′. For every variant both the N1 and both the B15 isolates showed the expected sequence, confirming the existence of the variant in our samples.

Mobile elements. We used BLAST to search for short flanking sequences of six SCC-associated loci (22) in C1285 using a word size of 16. As a control, we repeated the searches in MRSA252 (21) and MSSA476. We used ClustalX (35) to align each pair of SCC direct repeats in MRSA252 and MSSA476 to C1285. We used xbase (36) to align C1285 against MRSA252 and MSSA476 using MUMMER (37) and annotate the genome. The alignment in the SCC region was inspected using the Artemis Comparison Tool (38). We used BLAST to search for *S. aureus* transposons Tn552 and Tn554 in C1285. We used Stampy (28), MUMMER (37) and Mauve (39) to search for MSSA476 phages φSa3 and φSa4 and pathogenicity islands νSaα and νSaβ. We searched for novel prophages using Prophage Finder (40).

Population genetics analysis. We employed a permutation test for recombination that detects any correlation between physical distance and linkage disequilibrium (41). We inferred tree topology and branch lengths using maximum likelihood (ML) under the assumption of no repeat mutation and homogeneous mutation rates. We used the ML tree

to reconstruct haplotypes. We performed Bayesian coalescent inference to estimate evolutionary parameters including the molecular rate using BEAST (42), assuming constant population size and the HKY mutation model (43). All validated SNPs were included, together with 1% of invariant nucleotides. Separate analyses of participant P, Q and R genomes were undertaken, along with separate analyses of ENC alone and LNC and LBC together. Further analyses of each sample (N1, N2, N4, N6, N8, N12 and B15) within participant P were undertaken to estimate diversity ($\theta = 2N_e g\mu$) for each group separately by fixing μ. For the analysis of participant Q sequences there was insufficient power to estimate μ; instead μ was fixed at the rate estimated for participant R. In all cases, we assumed an improper uniform prior on $N_e g$ (the product of effective population size and generation length), an improper uniform prior on μ (mutation rate per day, unless fixed), a uniform prior on nucleotide frequencies, and a log-normal prior on κ (transition:transversion ratio) with mean 1 and standard deviation 1.25 on the logarithmic scale. Pairs of chains of 10 million iterations each were run, sampled every 1,000 iterations with a burn-in of 100,000 iterations removed before merging the chains to obtain final results. We quote the posterior median and (2.5%, 97.5%) quantiles as point estimates and credible intervals respectively. To obtain the maximum clade credibility tree using BEAST, we employed an outgroup constructed using 1% of the fixed differences between MSSA476 and the host-specific internal reference, which allowed us to infer the direction of mutation. To remedy the strong leverage the outgroup sequence has on estimates of the molecular rate, we assumed an uninformative improper uniform prior on the sampling date of the outgroup sequence. A pair of chains of 400 million

iterations were run, sampled every 10,000 iterations with a burn-in of 100,000 iterations removed before merging the chains to obtain final results.

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

Acknowledgements

We thank L. O'Connor, A. S. Walker, P. Piazza and the Oxford MRC High Throughput Sequencing Hub team. This study was supported by the Oxford NIHR Biomedical Research Centre and the UKCRC Modernising Medical Microbiology Consortium, the

latter funded under the UKCRC Translational Infection Research Initiative supported by Medical Research Council, Biotechnology and Biological Sciences Research Council and the National Institute for Health Research on behalf of the Department of Health (Grant G0800778) and the Wellcome Trust (Grant 087646/Z/08/Z). We acknowledge the support of Wellcome Trust core funding (Grant 090532/Z/09/Z). TEP and DWC are NIHR Oxford Biomedical Research Centre senior investigators. Ethical approval for the carriage study was obtained from the Oxfordshire B Oxfordshire Research Ethics Committee (reference number 08/H0605/102).

References

1. Pybus OG, Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet* 10:540-550.
2. Harris SR, et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327:469-474.
3. Mwangi MM, et al. (2007) Tracking the *in vivo* evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. *Proc Natl Acad Sci USA* 104:9451-9456.
4. Lemey P, et al. (2007) Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Comput. Biol.* 3:e29.
5. Connor RI, et al. (1997) Change in coreceptor use correlates with disease progression in HIV-1–infected individuals. *J Exp Med* 185:621-628.
6. Drake JW, et al. (1998) Rates of spontaneous mutation. *Genetics* 148:1667-1686.

7. Pérez-Losada M, et al. (2007) Temporal trends in gonococcal population genetics in a high prevalence urban community. *Infect Genet Evol* 7:221-278.
8. Boye K, Westh H (2011) Variation in *spa* types found in consecutive MRSA isolates from the same patients. *FEMS Microbiol Lett* 314:101-105.
9. Bayliss CD (2009) Determinants of phase variation rate and the fitness implications of differing rates for bacterial pathogens and commensals. *FEMS Microbiol Rev* 33:504-520.
10. Thwaites G, Gant V (2011) Are bloodstream leukocytes Trojan Horses for the metastasis of Staphylococcus aureus? *Nat Rev Microbiol* 9:215-222.
11. Wertheim HF, et al. (2005) The role of nasal carriage in *Staphylococcus aureus* infections. *Lancet Infect Dis* 5:751-762.
12. von Eiff C, et al. (2001) Nasal carriage as a source of *Staphylococcus aureus* bacteremia. *N Engl J Med* 344:11-16.
13. Foster TJ (2005) Immune evasion by staphylococci. *Nat Rev Genet* 3:948-958.
14. Goerke C, Wolz C (2004) Regulatory and genome plasticity of *Staphylococcus aureus* during persistent colonization and infection. *Int J Med Microbiol* 294:195-202.
15. Edwards AM, Massey RC (2011) How does *Staphylococcus aureus* escape the bloodstream? *Trends Microbiol* 19:184-190.
16. Kennedy AD, et al. (2008) Epidemic community-associated methicillin-resistant *Staphylococcus aureus*: recent clonal expansion and diversification. *Proc Natl Acad Sci USA* 105:1327-1332.

17. Yang J, Tauschek M, Robins-Browne RM (2011) Control of bacterial virulence by AraC-like regulators that respond to chemical signals. *Trends Microbiol* 19:128-135.
18. Schweizer ML, et al. (2011) Increased mortality with accessory gene regulator (agr) dysfunction in *Staphylococcus aureus* among bacteremic patients. *Antimicrob Agents Chemother* 55:1082-1087.
19. Enright MC, et al. (2000) Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus. *J Clin Microbiol* 38:1008-1015.
20. Skråmm I, Moen AEF, Bukholm G (2011) Nasal carriage of *Staphylococcus aureus*: frequency and molecular diversity in a randomly sampled Norweigan community population. *APMIS*, in press.
21. Holden MTG, et al. (2004) Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci USA* 101:9786-9791.
22. Noto MJ, et al. (2009) Gene acquisition at the insertion site for SCCmec, the genomic island conferring methicillin resistance in *Staphylococcus aureus*. *J Bacteriol* 190:1276-1283.
23. Smyth DS, et al. (2009) Population structure of a hybrid clonal group of methicillin-resistant *Staphylococcus aureus*, ST239-MRSA-III. *PLoS ONE* 5:e8582.

24. Nübel U, et al. (2010) A timescale for evolution, population expansion, and spatial scale of an emerging clone of methicillin-resistant *Staphylococcus aureus*. *PLoS Pathog* 6: e1000855.
25. Fantappie L, Scarlato V, Delany I (2011) Identification of the *in vitro* target of an iron-responsive AraC like protein from meningococcus that is in a regulatory cascade with Fur. *Microbiol,* in press.
26. Harmsen D, et al. (2003) Typing of methicillin-resistant *Staphylococcus aureus* in a university hospital setting by using novel software for *spa* repeat determination and database management. *J Clin Microbiol* 41:5442-5448.
27. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821-829.
28. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 21:936-939.
29. Li H, et al. (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25:2078-2079.
30. Iqbal Z, et al. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet*, in press.
31. Bensom G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucl Acids Res* 27:573-580.
32. Lindsay JA, et al. (2006) Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes. *J Bacteriol* 188:669-676.

33. Jarraud S, et al. (2002) Relationships between *Staphylococcus aureus* genetic background, virulence factors, agr groups (alleles), and human disease. *Infect Immun* 70:631-641.
34. Tristan A, et al. (2007) Virulence determinants in community and hospital meticillin-resistant *Staphylococcus aureus*. *J Hosp Infect* 65:S105-S109.
35. Thompson JD, et al. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl Acids Res* 25:4876-4882.
36. Chaudhuri RR, et al. (2008) xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucl Acids Res* 36:D543-6.
37. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.
38. Carver TJ, et al. (2005) ACT: the Artemis Comparison Tool. *Bioinformatics* 21:3422-3423.
39. Darling AE, Mau B, Perna NT (2010) ProgressiveMauve: multiple genome alignment with gene gain, loss, and rearrangement. *PLoS ONE* 5:e11147.
40. Bose M, Barber RD (2006) Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol* 6:0020.
41. Wilson DJ, McVean G (2006) Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* 172:1411-1425.
42. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214.

43. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22:160-174.

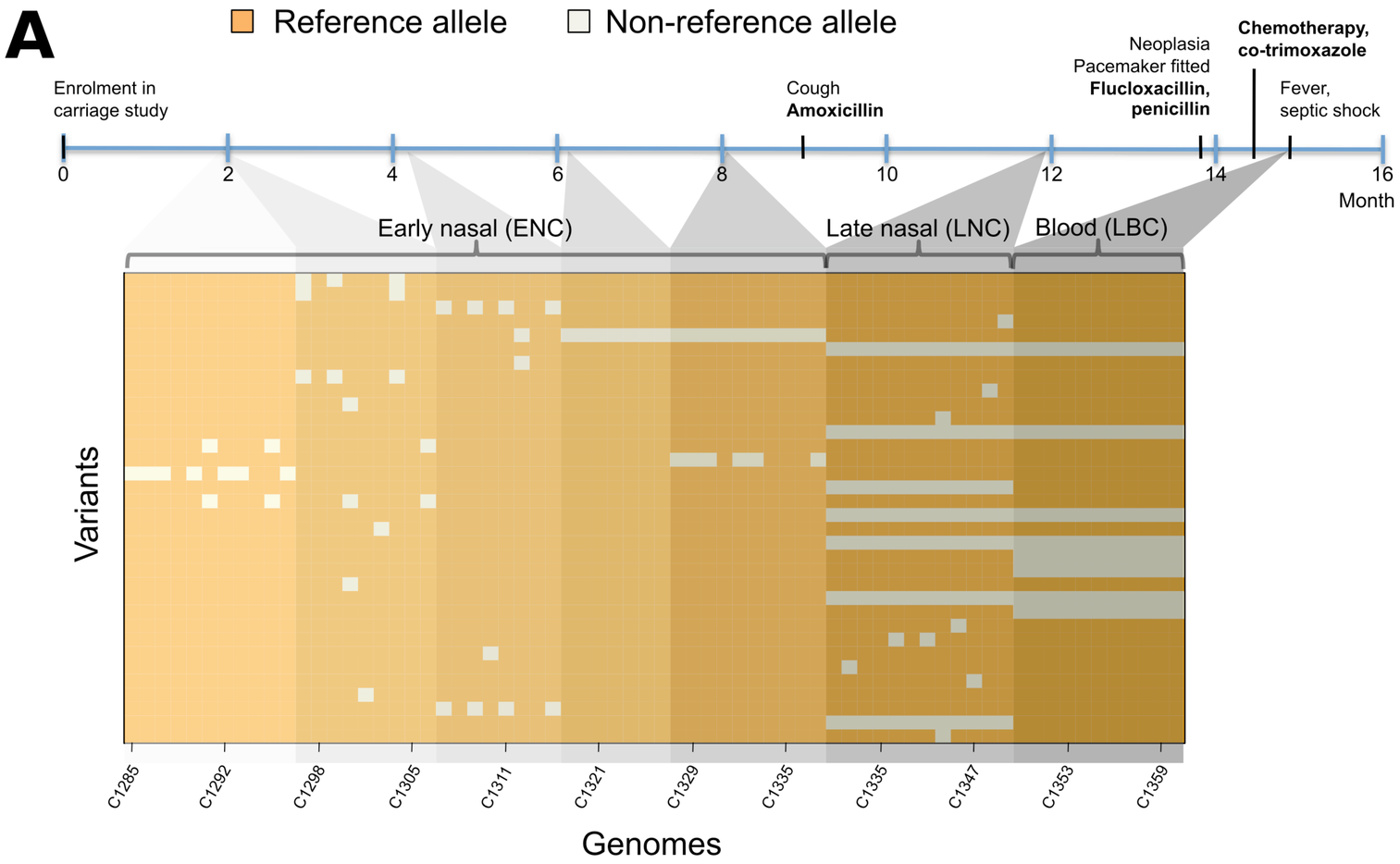
Figure Legends

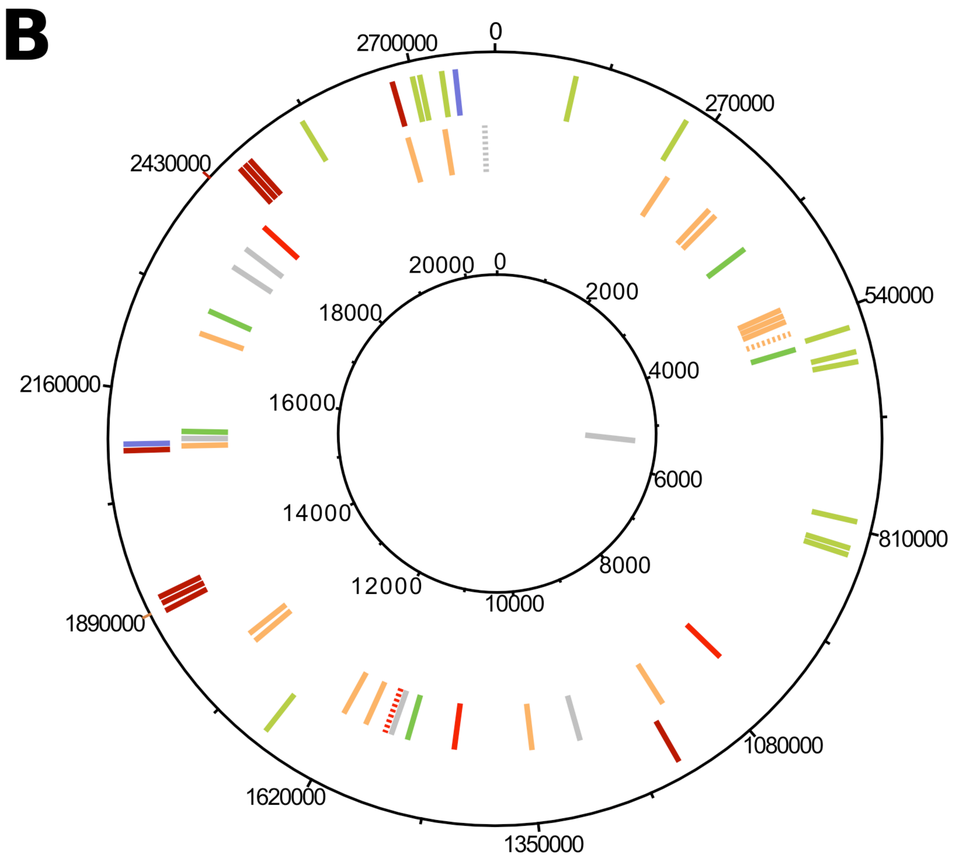
1. Molecular diversity during progression from carriage to disease in participant P. (A) Sampling frame, variants, and the timeline of disease progression. Seven groups of colonies sequenced from six nasal swabs and blood culture are shaded from light to dark grey. (B) The location of virulence factors and variants in the chromosome (outer ring) and plasmid (inner ring). Positions are inferred by mapping to MSSA476 and pSAS. From the outermost track inwards: i, Virulence-associated surface proteins (dark red), toxins (olive green) and regulatory genes (dark blue) identified in the bacterial chromosome. ii, Variants detected in the bacterial chromosome by type: synonymous (green), non-synonymous (orange), premature stop codon (red), intergenic (grey). Solid lines represent SNPs and dashed lines represent indels. iii. Variants detected in the plasmid (same color scheme as track 2). (C) The maximum likelihood unrooted tree relating all sequences. Nodes represent genotypes, where area is proportional to sample frequency, and small black circles represent hypothetical intermediate genotypes. Shading within the circles indicates the sample, where darker corresponds to later samples as in part (A). Edges represent mutations, color-coded as in part (B) track ii. The ordering of mutations among hypothetical genotypes is arbitrary. Numbers represent genotypes observed more than once for cross-reference with Fig. 2.
2. Bayesian coalescent tree.

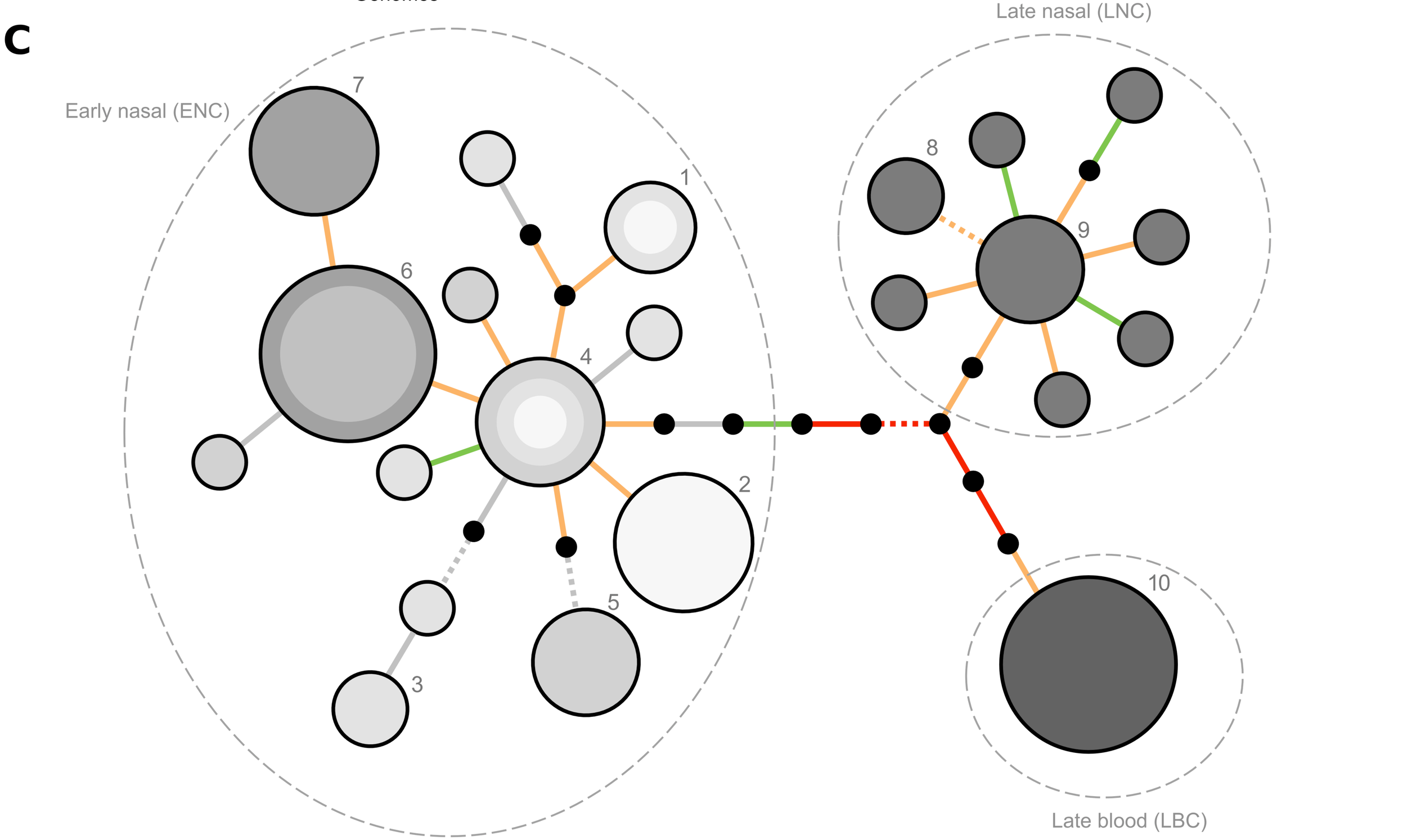
The maximum clade credibility tree representing the genealogy of sequences in the study, reconstructed from SNPs using BEAST. Genotypes are enumerated as per Fig. 1c. SNPs (filled circles) and indels (open circles) are superimposed on the tree, color-coded by type: synonymous (green), non-synonymous (orange), premature stop codon (red), intergenic (grey). The ordering of mutations within a branch is arbitrary.

Table Legends

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


A
Reference allele
Non-reference allele
Enrolment in carriage study
Cough
Amoxicillin
Neoplasia
Pacemaker fitted
Flucloxacillin, penicillin
Chemotherapy, co-trimoxazole
Fever, septic shock
0
2
4
6
8
10
12
14
16
Month
Early nasal (ENC)
Late nasal (LNC)
Blood (LBC)
Variants
C1285
C1292
C1298
C1305
C1311
C1321
C1329
C1335
C1335
C1347
C1353
C1359
Genomes


B
0
270000
540000
810000
1080000
1350000
1620000
1890000
2160000
2430000
2700000
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000


C
Early nasal (ENC)
Late nasal (LNC)
Late blood (LBC)
1
2
3
4
5
6
7
8
9
10

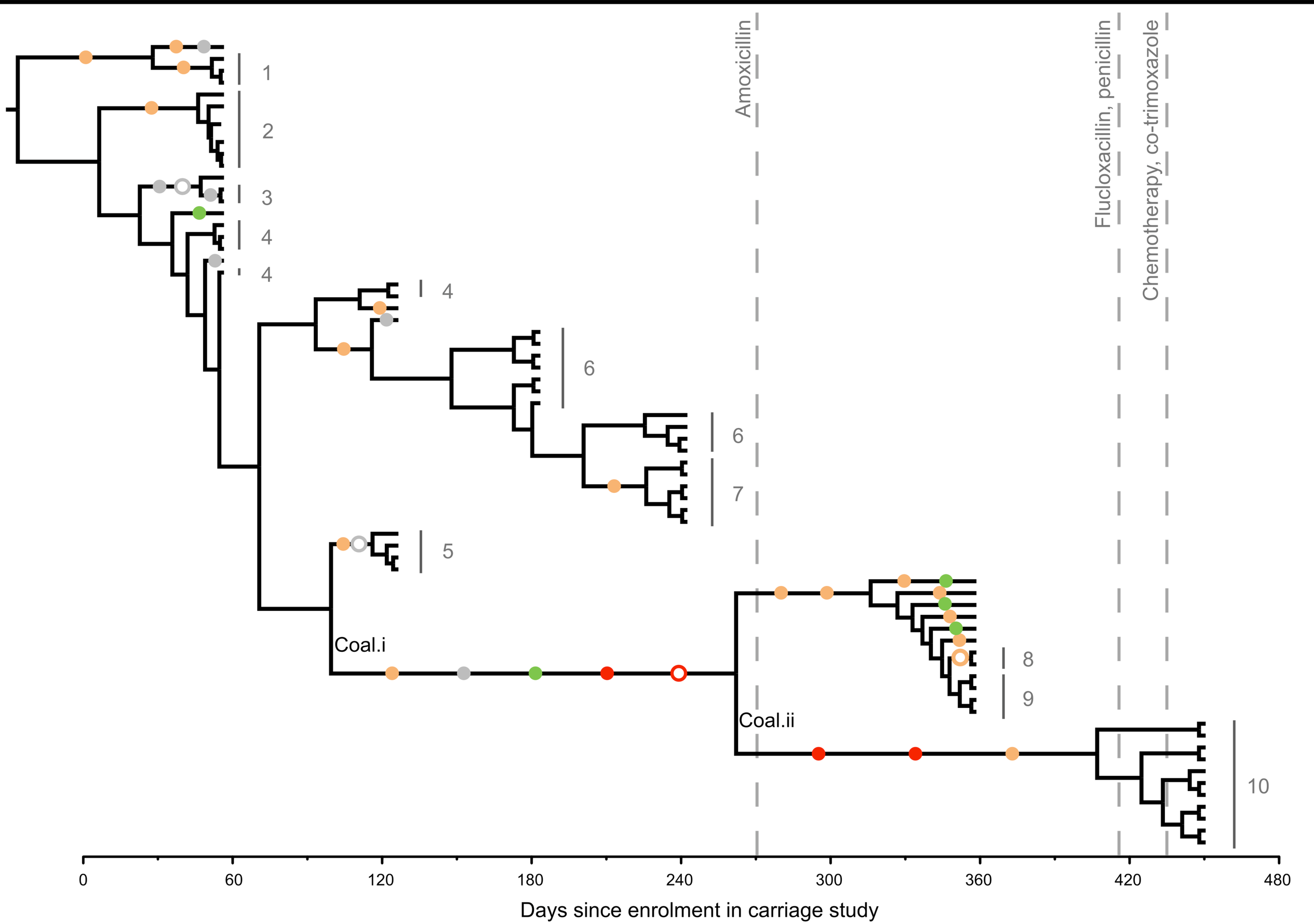

Amoxicillin
Flucloxacillin, penicillin
Chemotherapy, co-trimoxazole
1
2
3
4
4
4
6
6
7
5
Coal.i
8
9
Coal.ii
10
0
60
120
180
240
300
360
420
480
Days since enrolment in carriage study

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

Branches	Number of mutations		*p* value (Fisher's exact test)
	Protein-truncating	Other	
From ENC to LBC	4	4	-
All others in participant P	0	26	0.0015
All others in participant Q	3	45	0.0055
All others in participant R	4	44	0.0103
All others in P, Q and R	7	115	0.0017