Title: Longitudinal genomic surveillance of MRSA reveals extensive transmission in hospitals and the community

Authors: Francesc Coll[1]*, Ewan M. Harrison[2], Michelle S. Toleman[2,5,6], Sandra Reuter[2], Kathy E. Raven[2], Beth Blane[2], Beverley Palmer[3], A. Ruth M. Kappeler[3,4], Nicholas M. Brown[3,5], M Estée Török[2,5], Julian Parkhill[6], Sharon J. Peacock[1,2,5,6]*

Affiliations:

[1]London School of Hygiene & Tropical Medicine, UK.

[2]University of Cambridge, UK

[3]Public Health England, UK

[4]Papworth Hospital NHS Foundation Trust, UK

[5]Cambridge University Hospitals NHS Foundation Trust, UK

[6]Wellcome Trust Sanger Institute, UK

*Corresponding authors: francesc.coll@lshtm.ac.uk and sharon.peacock@lshtm.ac.uk

One Sentence Summary: Integrated longitudinal genomic and epidemiological surveillance of methicillin-resistant *Staphylococcus aureus* (MRSA) in the East of England revealed extensive transmission in hospitals and the community, and provide evidence for the need to review infection control policy and practice.

Abstract: Snapshots of MRSA transmission have been provided by genome sequencing in the context of suspected outbreaks and isolated hospital wards. Scale-up to populations is now required to establish the full potential of this technology for surveillance. We prospectively identified all individuals over 12 months who had at least one MRSA positive sample processed by a routine diagnostic microbiology laboratory in the East of England, which received samples from three hospitals and 75 general practitioner (GP) practices. We sequenced at least one MRSA isolate from 1,465 individuals (2,282 MRSA isolates) and recorded epidemiological data. An integrated epidemiological and phylogenetic analysis revealed 173 transmission clusters containing between 2 and 44 cases and involving 598 people (40.8%). Of these, 118 clusters (371 people) involved hospital contacts alone, 27 clusters (72 people) involved community contacts alone, and 28 clusters (157 people) had both types of contact. Community- and hospital-associated MRSA lineages were equally capable of transmitting in the community, with instances of spread in households, long-term care facilities (LTCF) and GP practices. This provides a comprehensive picture of MRSA transmission in an entire sampled population, and evidence to guide a review of infection control policy and practice.

Introduction

Staphylococcus aureus is responsible for a high proportion of community-associated invasive and soft-tissue infections, and is a leading cause of healthcare-associated infections *(1)*. This burden is compounded by MRSA infection, which results in increased mortality and hospitalization costs and longer hospital stays compared to methicillin-susceptible *S. aureus* infections *(2)*. Successful reduction of MRSA infection rates depends on preventing MRSA transmission and detecting and containing outbreaks *(3)*. Understanding the settings and circumstances under which MRSA evades current infection control measures is central to designing new strategies to reduce transmission.

MRSA carriage and infection has historically been associated with healthcare settings. Recent studies have demonstrated the value of applying whole-genome sequencing (WGS) to define the spread of MRSA *(4–10)* and a range of other pathogens in hospitals. WGS provides the ultimate resolution to discriminate between bacterial isolates, and when combined with epidemiological data enables the reconstruction of transmission networks. Previous studies have largely focused on suspected outbreaks *(4–6)*, or transmission in high-risk settings such as intensive care units *(7–10)*. These snapshots have confirmed the potential of WGS to confirm or refute outbreaks, but the value that could be derived from applying this to entire populations, including those that bridge the divide between hospitals and the community, is unknown. Here, we report the findings of a 12-month prospective study of all MRSA-positive individuals detected by a large diagnostic microbiology laboratory in England in which an integrated analysis of epidemiological and sequence data provided a holistic picture of MRSA transmission.

Results

Study participants and isolates

We identified 1,465 MRSA-positive individuals in the East of England over a 12-month period by screening all samples submitted to a diagnostic microbiology laboratory by three hospitals and 75 GP practices (see Fig. 1 for geographical distribution). Cases had a median age of 68 years (range newborns to 101 years, interquartile range (IQR) 46 to 82 years). We sequenced 2,282 isolates cultured from their multisite screens (n=1,619) or diagnostic specimens (n=663), which equated to one isolate from 1,006 cases and a median of 2 isolates (range 2-15, IQR 2 to 3) from 459 cases (see Supplementary Methods for rationale for selecting isolates for sequencing and fig. S1 for number of isolates sequenced per case). Around 80% of sequenced MRSA isolates were from samples submitted by the three study hospitals (1,453 multisite screens and 372 diagnostic specimens), with the remainder submitted by GP practices (166 multisite screens and 291 diagnostic specimens). Multi-locus sequence types (STs) were derived from sequence data, which revealed that the majority of isolates belonged to clonal complex (CC) 22 (1,667/2,282, 73%), the predominant healthcare-associated lineage in the United Kingdom (UK) *(11)*. This was followed in frequency by CC30 (n=129, 5.6%), CC5 (n=108, 4.7%), CC1 (n=105, 4.6%) and CC8 (n=87, 3.8%) (see table S1 for CC designation of the entire collection). Supplementary Methods provides a detailed description of the patient data collected, microbiology, sequencing methodology and sequence data analyses, and fig. S2 shows a flowchart summarizing the data types used and analyses.

Integration of genomic and epidemiological data

We initially divided the 2,282 MRSA into clusters containing isolates that were no more that 50 single-nucleotide polymorphisms (SNPs) different based on core genome comparisons.

Supplementary Methods describes the rationale for the cut-off used. This led to the identification of 173 separate phylogenetic clusters. MRSA isolated from more than half of cases (785/1465, 53.6%) were genetically linked to MRSA from at least one other case based on isolates belonging to the same cluster. The next step was to apply epidemiological data (hospital admission and ward movement data, GP registration and residential postcode) to this clustering framework to determine links between cases within each cluster, which ignored the traditional categorization of lineages as community- or hospital-associated. Figure S3 provides an overview of how the bacterial phylogeny and patient epidemiological data were integrated to define and classify transmission clusters. This revealed that 598/785 (76.2%) cases had an identifiable MRSA-positive contact with at least one other study case in a hospital setting and/or in the community (Table 1).

It is possible for epidemiological links between MRSA-positive individuals to arise by chance when MRSA carriers are admitted to hospital wards or other healthcare facilities with a high patient turnover and/or a proportionately higher prevalence of MRSA cases than the hospital or community averaged baseline. To assess the potential impact of this we determined the strength of epidemiological links between people with genetically unrelated isolates (separated by more than 50 SNPs). This was achieved by a systematic pairwise comparison of 1040 cases with MRSA CC22. A total of 540,280 unique pairwise case comparisons were made, of which 534,417 had more than 50 SNPs (table S2). The instances of shared wards, GP practices, and postcodes was uncommon (wards/GP practices) or very rare (postcodes) for case-pairs positive for unrelated CC22 MRSA (table S2). This analysis led us to classify shared postcodes (present in 0.04% of genetically unrelated cases), GP practice and ward contacts (<1% of genetically unrelated cases) other than the Accident and Emergency Department (A&E) (6.91%) as strong

epidemiological links. Admission to the same hospital (particularly hospital A) was common in unrelated cases and considered a weak epidemiological link.

Each case was paired with the individual whose MRSA isolate was the closest genetic match, after which the genetic distance between each MRSA pair was plotted against six different categories of epidemiological contact (Fig. 2). This demonstrated a direct relationship between bacterial relatedness and strength of epidemiological contact.

Evidence of MRSA transmission in the community

Twelve per cent of cases (72/598) with both bacterial and epidemiological links could be resolved into 27 distinct community transmission clusters. MRSA lineages regarded as community-associated (CA-MRSA) - which in the UK include CC1, CC5, CC8, CC45 and CC80 - were associated with 9 separate community transmission clusters (Table 1). However, most community clusters involved hospital-associated lineages (17 separate CC22 clusters involving 50/72 cases (69%), and one CC30 cluster involving 3/72 cases (4%)). To contextualize the MRSA CC22 isolates associated with transmission in the community we constructed a phylogenetic tree containing all CC22 study isolates. This showed that CC22 associated with community clusters were scattered throughout the phylogenetic tree, interspersed with clusters associated with cases with hospital contacts alone (Fig. 3). This indicates that CC22 isolates that are transmitted in the community belong to the wider CC22 population, with no evidence for specific genetic subsets. We also identified transmission clusters relating to three independent GP practices, the largest of which contained 13 cases. All cases with shared postcodes were further investigated to determine whether they shared a residential address. This confirmed that MRSA transmission had occurred in at least eleven separate households (25 cases) and in eight

LTCFs (22 cases) (Table 1). A pictorial representation of exemplars of transmission at a GP practice, LTCF and household is shown in fig. S4, A-C.

Evidence of MRSA transmission in hospitals

More than half of cases with epidemiological and bacterial genomic links (371/598, 62%) resided in transmission clusters with hospital contacts, of which 255 cases had ward contacts. The 371 cases were resolved into 118 different clusters each involving between 2 and 44 individuals (Table 1). We narrowed down further investigation to those clusters that contained five or more patients (9 clusters, see table S3 for details), and evaluated these for instances of direct ward contact (same ward, overlapping admission dates) or indirect ward contact (same ward, no overlap in admission dates). Where available, the presence of a negative MRSA culture followed by a positive MRSA culture was interpreted as additional evidence of hospital acquisition. The specific ward where MRSA had been putatively acquired could be determined in 3 of the 9 clusters, one of which is depicted in Fig 4A. This ward-centric pattern occurred in two different hospitals and across different clonal complexes (CC22, CC30 and CC15). Of note, we observed that there was a time delay between presumptive acquisition date and first clinical detection of MRSA-positivity in most cases (6/8, 3/4 and 3/5 patients). For the remaining six hospital clusters, multiple wards in the same hospital were plausible places of acquisition. We also observed a pattern of transmission that centered around specific individuals, in which the movement of a single, persistently MRSA positive index patient through multiple wards resulted in MRSA acquisition by numerous others patients. This patient-centric pattern of transmission was identified in three transmission clusters (Fig. 4B, Fig S2 E & F) and was observed in two different hospitals and for two clonal complexes (CC22 and CC30). Acquisition by other cases was associated with a high rate of indirect ward acquisition.

MRSA transmission at the hospital-community interface

We identified 28 clusters (157 cases) that contained a mixture of people with community and hospital epidemiological links (Table 1). Further analysis of 15 clusters that contained five or more cases (detailed in table S3) revealed instances of community-onset transmission followed by onward nosocomial dissemination, and hospital-onset transmission followed by nosocomial and community spread in CC30 and CC22 clusters. A pictorial representation of exemplars of these transmission patterns is shown in fig. S4, D-F.

Discussion

Our findings have important implications for infection control policy and practice. MRSA transmission in our study population was not attributable to large nosocomial outbreaks, but resulted from the cumulative effect of numerous clinically unrecognized episodes. We detected 173 separate genetic clusters that mapped to numerous different locations over the course of 12 months, which is indicative of repeated lapses in infection control. There are several explanations for extensive unrecognized transmission, including lack of hospital discharge swabbing, and the fact that place of acquisition is often different to the place of detection and separated by a period of days, weeks or months. This indicates the need for outbreak investigations to widen their scope in time and place when considering potential MRSA contacts.

Standard infection control practice centered on a ward-based approach may also fail to detect the impact of longitudinal patient-centric transmission. We identified a critical role for some persistent carriers who spread MRSA in multiple wards during complex healthcare pathways. This frequently involved indirect transmission, in which apparent acquisition by a new case

occurred after the index case had left the ward, which is suggesting of environmental contamination or colonized healthcare workers. Further studies are needed to identify host factors responsible for persistent carriage associated with a high risk of MRSA transmission, to facilitate risk stratification and targeted allocation of isolation facilities where these are a limited resource.

It is generally accepted that the majority of MRSA lineages have either become adapted to persist and spread in hospitals, or are sufficiently fit to compete with other *S. aureus* lineages associated with community-associated carriage *(12)*. CC22 is the predominant healthcare-associated lineage MRSA lineage in the UK (~70%) followed in frequency by CC30, and most on-going MRSA transmission is assumed to occur in healthcare settings. We expected that most clusters caused by CC22 and CC30 MRSA would map to hospitals, but instead found considerable CC22 transmission in the community. Furthermore, clusters associated with community transmission of MRSA CC22 were distributed across the CC22 phylogeny and were interspersed with hospital-related clusters. This provides definitive evidence for the spread of so-called hospital-associated lineages such as CC22 through transmission networks that include the community. The repeated introduction of MRSA from the community into hospitals and vice versa signals the need for more robust action to detect and tackle community-associated carriage.

By including patient epidemiological information, we found that residential postcodes and GP registration information were strong epidemiological markers of MRSA transmission. Sharing the same postcode and/or GP practice by two or more MRSA-positive patients often indicated an outbreak, some of which spanned over several months. Our findings support the routine collection of postcodes and GP registration as an integral part of routine surveillance to capture

putative MRSA outbreaks in the community. This could guide a targeted approach to the use of WGS to confirm or refute transmission and direct infection control interventions that curtails further dissemination.

We acknowledge several limitations of this study. The study design did not include longitudinal or discharge MRSA screening in hospitals, or screening of environmental reservoirs and healthcare workers. Furthermore, sampling of the community was opportunistic and relied on samples submitted to the diagnostic microbiology. We acknowledge that this would mean failure to detect some MRSA carriers involved in our transmission clusters, and that undetected carriers results in incomplete transmission routes being reconstructed. Non-sampled carriers explain why the MRSA isolate from 680 cases was not linked to the MRSA from any other case, and why 193 cases whose isolate resided in a genetic cluster had no identifiable epidemiological contact. Despite detecting multiple transmission clusters, we are also likely to have underestimated the full extent of MRSA transmission attributable to nosocomial and community sources because of under-sampling of the entire population served by the diagnostic laboratory at CUH.

In conclusion, we provide evidence for the value of integrated epidemiological and genomic surveillance of a population that access the same healthcare referral network in a substantial geographic region of England. The large number of patients screened here allowed us to sample MRSA lineages that are not dominant in the UK but are endemic in other areas of the world including USA300 (prevalent in the United States) *(13)*, the European CA-MRSA CC80 *(14)*, and the Taiwanese CC59 clone (prevalent in Asia) *(15)*. The identification of transmission clusters involving these lineages in hospitals, the community and at the hospital-community interface suggest that our findings are likely to be more widely applicable.

Materials and Methods

Experimental design

We conducted a 12-month prospective observational cohort study between April 2012 and April 2013 to identify consecutive individuals with MRSA-positive samples processed by the Clinical Microbiology and Public Health Laboratory at the Cambridge University Hospitals NHS Foundation Trust (CUH). This facility received samples from three hospitals (referred to as A, B and C) and 75 GP practices in the East of England. All hospital in-patients were routinely screened for MRSA on admission to hospital, and screening was repeated weekly in critical care units. Compliance with mandatory admission screening at the three study hospitals was 85-90% (personal communication, Dr. Nicholas Brown, CUH). Additional clinical specimens were taken as part of routine clinical care. In the community, there was no formal MRSA screening and specimens were taken by GPs or community nursing teams for clinical purposes, meaning that coverage was not complete. Epidemiological data (including hospital ward stays and residential postcodes) were recorded for all MRSA-positive cases. Detailed methodology is provided in Supplementary Methods, and a flowchart summarizing the data types and analyses undertaken is shown in fig. S2. The study protocol was approved by the National Research Ethics Service (ref: 11/EE/0499), the National Information Governance Board Ethics and Confidentiality Committee (ref: ECC 8-05(h)/2011), and the Cambridge University Hospitals NHS Foundation Trust Research and Development Department (ref: A092428).

DNA sequencing and genomic analyses

A total of 3,053 MRSA isolates were collected during the study, of which 2,320 were selected for WGS. A detailed description of the rationale for selecting isolates for sequencing and

genomic methodologies is provided in Supplementary Methods. In brief, DNA was extracted, libraries prepared and 100-bp paired end sequences determined for 2,320 isolates on an Illumina HiSeq2000, as previously described *(11)*. Of these, 2,282 were further analysed after passing quality control (see Supplementary Methods). Genomes were de novo assembled using Velvet *(16)*. STs were derived from assemblies and CCs assigned. All isolates assigned to the same CC were mapped using SMALT (http://www.sanger.ac.uk/science/tools/smalt-0) to the most closely related reference genome. SNPs were identified from BAM files using SAMTOOLS *(17)*. SNPs at regions annotated as mobile genetic elements were removed from whole-genome alignments and maximum likelihood trees created using RAxML *(18)* for each CC. Pairwise genetic distances between isolates of the same CC were calculated based on the number SNPs in the core genome. Sequence data were submitted to the European Nucleotide Archive (www.ebi.ac.uk/ena) under the accession numbers listed in Supplementary Data 1.

Epidemiological analysis

We established epidemiological links between each pair of MRSA-positive individuals (termed case-pairs) through a systematic comparison. Hospital contacts were categorized as follows: *direct ward contact* if a case-pair was admitted to the same ward with overlapping dates of admission; *indirect ward contact* if admitted to the same ward with no overlapping dates; *direct hospital-wide contact* if admitted to the same hospital in different wards with overlapping dates, and *indirect hospital-wide contact* if admitted to the same hospital in different wards with no overlapping dates. We identified episodes of hospital admission for each case in the 12-month period prior to their first MRSA-positive sample. Information on outpatient clinic appointments was not available. *Community contact* was classified if cases shared a postcode or had their MRSA-positive sample submitted by the same GP practice. Community contacts were further

categorized as *household contact* if people shared a residential address; *LTCF contact* if they lived in the same LTCF; or *GP contact* if they were registered with the same GP practice. Information on GP visits were not available other than that recorded for cases with MRSA swabs collected at GP practices. In a few instances, cases shared the same postcode but lived at a different residential address. In a minority of cases, patient addresses could not be retrieved from clinical records and were classified as 'unresolved'. We studied cases positive for MRSA CC22 to determine the frequency of different types of epidemiological contact among genetically unrelated cases, using a pairwise SNP distance greater than 50 SNPs. This analysis led us to consider epidemiological links as strong if they were ward contacts (other than A&E, GP contacts or shared postcodes, and weak if they were hospital-wide contacts and A&E (see Supplementary Methods for details).

Identification of putative MRSA transmission

Selecting a SNP cut-off to define MRSA transmission clusters was informed by two independent lines of evidence. First, we established the genetic diversity of the same MRSA clone in a single individual (pool of diversity) in 26 cases with more than one isolate (range 2 to 3, median 2) from independent samples cultured on the same day. The maximum genetic distance of MRSA in each case ranged from 0 to 41 SNPs (median 2, IQR 1 to 3), which is comparable to the maximum within-host diversity reported elsewhere *(19–21)*. In parallel, we selected the single largest phylogenetic cluster containing isolates from cases with strong epidemiological links (13 cases, a putative outbreak) and established that the pairwise genetic distance between cases ranged from 0 to 48 SNPs. We constructed CC-based phylogenetic trees and then sub-divided each tree into clusters based on a SNP distance of no more than 50, and looked for hospital and community contacts between cases residing in the same genetic cluster. Clusters were

categorized as containing community contacts alone; hospital contacts alone; community AND hospital contacts; or no known hospital/community contacts. For clusters with hospital and/or community contacts involving five or more cases, we incorporated individual patient movement data (for in-patients), sampling dates, MRSA screen results and bacterial phylogeny to identify the most plausible MRSA source. Supplementary Materials and Methods and figs. S2 and S3 describe in more detail how genomic and epidemiological data were integrated to identify and classify transmission clusters.

Supplementary Materials

Materials and Methods

Fig. S1. Number of isolates sequenced per patient

Fig. S2. Flowchart summarizing data types and analyses

Fig. S3. Integration of genomic and epidemiological data to identify transmission clusters

Fig. S4. Six examples of transmission clusters in different settings

Fig. S5. Number of heterozygous sites in the core genome per isolate

Fig. S6. Within-host diversity over time and at a single time point

Table S1. Proportion of isolates in different clonal complexes

Table S2. Frequency of epidemiological contacts among genetically unrelated cases

Table S3. Epidemiological classification of transmission clusters containing 5 or more cases

Data file S1. SupplementaryData1.xlsx

References:

1. F. D. Lowy, Staphylococcus aureus Infections, *N Engl J Med* **339**, 520–532 (1998).
2. L. K. Yaw, J. O. Robinson, K. M. Ho, A comparison of long-term outcomes after meticillin-

resistant and meticillin-sensitive Staphylococcus aureus bacteraemia: an observational cohort study, *Lancet Infect Dis* **14**, 967–975 (2014).

3. M. C. J. Bootsma, O. Diekmann, M. J. M. Bonten, Controlling methicillin-resistant Staphylococcus aureus: Quantifying the effects of interventions and rapid diagnostic testing, *Proc Natl Acad Sci USA* **103**, 5620–5625 (2006).

4. C. U. Köser, M. T. G. Holden, M. J. Ellington, E. J. P. Cartwright, N. M. Brown, A. L. Ogilvy-Stuart, L. Y. Hsu, C. Chewapreecha, N. J. Croucher, S. R. Harris, M. Sanders, M. C. Enright, G. Dougan, S. D. Bentley, J. Parkhill, L. J. Fraser, J. R. Betley, O. B. Schulz-Trieglaff, G. P. Smith, S. J. Peacock, Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak, *N Engl J Med* **366**, 2267–2275 (2012).

5. S. R. Harris, E. J. Cartwright, M. E. Török, M. T. Holden, N. M. Brown, A. L. Ogilvy-Stuart, M. J. Ellington, M. a Quail, S. D. Bentley, J. Parkhill, S. J. Peacock, Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study, *Lancet Infect Dis* **13**, 130–136 (2013).

6. L. Senn, O. Clerc, G. Zanetti, P. Basset, N. C. Gordon, A. E. Sheppard, D. W. Crook, R. James, H. A. Thorpe, E. J. Feil, S. Blanc, The Stealthy Superbug: the Role of Asymptomatic Enteric Carriage in Maintaining a Long-Term Hospital Outbreak of ST228 Methicillin-Resistant Staphylococcus aureus, *mBio* **7**, 1–9 (2016).

7. U. Nübel, M. Nachtnebel, G. Falkenhorst, J. Benzler, J. Hecht, M. Kube, F. Bröcker, K. Moelling, C. Bührer, P. Gastmeier, B. Piening, M. Behnke, M. Dehnert, F. Layer, W. Witte, T. Eckmanns, MRSA Transmission on a Neonatal Intensive Care Unit: Epidemiological and Genome-Based Phylogenetic Analyses, *PLoS ONE* **8** (2013), doi:10.1371/journal.pone.0054898.

8. S. Long, S. Beres, R. Olsen, J. Musser, Absence of Patient-to-Patient Intrahospital Transmission of Staphylococcus aureus as Determined by Whole-Genome Sequencing, *mBio* **5**, 1–10 (2014).

9. J. R. Price, T. Golubchik, K. Cole, D. J. Wilson, D. W. Crook, G. E. Thwaites, R. Bowden, a. S. Walker, T. E. a Peto, J. Paul, M. J. Llewelyn, Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of staphylococcus aureus in an intensive care unit, *Clin Infect Dis* **58**, 609–618 (2014).

10. S. Y. C. Tong, M. T. G. Holden, E. K. Nickerson, B. S. Cooper, C. U. Köser, A. Cori, T. Jombart, S. Cauchemez, C. Fraser, V. Wuthiekanun, J. Thaipadungpanit, M. Hongsuwan, N. P. Day, D. Limmathurotsakul, J. Parkhill, S. J. Peacock, Genome sequencing defines phylogeny and spread of methicillin-resistant Staphylococcus aureus in a high transmission setting, *Genome Res* **25**, 111–118 (2015).

11. S. Reuter, M. E. Török, M. T. Holden, R. Reynolds, K. E. Raven, B. Blane, T. Donker, S. D. Bentley, D. M. Aanensen, H. Grundmann, E. J. Feil, B. G. Spratt, J. Parkhill, S. J. Peacock, Building a genomic framework for prospective MRSA surveillance in the United Kingdom and the Republic of Ireland, *Genome Res* **26**, 263–270 (2016).

12. J. Knox, A.-C. Uhlemann, F. D. Lowy, Staphylococcus aureus infections: transmission within households and the community, *Trends Microbiol* **23**, 437–444 (2015).

13. M. S. Toleman, S. Reuter, F. Coll, E. M. Harrison, B. Blane, N. M. Brown, M. E. Török, J. Parkhill, S. J. Peacock, Systematic Surveillance Detects Multiple Silent Introductions and

Household Transmission of Methicillin-Resistant Staphylococcus aureus USA300 in the East of England, *J Infect Dis* **214**, 447–453 (2016).

14. M. Stegger, T. Wirth, P. S. Andersen, M. Stegger, T. Wirth, P. S. Andersen, R. L. Skov, A. De Grassi, M. Simões, A. Tristan, Origin and Evolution of European Methicillin-Resistant Staphylococcus aureus Origin and Evolution of European Community-Acquired Methicillin-Resistant Staphylococcus aureus, *mBio* **5**, 1–12 (2014).

15. M. J. Ward, M. Goncheva, E. Richardson, P. R. McAdam, E. Raftis, A. Kearns, R. S. Daum, M. Z. David, T. L. Lauderdale, G. F. Edwards, G. R. Nimmo, G. W. Coombs, X. Huijsdens, M. E. J. Woolhouse, J. R. Fitzgerald, Identification of source and sink populations for the emergence and global spread of the East-Asia clone of community-associated MRSA, *Genome Biol* **17**, 160 (2016).

16. D. R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs., *Genome Res* **18**, 821–9 (2008).

17. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools, *Bioinformatics* **25**, 2078–2079 (2009).

18. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies., *Bioinformatics* **30**, 1312–3 (2014).

19. T. Golubchik, E. M. Batty, R. R. Miller, H. Farr, B. C. Young, H. Larner-Svensson, R. Fung, H. Godwin, K. Knox, A. Votintseva, R. G. Everitt, T. Street, M. Cule, C. L. C. Ip, X. Didelot, T. E. A. Peto, R. M. Harding, D. J. Wilson, D. W. Crook, R. Bowden, Within-Host Evolution of Staphylococcus aureus during Asymptomatic Carriage, *PLoS ONE* **8**, 1–14 (2013).

20. O. C. Stine, S. Burrowes, S. David, J. K. Johnson, M. Roghmann, Transmission Clusters of Methicillin-Resistant Staphylococcus Aureus in Long-Term Care Facilities Based on Whole-Genome Sequencing., *Infect Control Hosp Epidemiol* **37**, 1–7 (2016).

21. G. K. Paterson, E. M. Harrison, G. G. R. Murray, J. J. Welch, J. H. Warland, M. T. G. Holden, F. J. E. Morgan, X. Ba, G. Koop, S. R. Harris, D. J. Maskell, S. J. Peacock, M. E. Herrtage, J. Parkhill, M. a. Holmes, Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission, *Nat Commun* **6**, 6560 (2015).

22. M. Boetzer, C. V Henkel, H. J. Jansen, D. Butler, W. Pirovano, Scaffolding pre-assembled contigs using SSPACE., *Bioinformatics* **27**, 578–9 (2011).

23. M. Boetzer, W. Pirovano, Toward almost closed genomes with GapFiller., *Genome Biol* **13**, R56 (2012).

24. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows–Wheeler transform, *Bioinformatics* **26**, 589–595 (2010).

25. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome., *Genome Biol* **10**, R25 (2009).

26. M. C. F. Prosperi, M. Ciccozzi, I. Fanti, F. Saladini, M. Pecorari, V. Borghi, S. Di Giambenedetto, B. Bruzzone, A. Capetti, A. Vivarelli, S. Rusconi, M. C. Re, M. R. Gismondo, L. Sighinolfi, R. R. Gray, M. Salemi, M. Zazzi, A. De Luca, A novel methodology for large-scale phylogeny partition., *Nat Commun* **2**, 321 (2011).

Acknowledgments: We thank Hayley Brodrick, Kim Judge, Harriet Giramahoro, and Marie Blackman-Northwood for technical assistance, Lois Mlemba for clinical data collection, the Wellcome Trust Sanger Institute Core Sequencing and Pathogen Informatics Groups, and David Harris for assisting in submitting sequence data to public databases.

Funding: Supported by grants from the UKCRC Translational Infection Research (TIR) Initiative, and the Medical Research Council (Grant Number G1000803) with contributions to the Grant from the Biotechnology and Biological Sciences Research Council, the National Institute for Health Research on behalf of the Department of Health, and the Chief Scientist Office of the Scottish Government Health Directorate (to Prof. Peacock); a Hospital Infection Society Major Research Grant; by Wellcome Trust grant number 098051 awarded to the Wellcome Trust Sanger Institute, and by Wellcome Trust 201344/Z/16/Z awarded to Francesc Coll. M.S.T. is a Wellcome Trust Clinical PhD fellow. M.E.T. is a Clinician Scientist Fellow, supported by the Academy of Medical Sciences and the Health Foundation and by the NIHR Cambridge Biomedical Research Centre.

Author contributions: M. E. T. and S. J. P. designed the study, wrote the study protocol and case record forms, obtained ethical and research and development approvals for the study, and supervised the data collection. N.M.B., R.M.K and B.P. were responsible for isolating and identifying MRSA in the diagnostic microbiology laboratory, and provided expert opinion relating to infection control. F.C. undertook the epidemiological and bioinformatic analyses with contributions from E.M.H., M.S.T. and S. R.. B.B. and K.E.R. conducted the laboratory work. J.P. supervised the genomic sequencing. F.C. and S.J.P. wrote the first draft of the manuscript. S.J.P. supervised and managed the study. All authors had access to the data and read, contributed and approved the final manuscript.

Competing interests: N.M.B. is on the advisory board for Discuva Ltd. All other authors declare that they have no conflicts of interest.

Data and materials availability: The whole genome sequences from this study have been deposited at European Nucleotide Archive under accession numbers listed in Supplementary Data 1.

Figures

Fig. 1. Map showing the study catchment area in the East of England.

The locations of hospitals (n=3), GP practices (n=75) and postcode districts are shown for the 1,465 study cases. Postcode districts are color-coded to show the number of MRSA positive cases sampled in each district. A total of 5,012,137 residents lived in the highlighted districts (16,240 km^2) according to the 2011 UK Census.

Fig. 2. Pairwise comparison between MRSA relatedness and type of patient contact.

For each case, the most closely related MRSA isolate from another case was identified and the epidemiological contact of each case-pair defined. The number of cases in each epidemiological category is shown as a function of the genetic distance. Panels A to D show the genetic distance distribution for cases with hospital contacts alone. Direct contact refers to a link in the same time and place (ward or hospital). Indirect contact refers to a link in the same place but different time. Panel E shows community contacts. Cases with neither hospital nor community contacts are shown in panel F. Only cases with MRSA isolates from clonal complexes found in at least one other patient in the population are shown (n=1,459).

Fig. 3. Transmission clusters color coded on the CC22 phylogeny.

Maximum likelihood tree generated from 34,600 SNP sites in the core genome is shown for 1,667 CC22 isolates. Colors refer to type of epidemiological links in clusters of genetically related isolates (maximum 50 SNPs) from multiple cases.

Fig. 4. Exemplars of two patterns of nosocomial MRSA spread

A. Ward-centric. Eight patients in this transmission cluster had ward contacts in ward B2 and B21, including admission overlaps. Of note, the putative epicentre of transmission was in wards B2 or B21 but the outbreak strain was isolated on later admissions in 6 of the 8 patients, 3 of which (1090, 727 and 762) were first detected at a different hospital (hospital A) from where they had putatively acquired this (hospital B).

B. Patient-centric. Six patients had stayed in wards visited by patient 388 (i.e. A49, A80 and A59) prior to their MRSA isolation date. Negative MRSA screens prior to entry to these wards for some patients (1288, 1057, 1488, 1377 and 942) further supports hospital acquisition. Isolates from patient 388 were the most basal in the phylogenetic tree and their diversity enclosed that of the other patients, providing further indicators for this patient being the potential source for the transmission cluster. Colored blocks other than grey represent ward contacts, which are labelled by a letter to denote the hospital (A or B) and a number that denotes the anonymised ward.

Table 1. Epidemiological classification of transmission clusters.

Columns are ordered based on decreasing proportion of isolates in each CC. Each cell shows the number of cases and (in brackets) the number of transmission clusters to which these cases were assigned. The number of transmission clusters in each category is the sum of those of its sub-categories. The same applies to the number of cases except for column 'CC22' and 'Overall'. A total of 7 cases had two different CC22 strains suggestive of mixed colonisation or strain replacement that linked them to two different transmission clusters. This explains why the total number of genetically clustered cases (n=578) is lower than the sum of cases in its sub-categories. CCs with genetically unrelated isolates or isolated once from the study population are not shown. 'Multiple hospitals' refers to epidemiological contacts from more than one of three study hospitals.

Epidemiological classification	Overall	CC22	CC30	CC5	CC1	CC8	CC45	CC59	CC80	CC15	CC361
Genetically unrelated cases	680	462	36	49	35	42	17	15	6	1	2
Genetically clustered with other cases	785	578	46	30	45	9	34	26	9	9	3
Genetically clustered & epidemiological contacts	598 (173)	449 (127)	36 (8)	20 (9)	33 (13)	4 (2)	24 (8)	21 (3)	2 (1)	8 (1)	3 (1)
Only community contacts	72 (27)	50 (17)	3 (1)	3 (1)	6 (3)	4 (2)	4 (2)	-	2 (1)	-	-
Different postcode Shared GP practice	14 (3)	10 (1)	-	-	2 (1)	-	2 (1)	-	-	-	-
Same postcode Shared Household	25 (11)	16 (7)	3 (1)	-	-	4 (2)	-	-	2 (1)	-	-
Same postcode Shared LTCF	22 (8)	20 (7)	-	-	-	-	2 (1)	-	-	-	-
Same postcode Different addresses	2 (1)	-	-	-	2 (1)	-	-	-	-	-	-
Same postcode Unresolved	9 (4)	4 (2)	-	3 (1)	2 (1)	-	-	-	-	-	-
Only hospital contacts	371 (118)	296 (91)	10 (3)	15 (7)	20 (8)	-	16 (5)	5 (2)	-	8 (1)	3 (1)
Ward contact	255 (64)	212 (52)	6 (1)	5 (2)	10 (4)	-	9 (2)	3 (1)	-	8 (1)	3 (1)
Hospital A	125 (41)	101 (35)	6 (1)	-	6 (2)	-	9 (2)	-	-	-	3 (1)
Hospital B	48 (14)	32 (10)	-	3 (1)	2 (1)	-	-	3 (1)	-	8 (1)	-
Hospital C	8 (4)	4 (2)	-	2 (1)	2 (1)	-	-	-	-	-	-
Multiple hospitals	75 (5)	75 (5)	-	-	-	-	-	-	-	-	-
Hospital-wide contact	118 (54)	85 (39)	4 (2)	10 (5)	10 (4)	-	7 (3)	2 (1)	-	-	-
Hospital A	97 (45)	70 (33)	2 (1)	8 (4)	8 (3)	-	7 (3)	2 (1)	-	-	-
Hospital B	6 (3)	2 (1)	2 (1)	-	2 (1)	-	-	-	-	-	-
Hospital C	8 (4)	6 (3)	-	2 (1)	-	-	-	-	-	-	-
Multiple hospitals	8 (2)	8 (2)	-	-	-	-	-		-	-	-
Both hospital and community contacts	156 (28)	104 (19)	23 (4)	2 (1)	7 (2)	-	4 (1)	16 (1)	-	-	-
Different postcode Shared GP practice	13 (2)	13 (2)	-	-	-	-	-	-	-	-	
Same postcode Shared Household	37 (9)	17 (3)	11 (3)	2 (1)	3 (1)	-	4 (1)	-	-	-	-
Same postcode Shared LTCF	56 (9)	36 (7)	-	-	4 (1)	-	-	16 (1)	-	-	-
Same postcode Different addresses	17 (3)	5 (2)	12 (1)	-	-	-	-	-	-	-	-
Same postcode Unresolved	33 (5)	33 (5)	-	-	-	-	-	-	-	-	-
Neither hospital nor community contacts	193	134	10	10	12	5	10	5	7	-	-
Total number of cases	1465	1040	82	79	80	51	51	41	15	9	5

Supplementary Materials for

Longitudinal genomic surveillance of MRSA reveals extensive transmission in hospitals and the community

Francesc Coll*, Ewan M. Harrison, Michelle S. Toleman, Sandra Reuter, Kathy E. Raven, Beth Blane, Beverley Palmer, A. Ruth M. Kappeler, Nicholas M. Brown, M Estée Török, Julian Parkhill, Sharon J. Peacock*

*Corresponding author: francesc.coll@lshtm.ac.uk and sharon.peacock@lshtm.ac.uk

This file includes:

Materials and Methods

Case data and sequencing strategy

Patient medical records were reviewed to collect routine epidemiological and microbiological data. This included date and ward of hospital admission and discharge. Information on ward transfers during the same admission were also collected from the hospital bed tracking system, what provided information on the location of methicillin-resistant *Staphylococcus aureus* (MRSA) cases in relation to other MRSA carriers and non-MRSA carriers. Full postcodes were obtained from electronic healthcare records. A full postcode in the United Kingdom (UK) designates a restricted area with a small number of addresses.

Selection of isolates for sequencing was performed as follows. The first MRSA isolate cultured from each case was sequenced. Some cases had more than one MRSA-positive culture during the same admission. If screening swabs and clinical samples were positive, an isolate from each of the two categories were sequenced. If MRSA was available from more than one clinical sample, the isolate from a sterile site sample took preference over isolates from colonized sites. All blood culture isolates were sequenced. Each admission was treated as an independent event and the rule base above reapplied.

Microbiology and DNA sequencing

MRSA was isolated by plating screening swabs onto Brilliance MRSA chromogenic medium (Oxoid, Basingstoke, UK) and from all other samples by plating onto Columbia Blood Agar (Oxoid, Basingstoke, UK). *S. aureus* was identified using a commercial latex agglutination kit (Pastorex Staph Plus, Bio Rad Laboratories, Hemel Hempstead, UK). One colony was randomly selected for downstream processing. Positive cases and their isolates were assigned a unique anonymous identification code. Bacterial DNA was extracted using the QIAxtractor

kit (QIAgen, Hilden, Germany), and anonymised DNA samples were transferred to the Wellcome Trust Sanger Institute for sequencing.

Genomic analyses

Genomes for the 2,282 MRSA isolates were de novo assembled using an iterative process involving Velvet *(1)* version 1.2.09, SSPACE *(2)* v2.0, GapFiller *(3)* v1.11, BWA *(4)* v0.7.12 and Bowtie *(5)* v1.1.0 as explained elsewhere *(11)*.

Thirty-eight isolate genomes were excluded including 6 isolates that were found to be *Staphylococcus haemolyticus* (n=5) or *Staphylococcus sciuri* (n=1); 4 isolates were *S. aureus* but were not MRSA (i.e. *mecA* and *mecC* negative); 8 isolates were duplicates; 6 isolates had missing epidemiological or other information; 2 genomes had a low number of reads (<1,000,000); 8 MRSA genomes had a high number of heterozygous sites (>10,000 see fig. S5) and very fragmented assemblies, suggestive of multiple MRSA clones being sequenced; and 4 genomes had an abnormal position in the phylogeny (suggestive of contamination).

Sequence types (STs) were derived from de novo assemblies by extracting all seven *S. aureus* MLST loci and comparing them to the PubMLST database (www.PubMLST.org). Clonal complexes (CCs) were derived from the allelic profile, allowing up to two allele mismatches from the reference ST. All isolates belonging to the same clonal complex (CC) were mapped using SMALT v0.7.4 to a closely related reference genome, as follows: ST22 strain HO 5096 0412 (accession number HE681097), ST30 strain MRSA252 (BX571856), ST5 strain N315 (BA000018), ST1 strain MW2 (BA000033), ST8 strain USA300 FPR3757 (CP000255), ST45 strain CA-347 (CP006044.1), ST59 strain M013 (CP003166.1) and ST80 strain 11819-97 (CP003194.1). In addition, all isolates were mapped to the ST22 reference genome.

Single-nucleotide polymorphisms (SNPs) were called from BAM files using SAMTOOLS v0.1.19 using default parameters. Heterozygous sites were called if the non-reference allele was represented by 20 to 80% of the reads. Supplementary Data 1 provides a full list of isolates, accession numbers, STs and quality control (QC) metrics.

Whole genome alignments were created for CC22, CC30, CC5, CC1, CC8, CC45, CC59 and CC80 using SNPs derived from mapping of isolates to their corresponding reference genome. Whole genome alignments for CCs containing fewer than 10 cases (CC15, CC130, CC361, CC72, CC12 and CC88) were created with SNPs derived from mapping to the ST22 EMRSA15 strain reference genome, since a genetically close reference genome was not available. Chromosomal regions containing mobile genetic elements were removed from the alignments and maximum likelihood trees created using RAxML v7.8.6 for each of the fourteen CCs. Pairwise genetic distances between isolates of the same CC were calculated based on the number SNPs in the core genome. These analyses were not conducted for STs/CCs which only occurred once (ST97, ST152, CC398, ST425, ST779 and ST1943).

Within-host genetic diversity

Within-host diversity of MRSA was determined for 346 cases who had more than one CC22 isolate sequenced (n=970 isolates). The pairwise SNP distance in the core genome was determined for each pair of isolates from the same case. Figure S6A shows SNP distances binned by time between isolate collection times. Isolates collected within one month had a median genetic distance of 2 SNPs and isolates collected within 9 to 12 months apart had a median of 6 SNPs. We detected variability around these medians and hypothesized that a pre-existing pool of diversity within individuals could be responsible for this. We therefore

evaluated cases having multiple isolates sequenced on the same day (n=26) and calculated the maximum within-host diversity (fig. S6B). The pool of diversity ranged from 0 to 41 SNPs.

Identification and classification of transmission clusters

Phylogenetic trees for each CC were partitioned into clusters using a "depth-first-search" algorithm *(6)*. We used a maximum genetic distance of 50 SNPs at the core genome as a threshold to define clusters based on two criteria: (a) it is greater than the maximum within-host SNP distance found in this study (fig. S6) (i.e. 41 SNPs), which in turn is comparable to that already reported *(7–9)*, and (b) it is greater than the maximum SNP distance observed among isolates of different cases in the largest phylogenetic clade associated with strong epidemiological links (fig. S2A). We then focused on clusters harboring multiple cases and studied epidemiological links – ward, hospital and community contacts – among them. In the context of this study, a transmission cluster was defined as two or more cases whose MRSA isolates were genetically related - forming a monophyletic clade in the tree of a maximum of 50 SNPs – and with epidemiological links supporting transmission. See fig. S2 and S3 for a schematic representation of how genomic and epidemiological data were integrated to identify transmission clusters.

Inference of transmission routes

For transmission clusters involving five or more people, case location was plotted over the time of the study using R. Sample collection times, MRSA screen results and the phylogeny were taken into account to visually identify the most plausible ward/source where MRSA was putatively acquired.

Supplementary Figures

Fig. S1. Number of isolates sequenced per patient

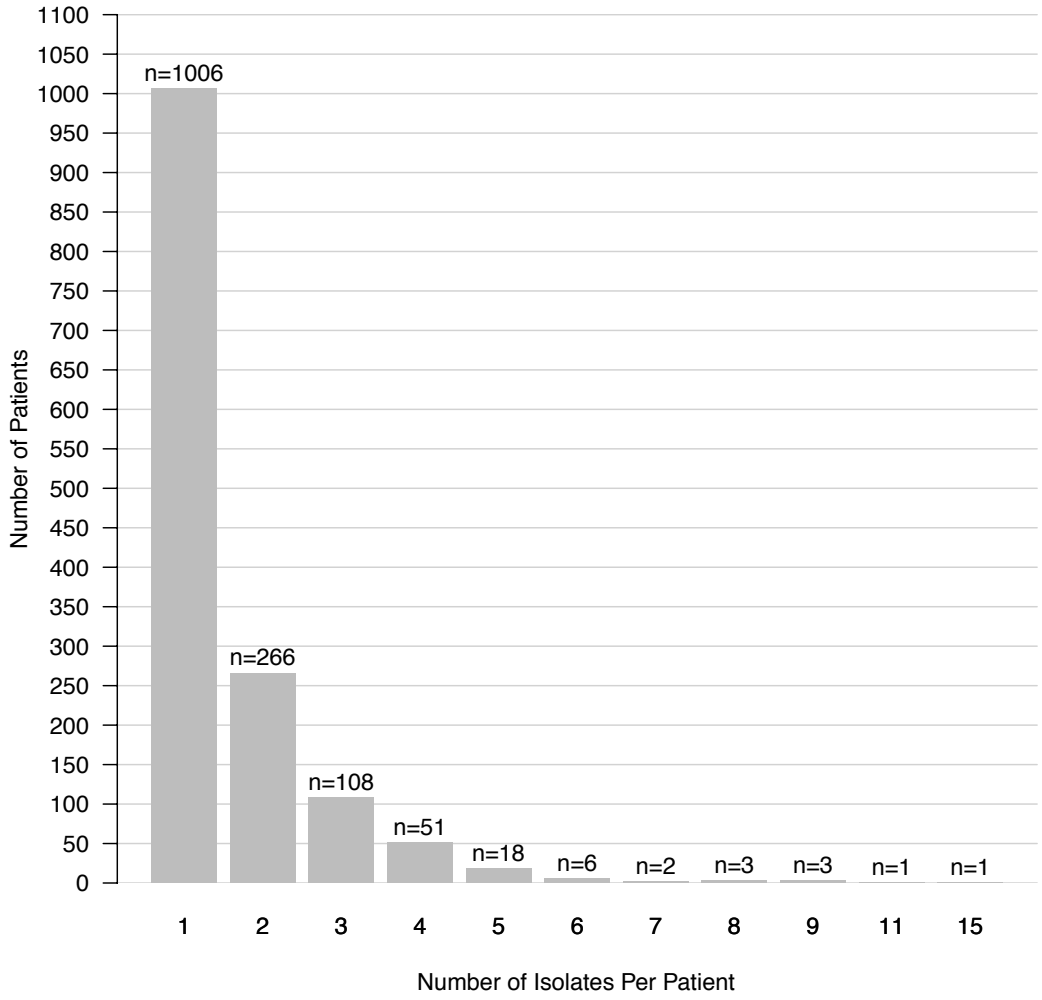
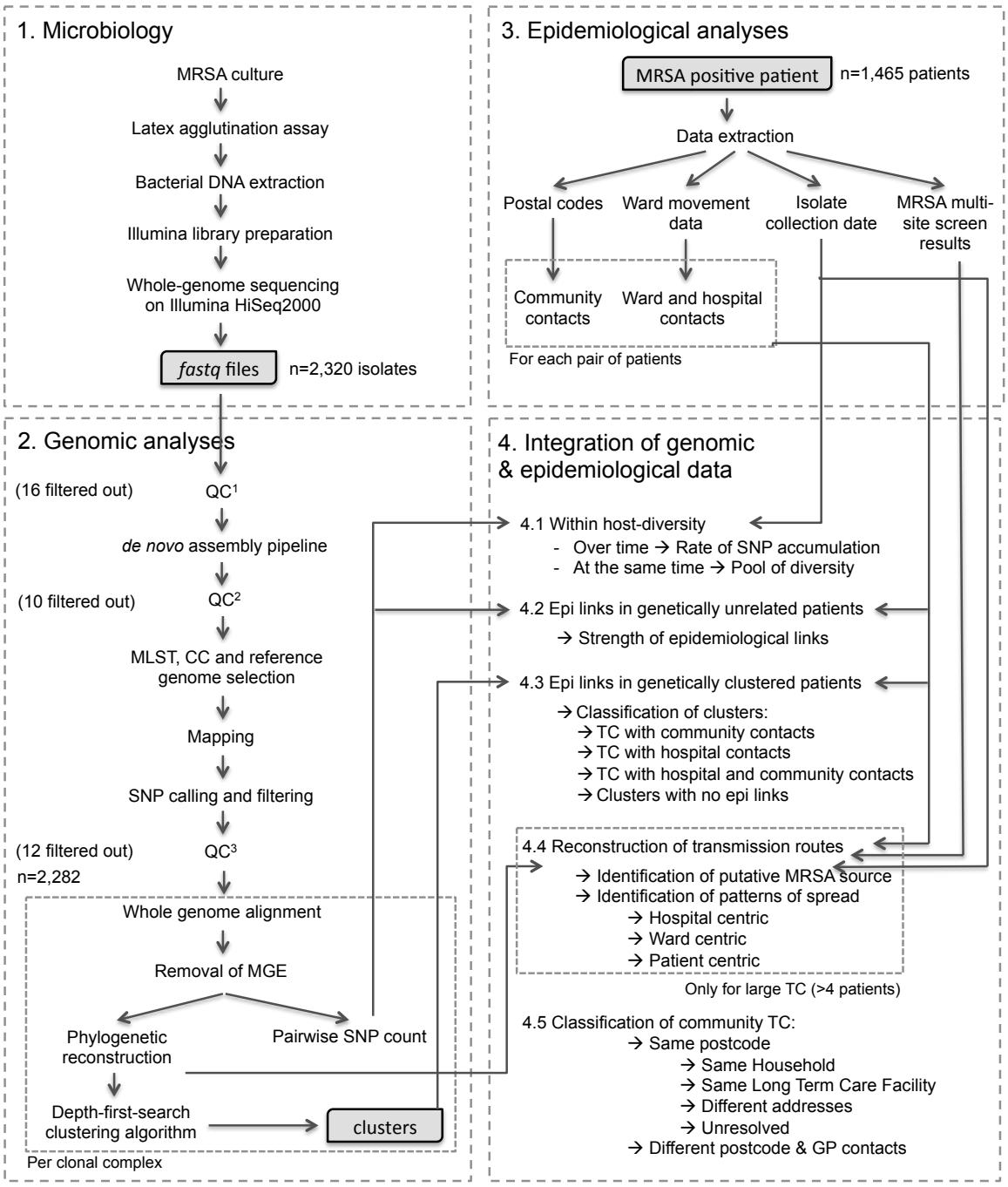


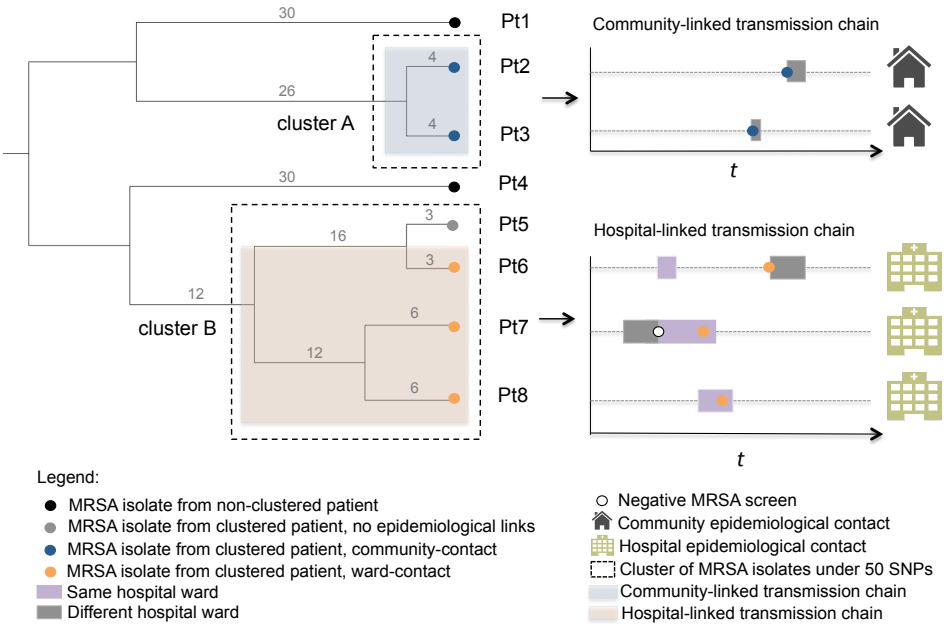
Fig. S2. Flowchart summarizing data types and analyses



QC^1 refers to samples discarded due to low number of reads (n=2) and those that were duplicates or had missing information (n=14), QC^2 to samples that were not MRSA (n=4) or *S. aureus* (n=6) and QC^3 to samples filtered out due to high number of heterozygous sites (n=8, fig. S5) or abnormal position in the phylogeny (n=4). Abbreviations: QC, quality

control; MGE, mobile genetic elements; MLST, multi-locus sequence type; TC, transmission cluster.

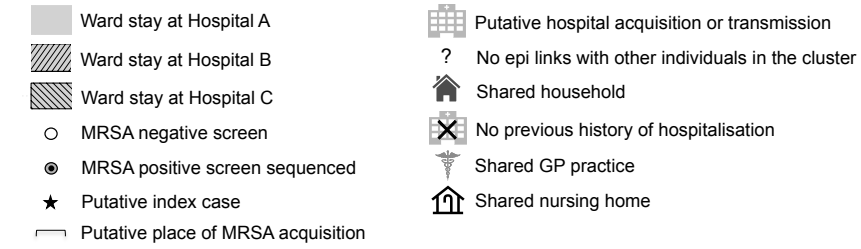
Fig. S3. Integration of genomic and epidemiological data to identify transmission clusters



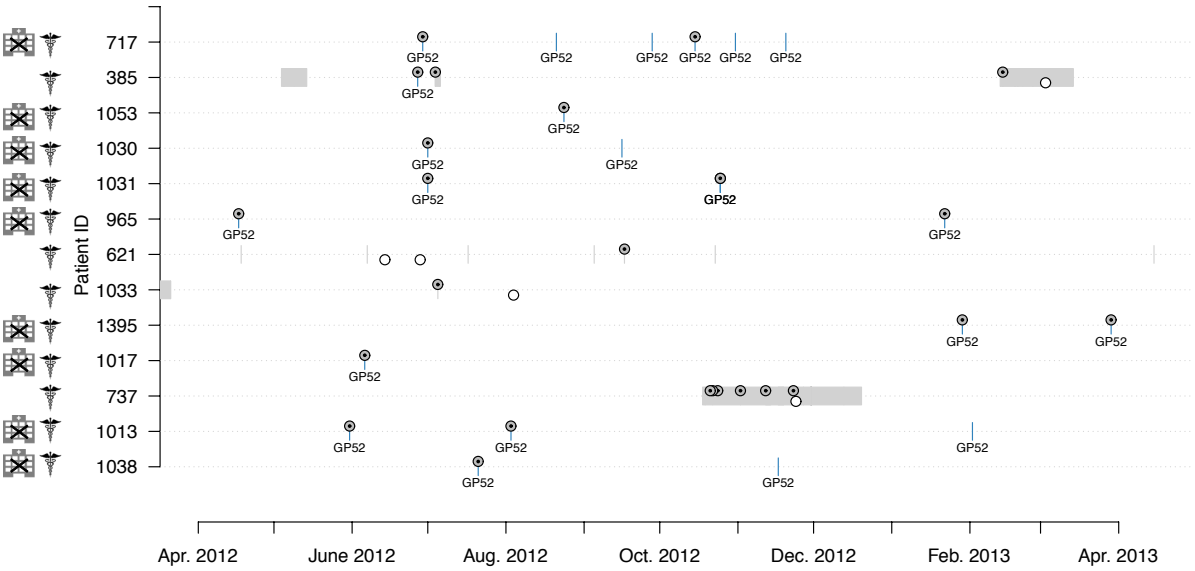
A hypothetical example is shown to illustrate how bacterial phylogeny and patient epidemiological data were used to define and classify transmission clusters. First, phylogenetic trees were partitioned into clusters based on a maximum genetic distance of 50 SNPs in the core genome. Numbers on branches show SNP distances. The diagram shows two genetic clusters that contain MRSA isolates from two and four cases respectively. Pt1 and Pt4 isolates were not clustered and thus classified as 'genetically unrelated'. Epidemiological links between cases within the same cluster were then defined. In cluster A, Pt2 and Pt3 had no previous hospital contact but shared the same residential postcode and was classified as a 'community' cluster. In cluster B, three out of four cases had stayed in the same hospital ward and did not share the same postcode, forming a 'hospital' cluster. In some cases (e.g. Pt7), hospital acquisition was supported by a previous negative MRSA screen. Pt5 had no epidemiological links despite being clustered and was therefore not included in the transmission cluster.

Fig. S4. Six examples of transmission clusters in different settings

LEGEND KEY: Colored blocks other than grey represent ward contacts, which are labelled by a letter to denote hospital or GP (A, B or GP) and a random number that denotes the anonymized ward or GP.

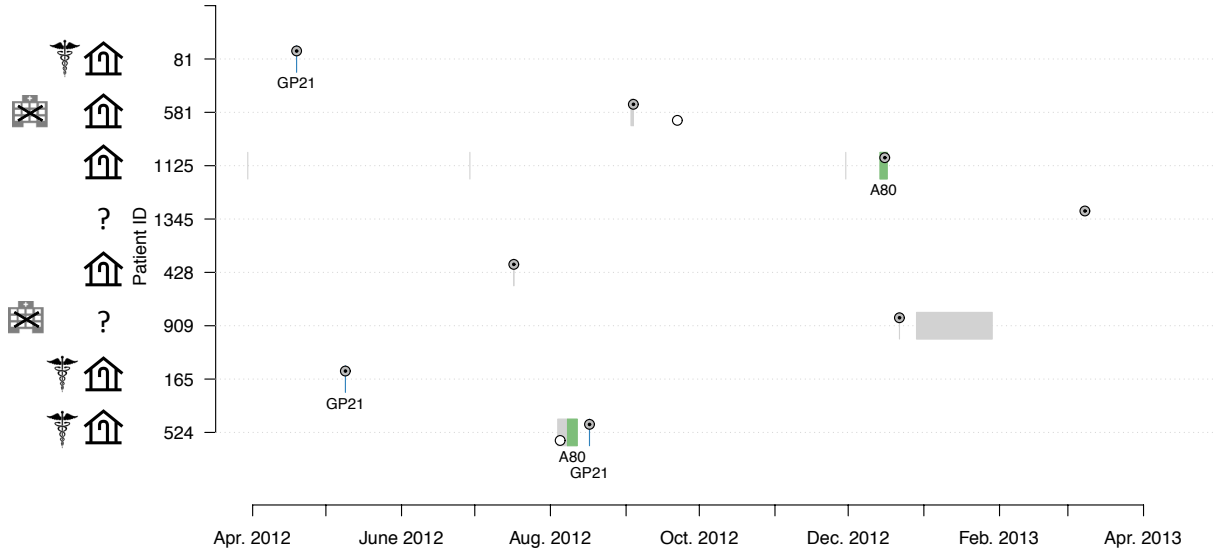


A. Transmission cluster centered around a GP practice (CC22)



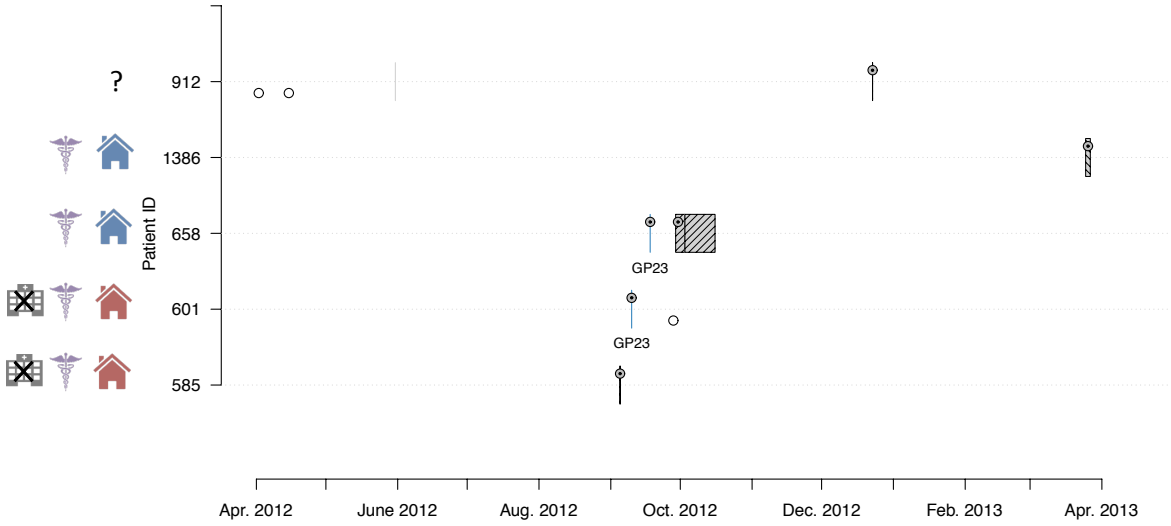
The median bacterial pairwise genetic distance between cases was 16 SNPs (range 0 – 48). 10/13 cases had samples taken at the same GP practice (GP52). 9/13 cases had no recorded hospital admission in the previous year, ruling out recent nosocomial acquisition. All cases had different postcodes except for 621 and 1395. Cases 621, 737 and 1033 all had MRSA samples taken at hospital A but were also registered to GP52.

B. Community transmission cluster at a LTCF (CC22)



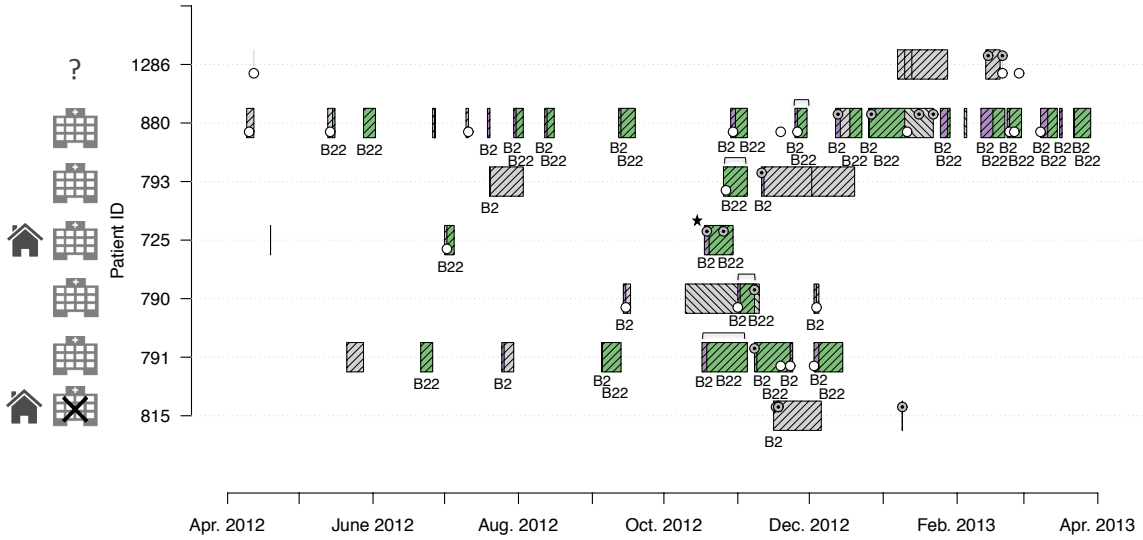
The median bacterial pairwise genetic distance between cases was 27 SNPs (range 5 – 46). Six of 8 cases resided in the same LTCF (excluding 909 and 1345). In three cases (patient 81, 165 and 524), the MRSA-positive sample was submitted by the same GP practice (GP21), while the remainder were taken at hospital A. No epidemiological links were identified for cases 909 and 1345.

C. Community transmission cluster involving households and a GP practice (CC22)



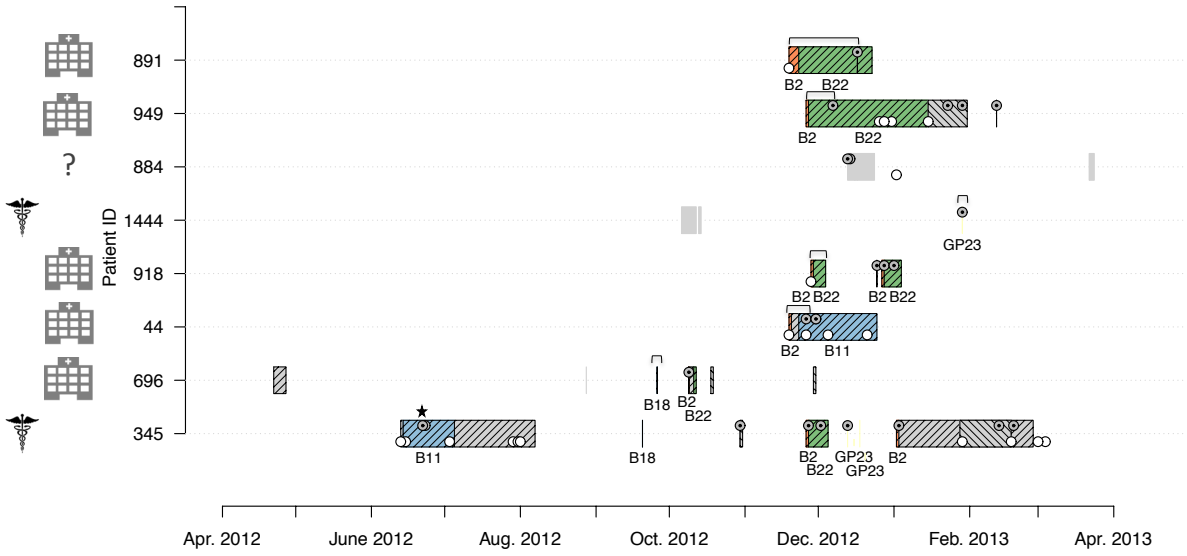
The median bacterial pairwise genetic distance between cases was 47 SNPs (range 5 – 48). 4/5 cases had community epidemiological links. Cases 658 and 1386 shared a residential address, and cases 585 and 601 shared a residential address. Cases 658, 1386, 585 and 601 were all registered with the same GP (GP23). Case 912 has no identifiable epidemiological link to the other cases.

D. Transmission cluster with both community and hospital epidemiological links (CC30)



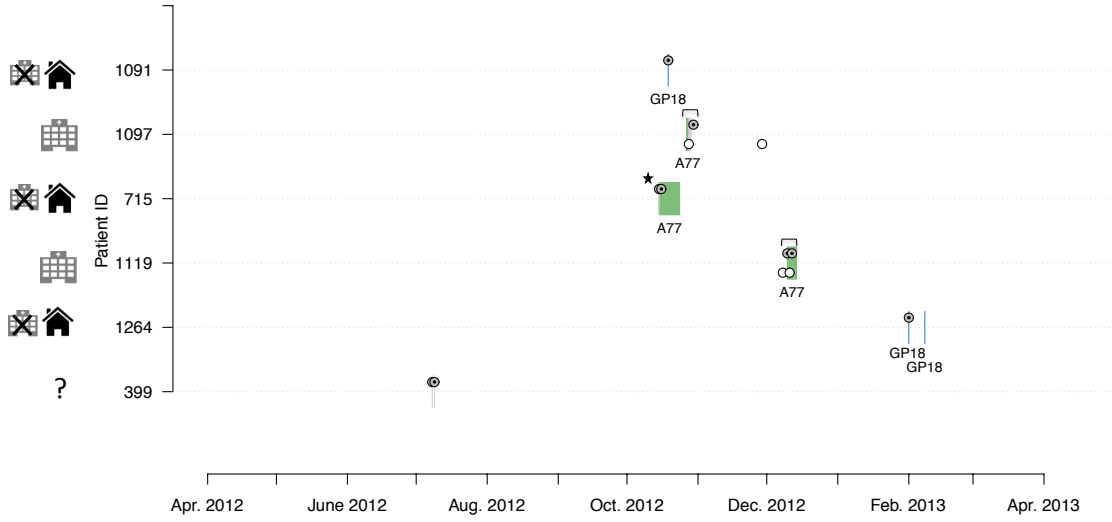
The median bacterial pairwise genetic distance between cases was 2 SNPs (range 0 – 5). 5/7 cases had ward contacts on ward B2 and B22, including admission overlaps in ward B22. Cases 815 and 725 shared a residential address. The transmitted MRSA strain was isolated in all cases after case 725 was admitted into hospital, suggesting that this case may have introduced MRSA into this setting. Case 815 had no recorded hospital admission in the previous year, ruling out recent nosocomial acquisition.

E. Transmission cluster with both community and hospital epidemiological links (CC22)



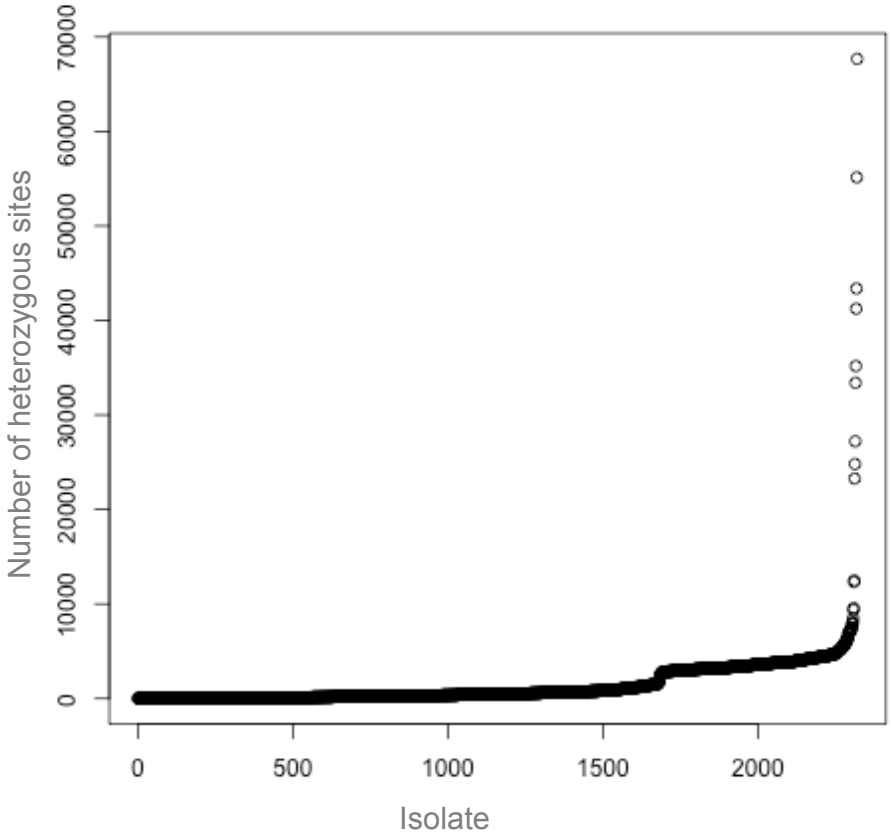
The median bacterial pairwise genetic distance between cases was 3 SNPs (range 0 – 8). Patient 345 appears to be the source in the transmission cluster as their isolates predated others, occupied the most basal position in the phylogenetic tree, and enclosed those from the other seven cases. Patient 345 appears to have acquired MRSA at hospital B as two negative screens preceded the first positive sample. All other cases except 884 are epidemiologically linked to patient 345 including contact on ward B18 (case 696), B2/B22 (cases 918, 949, 891), B11 (case 44) and GP23 (case 1444). Case 345 was colonized by the same MRSA strain for 9 months.

F. Transmission cluster with both community and hospital epidemiological links. Household transmission followed by onward hospital transmission. (CC30)



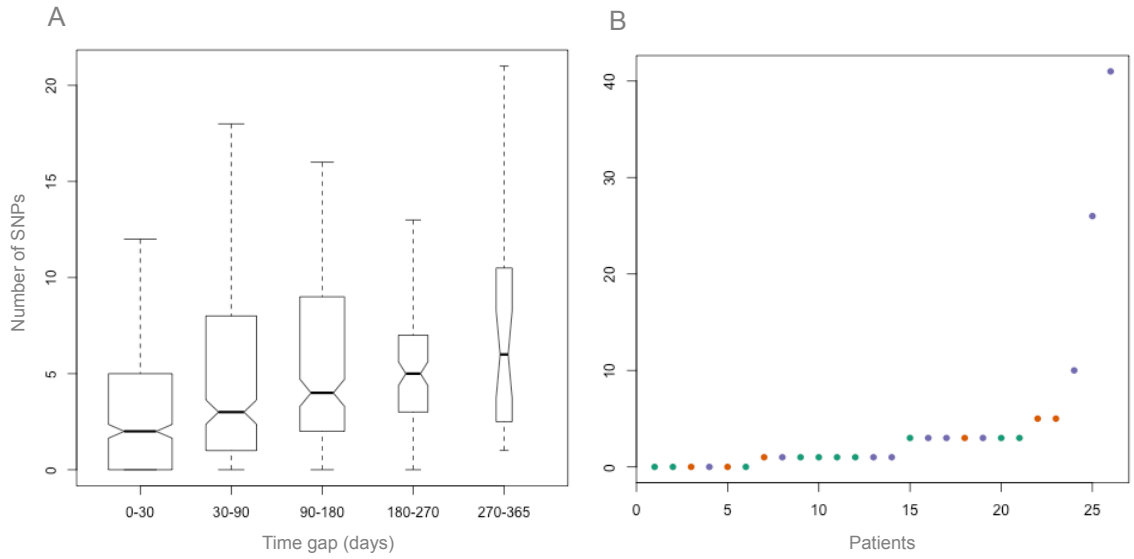
The median bacterial pairwise genetic distance between cases was 3 SNPs (range 0 – 12). Cases 1091, 1264, 715 had no hospital contact in the past year but shared a residential address. Cases 1119 and 1097 were admitted to ward A77 after patient 715 had been discharged (indirect ward contact). MRSA isolates from cases 1119, 1097 and 715 were 0 SNPs apart, supporting recent transmission. Case 399 had no identifiable epidemiological link to the other cases.

Figure S5. Number of heterozygous sites in the core genome per isolate



A total of 11 genomes (3 *Staphylococcus haemolyticus* and 8 *Staphylococcus aureus*) with >10,000 heterozygous sites were assumed to be contaminated and were discarded from further analyses.

Figure S6. Within-host diversity over time and at a single time point



A. Genetic distance between pairs of isolates taken from the same case. A total of 346 cases with 970 CC22 isolates were used in the analysis. On the x-axis, the time gap in days between isolate collection times is shown. On the y-axis, SNP distances between pairs of isolates from the same patient. The width of the box plots reflects the number of data points.

B. Maximum genetic distance observed among MRSA isolates taken from the same patient on the same day. A total of 26 cases with 55 CC22 isolates were used in the analysis (2 to 3 isolates per case). The color of each point indicates the specimen type from which isolates being compared were cultured (green for screen vs. screen, orange for clinical vs. clinical, and purple for screen vs. clinical sample). On the x-axis, case identifiers are shown.

Supplementary Tables

Table S1. Proportion of isolates in different clonal complexes

CC	Number of Isolates	Percentage of Isolates (%)	Number of individuals
22	1667	73.05	1040
30	129	5.65	82
5	108	4.73	79
1	105	4.60	80
8	87	3.81	51
45	70	3.07	51
59	50	2.19	41
80	19	0.83	15
361	11	0.48	5
130	10	0.44	9
15	10	0.44	9
72	5	0.22	3
97	2	0.09	1
12	2	0.09	2
88	2	0.09	2
152	1	0.04	1
1943	1	0.04	1
398	1	0.04	1
425	1	0.04	1
779	1	0.04	1
Total	2282	100	1465*

*A total of 9 cases had isolates from two different CCs. In one case three different CCs were sequenced.

Table S2. Frequency of epidemiological contacts among genetically unrelated cases

Hospital-wide contacts	
Both cases not admitted to the same hospital	313,269 (58.61%)
Both cases admitted to Hospital A	195,222 (36.2%)
Both cases admitted to Hospital B	14,295 (2.67%)
Both cases admitted to Hospital C	11,631 (2.18%)
Ward or GP contacts	
Both cases did not visit the same ward or GP practice	453,753 (84.90%)
Both cases seen as a patient in A&E	37,340 (6.99%)
Both cases admitted to the same ward or GP	2,933 or less (0.55% or less)
Postcode contacts	
Postcodes not available for comparison	64,898 (12.14%)
Both cases had different postcodes	469,309 (87.82%)
Both cases had the same postcode	210 (0.04%)

A total of 1,040 cases with CC22 isolates were available, resulting in 540,280 case-pair comparisons. Of these, 534,417 case pairs had genetically unrelated MRSA isolates, i.e. more than 50 SNPs apart. The table shows the frequency of three different types of epidemiological contacts among these 534,417 comparisons.

Table S3. Epidemiological classification of transmission clusters containing 5 or more cases

CC	TC	Cases	Classification	Transmission Pattern	Figure
CC22	285-B	44	Hospital	Hospital-centric (hospital A)	Not Shown
CC22	285-A	21	Hospital	Ward-centric (hospital B)	Figure 2A
CC22	194-A	10	Hospital	Hospital-centric (hospital A)	Not Shown
CC22	248-A	8	Hospital	Hospital-centric (hospital B)	Not Shown
CC15	1-A	8	Hospital	Ward-centric (hospital B)	Not Shown
CC30	14-A	6	Hospital	Ward-centric (hospital A)	Not Shown
CC22	53-A	6	Hospital	Hospital-centric (hospital A)	Not Shown
CC45	6-A	6	Hospital	Hospital-centric (hospital A)	Not Shown
CC22	94-A	5	Hospital	Hospital-centric (hospital A)	Not Shown
CC22	144-A	13	Community	Centred at GP practice	Fig S2 A
CC22	242-A	6	Community	Centred at LTCF	Fig S2 B
CC59	4-A	16	Hospital/Community	Centred at two LTCF + hospital wide contacts in hospital A	Not shown
CC30	4-A	12	Hospital/Community	Centred at household + ward contacts in hospital B	Fig S2 D
CC22	241-A	11	Hospital/Community	Centred at LTCF + ward contacts in hospital A	Not Shown
CC22	227-A	11	Hospital/Community	Unresolved community contact + ward contacts in hospital A	Not Shown
CC22	113-A	8	Hospital/Community	Centred at two households + hospital-wide contacts in hospital A	Not Shown
CC22	237-A	7	Hospital/Community	Centred at GP + patient centric in hospital B	Fig S2 E
CC22	106-A	7	Hospital/Community	Unresolved community contact + ward contacts, patient centric in hospital A	Figure 2B
CC22	74-A	7	Hospital/Community	Unresolved community contact + ward contacts in hospital B	Not Shown
CC22	138-A	6	Hospital/Community	Centred at GP practice + ward contacts in hospital A	Not Shown
CC22	79-A	5	Hospital/Community	Unresolved community contact + ward contacts in hospital B	Not Shown
CC22	105-A	5	Hospital/Community	Centred at household + ward contacts in hospital A	Not Shown
CC22	48-A	5	Hospital/Community	Centred at LTCF + ward contacts in hospital A	Not Shown
CC22	125-A	5	Hospital/Community	Centred at LTCF + hospital-wide contacts in hospital A	Not Shown
CC30	1-A	5	Hospital/Community	Centred at household + patient centric in hospital A	Fig S2 F
CC22	259-A	5	Hospital/Community	Centred at LTCF + ward contacts in hospital A	Not Shown