Aanensen, David M.; Feil, Edward J.; Holden, Matthew T G; Dordel, Janina; Yeats, Corin A.; Fedosejev, Artemij; Goater, Richard; Castillo-Ramirez, Santiago; Corander, Jukka; Colijn, Caroline; Chlebowicz, Monika A.; Schous, Leo; Heck, Max; Pluister, Gerlinde; Ruimy, Raymond; Kahlmeter, Gunnar; Ahman, Jenny; Matuschek, Erika; Friedrich, Alexander W.; Parkhill, Julian; Bentley, Stephen D.; Spratt, Brian G.; Grundmann, Hajo; Krziwanek, Karina; Stumvoll, Sabine; Koller, Walter; Denis, Olivier; Struelens, Marc; Nasheva, Dimitr; Budimir, Ana; Kalenic, Smilja; Pieridou-Bagatzouni, Despo; Jakubu, Vladimir; Zemplíkova, Helena; Westh, Henrik; Larsen, Anders Rhod; Skov, Robert; Laurent, Frederic; Ettienne, Jerome; Strommenger, Birgit; Witte, Wolfgang; Vourli, Sofia; Vatopoulos, Alkis; Vainio, Anni; Vuopio-Varkila, Jaana; Fuzi, Miklos; Ungvári, Erika; Murchan, Stephan; Rossney, Angela; Miklasevics, Edvins

Whole-genome sequencing for routine pathogen surveillance in public health

Published in:
MBIO

DOI:
10.1128/mBio.00444-16

Published: 01/05/2016

Please cite the original version:

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.
Whole-Genome Sequencing for Routine Pathogen Surveillance in Public Health: a Population Snapshot of Invasive Staphylococcus aureus in Europe


Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom; The Centre for Genomic Pathogen Surveillance, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom; The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom; School of Medicine, University of St. Andrews, St. Andrews, United Kingdom; Pathogen Genomics, The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom; Department of Biology, Drexel University, Philadelphia, Pennsylvania, USA; Programa de Genómica Evolutiva, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de Mexico, Cuernavaca, Morelos, Mexico; Helsinki Institute for Information Technology HIIT, Aalto, Finland; Department of Mathematics, Imperial College London, London, United Kingdom; Department of Medical Microbiology, University Medical Center Groningen, Rijksuniversiteit Groningen, Groningen, The Netherlands; National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands; Centre Hospitalier Universitaire de Nice, Nice, France; EUCAST Development Laboratory, Växjö, Sweden; Department of Infection Prevention and Hospital Hygiene, Faculty of Medicine, University of Freiburg, Freiburg, Germany

D.M.A. and E.J.F. contributed equally to this work.

† We deeply regret the untimely loss of our dear friend and colleague Helmut Mittermayer, to whom we dedicate this paper.

ABSTRACT The implementation of routine whole-genome sequencing (WGS) promises to transform our ability to monitor the emergence and spread of bacterial pathogens. Here we combined WGS data from 308 invasive Staphylococcus aureus isolates corresponding to a pan-European population snapshot, with epidemiological and resistance data. Geospatial visualization of the data is made possible by a generic software tool designed for public health purposes that is available at the project URL (http://www.microreact.org/project/EkUvg9uY?tt=rc). Our analysis demonstrates that high-risk clones can be identified on the basis of population level properties such as clonal relatedness, abundance, and spatial structuring and by inferring virulence and resistance properties on the basis of gene content. We also show that in silico predictions of antibiotic resistance profiles are at least as reliable as phenotypic testing. We argue that this work provides a comprehensive road map illustrating the three vital components for future molecular epidemiological surveillance: (i) large-scale structured surveys, (ii) WGS, and (iii) community-oriented database infrastructure and analysis tools.

IMPORTANCE The spread of antibiotic-resistant bacteria is a public health emergency of global concern, threatening medical intervention at every level of healthcare delivery. Several recent studies have demonstrated the promise of routine whole-genome sequencing (WGS) of bacterial pathogens for epidemiological surveillance, outbreak detection, and infection control. However, as this technology becomes more widely adopted, the key challenges of generating representative national and international data sets and the development of bioinformatic tools to manage and interpret the data become increasingly pertinent. This study provides a road map for the integration of WGS data into routine pathogen surveillance. We emphasize the importance of large-scale routine surveys to provide the population context for more targeted or localized investigation and the development of open-access bioinformatic tools to provide the means to combine and compare independently generated data with publicly available data sets.

Received 5 April 2016 Accepted 12 April 2016 Published 5 May 2016


Editor Keith P. Klugman, Department of Global Health, Emory University

Copyright © 2016 Aanensen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license. Address correspondence to Hajo Grundmann, hajo.grundmann@uniklinik-freiburg.de.

This article is a direct contribution from a Fellow of the American Academy of Microbiology. External solicited reviewers: Jennifer Gardy, B.C. Centre for Disease Control; Geoffrey Coombs, Murdoch University, Australia.
Most bacterial infections of humans are caused by organisms that have a large population size and short generation times, and lineages with novel properties emerge and expand within observable time scales (1). Lineages that present a serious threat to public health are designated high-risk clones (HRCs) (2); these often combine enhanced virulence or transmission potential with multiple-antibiotic resistance. As HRCs are often difficult to treat and are associated with significant morbidity, mortality, and economic cost, they require targeted surveillance and containment at the population level. The considerable management challenge for public health microbiologists in tackling HRCs can be broken down into three tasks: (i) identification of public health risks posed by emerging and/or expanding HRCs, (ii) assessment of this risks by predicting important clinical and epidemiological consequences, and (iii) risk management through the implementation of prevention and control strategies. With the cost of sequencing entire bacterial genomes in steady decline and the development of powerful bioinformatic tools gathering pace, whole-genome sequencing (WGS) will inevitably be widely implemented for routine epidemiological surveillance. The dissemination of genomic data through established national and international networks of collaborating specialist laboratories will lead to increased awareness and shorter response times for HRCs. Ideally, these networks will be based on a shared bioinformatic infrastructure that links molecular data (genome sequences) and metadata (time, place, clinical details, and additional variables) with tools that help appraise the clinical and public health relevance of any given entry. This poses significant technical, ethical, and political challenges; however, the benefits of this fundamental shift in molecular epidemiology cannot be overstated. The efficient management and interrogation of genome data by public health and medical authorities may ultimately lead to the erosion of conventional reference diagnostic tasks such as identification to the species level, characterization of clinically important virulence or resistance phenotypes, and identification of outbreaks and inference of national or international transmission.

_Staphylococcus aureus_ represents an epidemiological paradigm because of the undisputed public health relevance of this species, which is characterized by multiple-antibiotic resistance and a potential for swift dissemination through health care, social, and farm animal production networks. Here we demonstrate the utility of WGS when applied to a continental-scale representative “population snapshot” by using a novel data visualization platform. We sequenced 308 _S. aureus_ isolates responsible for invasive infections that were recovered from 186 hospitals in 21 countries across Europe in a 6-month period. We consider three analytical strands: (i) a representative phylogeographic analysis that defines HRCs on a population level, (ii) an analysis of the dynamics of virulence and resistance carried by mobile genetic elements (MGEs), and (iii) an _in silico_ ascertainment of antibiotic resistance encompassing 19 antibiotic compounds of clinical relevance.

To underline the added value for public health decision making when WGS data are supplemented with epidemiological metadata collected through structured surveys, we developed a web application (Microreact) that allows easy access and visualization of our data by medical and public health audiences and is available at the project URL (http://www.microreact.org/project/EkUvg9uYtt=re). This tool can be used for any appropriate data set where a phylogenetic tree and associated metadata are available (see http://www.microreact.org).

**RESULTS**

We chose a random sample (n = 308; 10.6%) of isolates collected as part of a European structured survey of _S. aureus_ from invasive diseases (3). Sixty percent (n = 185) of these isolates were methicillin-sensitive _S. aureus_ (MSSA), and 40% (n = 123) were methicillin-resistant _S. aureus_ (MRSA). A total of 235,226 SNP sites within the core genome were identified by mapping against a single reference genome, HG 5096 0412 (sequence type 22 [ST22]). We have divided the analysis and interpretation of the results into three parts: (i) a broad phylogenetic analysis and more fine-scaled consideration of exemplar lineages of high public health relevance, (ii) the distribution and dynamics of the accessory genes conferring virulence and resistance traits in the context of individual HRCs, and (iii) a comparison of antibiotic susceptibility profiles ascertained by _in silico_ prediction from genome data with conventional susceptibility testing carried out in Staphylococcal Reference Laboratories (SRLs) and the European Committee on Antimicrobial Susceptibility Testing (EUCAST) Development Laboratory (EDL).

Broad and fine-scale phylogenetic analyses for the identification of high-risk clones. Analysis of core SNPs resolved the population into 6 major (CC5, CC22, CC8, CC30, CC45, CC15) and 10 minor clonal complexes (CCs) and STs (CC1, ST20, ST25, ST7, CC121, ST88, CC12, CC398, ST101, ST72), as previously defined by multilocus sequence typing (MLST)/eBURST (4) (Table 1; Fig. 1 and 2). The six major CCs are each represented by at least 24 isolates, whereas the minor lineages are represented by a minimum of four isolates and a maximum of 14. Subdivisions within a given CC and within individual STs are resolved by the data (Fig. 2E), which identify very closely related isolates resulting from recent expansion. We refer to these groups within CCs simply as clusters. These include many well-known MRSA “clones” previously defined by MLST, pulsed-field gel electrophoresis, and _spa_ typing.

The distribution of MRSA and MSSA isolates is not random with respect to CCs (Table 1). Ninety-three percent of the MRSA isolates belonged to only three CCs, CC22, CC5, and CC8, although these CCs represent only approximately half of the isolates (Table 1). Over 70% of the isolates in CC5, CC8, and CC22 are MRSA, whereas all of the isolates in CC15 are MSSA. CC45 shows moderate levels of MRSA (23%), and CC30 shows low levels (6%). The overall topology of the tree is highly consistent with previous studies (5) (Fig. 1 and 2). The impact of recombination on tree topology was investigated by computing the frequency of homoplasies (phylogenetic conflicts within the data). This confirmed previous work that revealed a high rate of homoplasy between the CCs (37.8% of the core SNPs) but very low levels when considering the variation within each of the CCs (0.62% of the core SNPs) (5). We are therefore confident that the fine-scaled phylogeographic inferences within individual CCs have not been seriously compromised by recombination. Intra-CC diversity was calculated simply as the mean number of SNP differences in all pairwise comparisons, and the average time since each pair of genomes shared a common ancestor was calculated on the basis of the published _S. aureus_ mutation rate (Table 1) (6–9). This revealed similar levels of diversity between CCs and that, on average, the common ancestor of any pair of genomes belonging to the same major CC existed from the mid-1930s (CC45) to the mid-1970s (CC15).

A key observation from our previous work on _spa_ typing was...
MLST is the supplemental material, with CC45 as an example. Although some of these isolates originate from areas close to the shared border (e.g., Ústí nad Labem in the Czech Republic and Grossenhain in Germany are just 122 km apart by road). The data thus point to the border between Germany and the Czech Republic as being a barrier to health care referral practices and thus MRSA spread.

The MRSA cluster corresponding to ST228 and related STs consists of 15 isolates harboring a type I SCCmec element. Nine of these isolates define a very tight subcluster corresponding to ST111 (n = 8) and ST1481 (n = 1). Eight of these nine isolates originate from Croatia, which is indicative of rapid epidemic radiation in that country, and three isolates from Split cluster together on the tree, reflecting geographic structuring on a national level. One ST111 isolate originates from Sweden, consistent with the probable importation of ST111 into Sweden from Croatia by travel. The MRSA ST125 cluster consists of eight MRSA isolates restricted to central and northern Spain. For further investigation, see the Microreact tool, where supplementary analysis and geographic detail can be explored.

**CC22.** CC22 contains epidemic MRSA-15 (EMRSA-15) (Fig. 4A), which is currently the most abundant and fastest growing health care-associated MRSA (HA-MRSA) clone in Europe (10). This clone accounts for the majority of the CC22 isolates (31/41; 76%) in our sample, which are characterized by an SCCmec type IV element and form a tightly clustered starlike phylogeny (Fig. 4A). Fourteen of these isolates originated from the
United Kingdom, 11 were from Germany, and 7 were from Portugal. The topology of the tree supports a United Kingdom origin of EMRSA-15, followed by separate introductions from the United Kingdom into Germany and Portugal, consistent with the work of Holden et al. (8). In the German cluster, isolates from Berlin show the fewest SNP differences from the United Kingdom isolates, implicating that city as the initial point of introduction from the United Kingdom (Fig. 4B). This clone subsequently spread to the surrounding cities of Kiel, Frankfurt (Oder), Hannover, and Magdeburg, supporting epidemiological observations made by the Robert Koch Institute during investigations in the 1990s (11). Similarly, the Portuguese CC22 cluster points toward Lisbon as the point of introduction, with subsequent spread to Braga (in the far north) and Coimbra (central). All of the CC22 isolates located basal to the EMRSA-15 subtree are MSSA. They form a more diverse population, genetically and geographically, and represent the MSSA reservoir from which EMRSA-15 emerged (8).

**CC30.** CC30 contains successful HA-MRSA and community-associated MRSA (CA-MRSA) epidemic clones, including EMRSA-16 and the Panton-Valentine leukocidin (PVL) toxin-positive Southwest Pacific (SWP) clone. Figure 5A shows the tree for the 32 CC30 isolates present in the current sample and illustrates that only two (6.5%) were MRSA. One of the MRSA isolates is closely related to the MRSA252 reference, which is a HA-EMRSA-16 isolate (ST36), and the other (ST1829) is related to the TCH60 reference, a representative isolate of the SWP clone (14). The majority of 28 MSSA isolates form a striking starlike phylogeny with little geographic structure (for detailed investigation of geographic origin, see the Microreact tool). To place this lineage in a broader context, we combined our data with those of McAdam et al. (7), who recently sequenced representatives of the three major CC30 epidemic lineages: EMRSA-16, SWP, and the historical MSSA phage type 80/81 clone (Fig. 5B). This revealed that the large MSSA radiation corresponds to a successful progenitor population from which HA-EMRSA-16 emerged. This MSSA cluster, referred to by MacAdam et al. as “other epidemic,” is responsible for a significant disease burden in the community and in hospitals and encompasses a variety of spa types. Given the public health significance of this cluster, we tentatively suggest the designation
EMSSA-ST30. All of the isolates in this cluster contain the SNPs in \( hla \) and \( agrC \) previously reported to attenuate virulence (18) (Fig. 5B). The presence of these SNPs among community-associated invasive disease isolates further challenges the suggestion that they play a role in nosocomial adaptation (18).

**Distribution and dynamics of accessory genes.** The accessory genome consists of genes that are variably present in a defined population and can be major drivers of adaptation (15). An understanding of the dynamics of key MGEs such as phages and the SCC\( \text{mec} \) elements, which carry determinants of the virulence and resistance phenotypes, is critical for monitoring the emergence and diversification of successful HRCs. We explored the diversity and distribution of accessory genes in our *de novo* assembled genomes by assigning genes to noncore homology groups (ncHGs). We categorized these groups according to the types of mobile elements with which they are associated, specifically, phages, SCC\( \text{mec} \) elements, plasmids, \( S. \text{aureus} \) pathogenicity islands, and transposons (Fig. 2). Fifty-seven percent of the ncHGs in the sample were phage associated, representing the most dominant category by far. The total number of ncHGs varies between different CCs; CC8 harbors the highest average number of ncHGs per isolate at 197, and CC15 has the lowest at 71 (see Fig. S5 in the supplemental material).

Figure 2D illustrates the variation in the accessory genome within single CCs, and Fig. 6 shows the total number of ncHGs shared by each pair of genomes. Many more ncHGs are shared by isolates belonging to the same CC than by isolates belonging to different CCs. This supports the concept of a “core variable” genome (16), representing genes that are universally present at the scale of an individual CC but are variably present or absent with respect to the species as a whole. We also noted cases of MRSA clusters within CCs that have acquired a distinctive repertoire of ncHGs, evident as dark squares in Fig. 6; for example, the ST239 cluster (CC8), the ST228 cluster (CC5), and the ST225 cluster (also CC5). The high level of consistency between phylogenetic relatedness and the ncHG repertoire is expected over these very fine phylogenetic scales because of common inheritance. However, this consistency rapidly decays with increasing core genome divergence, particularly for highly dynamic elements such as phages. It is evident from Fig. 6 that CC5 and CC8 have a high number of ncHGs in common, resulting in a single large square that encompasses both CCs. The large proportion of MRSA strains within CC5 and CC8 does not, by itself explain, this observation, as it is still apparent when SCC\( \text{mec} \) is excluded (data not shown).

Comparing the total number of ncHGs within MRSA and MSSA isolates in the same CC revealed that MRSA clusters contain more ncHGs than closely related MSSA genomes, even when the SCC\( \text{mec} \) elements were excluded (see Fig. S6 in the supplemental material). For example, considering CC5 and CC22,
MRSA isolates contain approximately 50% more ncHGs than MSSA isolates of the same complex. Although it is unclear whether the acquisition of SCCmec is the cause or the consequence of an increased propensity to integrate horizontally acquired DNA, this observation may be relevant to the finding that MRSA strains tend to contain many more resistance determinants than MSSA strains, conferring combined resistance to multiple classes of antibiotics.

Phages are important from a public health perspective, as they are carriers of virulence genes and resistance determinants. We assayed the distribution of seven phage types on the basis of the presence/absence of their integrase genes as described previously (17). The overall number of prophages per genome ranged from zero (9%) to five (0.3%), with a median of two (46%). Sa3int was the most commonly observed phage type, being present in 82% of our isolates, followed by Sa2int (32%), and Sa1int (27%). Sa4int was very rare, being observed in only four isolates (1.3%). The distribution of these phage types and associated cargo genes across the tree reveals important differences between the CCs and between different clusters belonging to the same CC (see Fig. S7 in the supplemental material). For example, the Sa3int prophage was found in all of the CCs, with the exception of CC15. This prophage can harbor four genes belonging to the immune evasion cluster, chp, sak, scn, and seal/sep, and in our data mostly harbored sak and scn, with the addition of either chp (e.g., CC22, CC45) or seal/sep (e.g., CC8, except USA300). Notably, chp and scn are common in CC15 even though the prophage itself is absent. The most likely explanation for this pattern is that the prophage was once integrated and then lost from the genome, leaving chp and scn behind.

SasX is a cell wall-anchored surface protein that is linked to the epidemiological success of ST239 in China and Southeast Asia (19). The φSPβ-like prophage that encodes SasX was identified among three ST239 isolates from Poland (see Fig. S7D). Two of these isolates cluster with the TW20 reference, which is represen-
tative of the “Asian clade” of ST239 (see Fig. S3), pointing to the introduction of these isolates into Poland from Asia. The third sasX-positive ST239 isolate in our data is more likely to be a native Polish variant and most likely acquired sasX from the imported Asian strains cocirculating in Poland. The data thus point to the horizontal dissemination, presumably via lysogenic conversion, of the phage-borne sasX virulence factor within Europe. Other examples of horizontal dissemination of virulence genes are evident in our data. We note only a single representative of USA300, an MRSA strain from Belgium that closely clusters with the reference USA300 genome and contains a type I arginine catabolic MGE (ACME) (see Fig. S2). However, we also note a single CC5 SCCmec type V MRSA isolate from Portugal that has also acquired a type I ACME region (Fig. 3; see Fig. S7D). The distribution of other toxin and virulence genes, such as the presence of tstH in the cluster we designated EMSSA-ST30, is illustrated in Fig. S7D and incorporated into the visualization platform at the project URL.

Multiple SCCmec types (I, II, III, IV, V, and IV) were present among the 120 MRSA isolates. Type IV elements are the most widely distributed among different lineages, being present in all five of the major CCs that contain MRSA strains. From the phylogenetic analyses, we estimate that a minimum of 20 independent acquisitions of this element have occurred in our sample. However, the rate of SCCmec acquisition clearly varies between CCs, whereas only a single acquisition of the SCCmec type IVh element is observed in CC22, in CC5, there have been a minimum of seven different acquisition events encompassing five different SCCmec types. For example, the Spanish ST125 cluster harbors both SCCmec type IVc and IVs elements, indicating multiple acquisitions. In contrast, no CC15 isolates have acquired SCCmec. The reversion of MRSA to MSSA because of the deletion of SCCmec appears to be a relatively rare event and was identified in only two CC8 isolates in our sample (see Fig. S2).

**WGS as a tool for predicting antibiotic resistance.** The EDL tested all of the isolates against 16 antibiotics. In addition, all MRSA isolates were tested against a further three antibiotics that are prescribed mainly for the treatment of infections caused by MRSA. A total of 5,288 in silico predictions of resistance/susceptibility were made and compared against the EUCAST reference results in a blinded fashion. Of these, 1,075 were predicted to be resistant, compared with 1,050 identified phenotypically as resis-

![FIG 5](https://example.com/figure5.png)

**FIG 5** (A) Phylogenetic reconstruction of CC30 isolates in the sample. Branch color indicates MSSA (green) or MRSA (red). Reference genomes are named and clusters are highlighted and named in accordance with the report of McAdam et al. (7). The dashed line indicates the long branch leading to three ST34 isolates that evolved through the acquisition of a 200-kb homologous replacement within the chromosome from an ST10/ST145-like parent (66). (B) Phylogenetic tree of combined CC30 data obtained from isolates from the study of McAdam et al. (7) and isolates from panel A. Colors and cluster names are as in panel A. Light gray shading indicates isolates carrying SNPs in the hla and agrC genes thought to restrict these lineages to health care settings. (C) Representation of successive clonal radiations within the recent evolutionary history of CC30. These radiations correspond to recognized HRCs, both contemporary and historic. The SWP clone is a historic diversified starlike expansion with relatively long branches. Phage type 80/81 probably emerged from within this starlike expansion, as did the current MSSA HRC, which we have termed EMSSA-ST30. Finally, EMRSA-16 emerged from within the successful EMSSA-ST30 lineage, resulting in a more recent, and tightly clustered, starlike expansion.
tant by the EDL, with concordance noted for 5,213 (98.6%) of the individual tests (Fig. 2; Table 2). However, there were some noteworthy exceptions. For the aminoglycoside amikacin, 21 isolates were falsely predicted to be resistant on the basis of the presence of the *aphA*-3 kanamycin resistance gene and *aadD*. It has previously been noted that the presence of these two aminoglycoside-modifying enzyme genes is a poor indicator of amikacin resistance (20, 21). Resistance to mupirocin was not predicted in five isolates that had inhibition zones (29 mm) close to the susceptibility breakpoint (≥30 mm). The reason for this discrepancy is likely to be normal variation or the choice of medium. There were also eight incorrect predictions of erythromycin susceptibility. In these cases, the result could readily be explained by loss of the *ermC*-carrying plasmid after sequencing but prior to phenotypic testing. Similarly, the loss of a BlaZ β-lactamase-carrying plasmid could explain the three false-negative predictions of penicillin resistance.

We also determined the concordance of the antimicrobial susceptibility test (AST) results provided by the European *Staphylococcus* Reference Laboratories (SRLs), which contributed the isolates to our sample, with the EDL reference data. The SRLs submitted a total of 2,252 AST results. Of these, 2,203 (97.8%) were concordant with the EUCAST test results (Fig. 2; Table 2). Thus, it is clear that a small amount of discordance between phenotypic tests can be expected, even when they are carried out by experienced reference laboratories.

We note that the degree of discordance between the *in silico* predictions and the EUCAST data is comparable to the degree of discor-
dance between the two sets of phenotypic data. This demonstrates that in silico predictions were at least as reliable, in terms of matching the gold standard, as the AST results generated by independent reference laboratories.

**DISCUSSION**

A consistent and comprehensive sampling frame is a crucial component of effective pathogen surveillance in public health, though in reality, this may involve significant logistical and political challenges. By taking a random sample of a larger collection assembled as part of a pan-European structured survey, we have minimized the problems associated with phylogenetic discovery bias (22) and are confident that our data capture a meaningful snapshot of the problems associated with phylogenetic discovery bias (22) and as part of a pan-European structured survey, we have minimized the problems associated with phylogenetic discovery bias (22) and are confident that our data capture a meaningful snapshot of the disease-causing population of *S. aureus* at the time of sampling. However, as with all sampling frameworks, important caveats remain. First, all of the isolates were recovered from invasive infections, primarily blood, meaning that we have underrepresented isolates causing skin and soft tissue infections. Second, laboratories submitted the first five MRSA isolates from individual patients. A small number of laboratories did not receive five MRSA isolates during the 6-month survey because of a very low frequency of MRSA disease in their regions, for example in Scandinavia, and these laboratories completed their quota with MSSA isolates.

Despite these caveats, the combined analysis of WGS with epidemiological metadata addresses the following three key elements of managing infectious disease threats in public health.

**Genetic population structure and identification of HRCs.** The WGS data demonstrate that the disease-causing population of *S. aureus* is readily partitioned into highly discrete subpopulations or CCs. These CCs vary in their potential to spawn HRCs such as EMRSA-15 (8) or USA300 (9, 22), and our analysis suggests that these differences may be related to the varying propensity of each CC to acquire exogenous DNA (see Fig. S5 in the supplemental material). Analysis of fine-scale genetic variation and spatial structuring of the well-characterized HRCs circulating in Europe during the sampling period helped identify key population level properties, or HRC signatures, that can be used to recognize candidate HRCs, even in the absence of detailed phenotypic data (Fig. 3 to 5; see Fig. S1 to S4). Central to this is a consideration of phylogenetic tree topology, relative abundance, and geographic structuring.

Perhaps the most striking example is EMRSA-15, which emerged from the CC22 population approximately 30 years ago (8) and has subsequently become pandemic. Phylogenetic analysis of ST22 reveals a comet-shaped phylogeny, with two distinct parts; the comet head consists of the starlike radiation of EMRSA-15, reflecting recent rapid expansion, while the tail represents the more diverse MSSA progenitor population (Fig. 4A).

We find multiple similar signatures of rapid clonal expansions that coincide with recognized successful clones such as ST225, ST228, and ST125 within CC5 (Fig. 3). We also note previously unrecognized clusters in Europe that constitute candidate HRCs in the MSSA population, the most notable example being the CC30 MSSA designated EMSSA-ST30 here. By combining our data with the WGS data from a previous study of well-known CC30 HRCs (Fig. 5B), we placed EMSSA-ST30 within a broader evolutionary context. This revealed that the EMRSA-16 HRC emerged from EMSSA-ST30, as hypothesized originally from MLST data (24). This highlights the importance of recognizing successful MSSA lineages not only as HRCs in their own right but also in their role as likely progenitors of EMRSA-15 (25).

**Assessing the risks posed by virulence and resistance determinants for public health.** CCs differ in their propensities to acquire and maintain accessory genome elements (see Fig. S1 in the supplemental material), a factor that may influence the emergence of HRCs. CC30 represents a good example of a lineage that has acquired elements carrying important virulence and resistance de-

---

**TABLE 2** Comparison of antibiotic resistances predicted by in silico and SRL test results against the EDL reference

<table>
<thead>
<tr>
<th>Antibiotic</th>
<th>No. of in silico predictions vs EDL results</th>
<th>No. of SRL vs EDL results</th>
<th>No. of in silico vs EDL results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of traits</td>
<td>False positive</td>
<td>False negative</td>
</tr>
<tr>
<td>Penicillin</td>
<td>308</td>
<td>269</td>
<td>4</td>
</tr>
<tr>
<td>Cefoxitin</td>
<td>308</td>
<td>123</td>
<td>3</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>308</td>
<td>122</td>
<td>2</td>
</tr>
<tr>
<td>Moxifloxacin</td>
<td>308</td>
<td>118</td>
<td>2</td>
</tr>
<tr>
<td>Amikacin</td>
<td>308</td>
<td>71</td>
<td>21</td>
</tr>
<tr>
<td>Gentamicin</td>
<td>308</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>Tobramycin</td>
<td>308</td>
<td>77</td>
<td>7</td>
</tr>
<tr>
<td>Erythromycin</td>
<td>308</td>
<td>105</td>
<td>5</td>
</tr>
<tr>
<td>Clindamycin</td>
<td>308</td>
<td>95</td>
<td>3</td>
</tr>
<tr>
<td>Tetracycline</td>
<td>308</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>Tigecycline</td>
<td>308</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fusidic acid</td>
<td>308</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>Linezolid</td>
<td>308</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mupirocin</td>
<td>308</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Rifampin</td>
<td>308</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Trimethoprim</td>
<td>308</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Teicoplanin</td>
<td>120</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vancomycin</td>
<td>120</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Daptomycin</td>
<td>120</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>5,288</td>
<td>1,075</td>
<td>50</td>
</tr>
</tbody>
</table>

*Only results for antibiotics tested by SRLs were compared.*
terminants. In addition to the phage-borne PVL toxin gene harbored in phage type 80/81 and the SWP clone, the cluster of isolates that we designated EMSSA-ST30 maintains the tstH gene encoding a superantigen that can cause toxic shock syndrome at high frequency (26) (see Fig. S7). This cluster is therefore reminiscent of, and likely descended from, the ST30 variants responsible for the tampon-associated toxic shock syndrome that emerged in the 1980s (27). This provides another line of evidence for the public health risk posed by this widespread clone.

The success of ST239 in China and much of Southeast Asia has been largely attributed to SasX, a surface-anchored protein that modulates host interactions and transmissibility and is carried on the ßPS-like prophage (19). In Poland, we found evidence that ST239 isolates from the successful Asian lineage spread repeatedly to Europe (12) but also that this sasX virulence determinant has been transduced from the European ST239 genetic background (see Fig. S3). This observation justifies heightened epidemiological vigilance with respect to this HRC, not only for particular epidemiologically successful clones but also for virulence determinants they can spread.

A key characteristic of HRCs has been the emergence of antibiotic resistance. WGS provides the opportunity to scan genomes for all known genetic determinants of antibiotic resistance. We examined the distribution of known resistance determinants and the extent to which it is possible to predict resistance profiles from sequence data in a blinded fashion. Whereas previous studies have shown a high concordance between phenotypic data and genotypic predictions (8, 13, 28), we sought to extend these studies by including a greater number of antibiotics (19 in total) and by including a representative sample of isolates that cause invasive disease among patients in European hospitals and that we consider the clinically most relevant representatives of the S. aureus population. Further, as discrepancies between the two methods could arise from uncertainties in the phenotypic data, as well as inaccurate predictions from genotyping, we compared two independent sets of phenotypic data, one contributed by various SRLs and the second generated at the EDL. Taking the EDL data to represent the “gold standard,” our genotypic predictions show higher rates of concordance with these data (EDL versus genotype, 98.6%) than do the data generated by the SRLs (EDL versus SRL, 97.8%, Table 2). We conclude that genotypic prediction is at least as reliable as routine phenotypic testing and that any discrepancies between the two approaches are just as likely to represent inaccuracies in the phenotypic testing as inaccurate genotypic predictions.

We do not, however, propose that genotypic prediction should replace phenotypic testing, as such a strategy would be vulnerable to the emergence of new and uncharacterized antibiotic resistance mechanisms. Moreover, in silico prediction for other organisms has proved more challenging, especially for Gram-negative bacteria (29), where the present understanding of the genetic basis of resistance is less comprehensive. Despite these limitations, WGS data clearly provide additional objective evidence that is not prone to heterogeneities inherent to conventional phenotypic test methods deployed in different reference laboratories. Moreover, in the longer term, WGS will deliver a digitized and cumulative record that will address the need for internationally agreed standards for collection of data and reporting on antibacterial resistance in human health and for harmonizing standards across medical, veterinary, and agricultural sectors as required by the WHO, the World Organization for Animal Health, and the Food and Agriculture Organization of the United Nations (30).

**Deployment of informed and targeted prevention and control strategies.** We have argued for a two-pronged approach to the identification of emerging HRCs and an assessment of the threats they pose to public health, (i) a consideration of clonal relatedness, abundance, and phylogeographic structure at the population level and (ii) mining of the accessory genome repertoire to ascertain likely virulence and resistance properties. These two perspectives should also be combined when considering appropriate containment measures. The recognition of an HRC through large-scale surveillance would enable the development of tailored rapid diagnostic tests based on distinguishing SNPs or accessory gene signatures. This will allow for selective screening and targeted containment by decontamination and isolation strategies that are easier to implement, more economical, and more likely to be effective than more generic procedures. The expedition of specifically tailored tests during outbreaks would allow health authorities or infection control practitioners to screen potential hosts for colonization or infection, thereby reducing the chance of onward transmission. WGS allows for rapid appraisal of outbreak sources and transmission pathways that will also help in weighing up infection control priorities. Moreover, information regarding the virulence gene repertoire associated with an emerging or outbreak HRC will inform clinicians and medical microbiologists of likely clinical manifestations.

Ultimately, the key to managing HRCs lies in making the data available in an open and intuitive format for infection control and public health audiences. This democratization of the data increases the collective power of interpretation (i.e., the identification of HRCs) while decreasing the necessity for local expertise in bioinformatics. The coupling of large-scale population sampling by WGS within open-access and freely available web resources empowers the community to identify clone-specific signatures (canonical SNPs and/or accessory genes), promoting the design of HRC-specific rapid and cost-effective molecular diagnostic tests (31).

With this communication, we aimed to demonstrate how the integration of WGS into epidemiological surveillance programs provides the means for both the early warning of emerging HRCs and a robust assessment of associated public health threats. However, the advent of the underlying sequencing technology addresses only a small part of the challenge of managing the threat from infectious disease. In order to exploit this technical advance to its maximum potential, two things must happen. First, national reference laboratories need to agree to, and abide by, common standards for the contribution of isolates to structured surveys, ideally at 5-year or shorter intervals, on pancontinental and, ideally, global scales. Second, platforms for bioinformatics need to be developed to allow intuitive management and exploration of the data. These data sets should conform to internationally curated standards for sets of genes and mutations that are recognized as key virulence or resistance determinants, and we propose the initiation of internationally curated data sets to act as a gold standard resource for genomic antimicrobial resistance determinants.

The establishment of cumulative databases will engender a far richer understanding of the detailed dynamics underpinning clonal emergence and replacement, national and international transmission, and the horizontal transfer of core genes and MGEs. The increasing public health threat from bacterial pathogens is, in large part, down to the ability of these organisms to rapidly adapt through the dissemination of genes and mobile elements. Our best
chance of managing these threats in the future is to emulate, as far as possible, this resource and data sharing through the development of international surveillance networks and a common database infrastructure.

MATERIALS AND METHODS

Sampling framework and bacterial isolates. From September 2006 to February 2007, 357 laboratories serving 450 hospitals in 26 countries collected nearly 3,000 MSSA and MRSA isolates from patients with invasive S. aureus infections, as described previously (3). Approximately 10% (n = 308) of these were randomly selected for sequencing. Isolate details are available for download at the project URL.

Genomic library creation and sequencing. For each sample, index-tagged libraries were prepared and sequenced in Illumina Genome Analyzer GAII cells with 54-base paired-end reads or Illumina HiSeq with 75-base-paired-end reads. Downstream processing utilized the index tag sequence information to assign reads to individual samples.

Reference genomes. To provide a wider context to the data, we utilized 26 fully annotated complete S. aureus reference strains. The strain names and accession numbers are included as follows: 04-2981, CP001844 (6); COL, CP000046 (32); ECT-R2, FR714927 (33); ED313, CP001996 (34); ED98, CP001781 (35); HO 0906 0412, HE681097 (8); JH1, CP000736 (36); JH9, CP000703 (36); JKD6008, CP002120 (37); JKD6159, CP002114 (38); LGA251, FR821779 (39); MRSAs2A, B5X71856; MSSA476, B5X71857 (40); Mu3, A0089324 (41); Mu50, BA000017 (42); MW2, BA000033 (43); N315, BA000018 (42); NCTC8325, CP000253 (44); Newman, AP009351 (45); RF122, AJ938182 (46); BB155, LN854566; ST398, AM990992 (47); TCH60, CP002110, TW20, FN433596 (48); USA300, FPR3757, CP000255 (23); USA300, TCH1516, CP000730 (49).

Detection of SNPs in the core genome. The paired-end reads for the survey isolates and the 26 publicly available genomes were mapped with SMALT (http://www.sanger.ac.uk/resources/software/smalt/) against the survey isolates and the 26 publicly available genomes were mapped with chromosome BX571857 (40); Mu3, AP009324 (41); Mu50, BA000017 (42); MW2, BA000033 (43); N315,BA000018 (42); NCTC8325,CP000253 (44); New- sequence information to assign reads to individual samples.

Phenotypic testing of antibiotic resistance. The antibiotic susceptibility of all S. aureus isolates was tested by the standardized EUCAST disk diffusion method in the EDL, Växjö, Sweden, without prior knowledge of the sequence data or metadata. Full methodological details are available in the EUCAST Disk Diffusion Test Manual, v 3.0, 2013 (http://www.eucast.org). Phenotypic resistance was defined by applying inhibition zone diameter epidemiological cutoff values (ECOFFs) and EUCAST clinical breakpoints (EUCAST breakpoint tables for interpretation of MICs and zone diameters, version 3.1, 2013 [http://www.eucast.org]). We pooled the intermediate and resistant categories into a single nonsusceptible category. For our sample, this partition was fully consistent with the classification into wild-type and non-wild-type isolates defined by the EUCAST ECOFFs. The antibiotics tested were penicillin, cefoxitin, ciprofloxacin, moxifloxacin, amikacin, gentamicin, tobramycin, erythromycin, clindamycin, tetracycline, tigecycline, fusidic acid, linezolid, mupirocin, rifampin, and trimethoprim. MRSA resistance to clinically relevant reserve antibiotics (teicoplanin, vancomycin, and daptomycin) not suitable for disk diffusion testing was determined by using the respective MIC ECOFFs (2.0, 2.0, and 1.0 mg/liter, respectively) with E tests (bioMerieux Clinical Diagnostics, Marcy l’Etoile, France). Phenotypic test results were dichotomized and grouped into wild-type and non-wild-type categories. Results of genomic predictions were compared with the phenotypic classifications as resistant or susceptible as defined by the ECOFFs. Moreover, results that were made available by the SRLs were equivalently dichotomized and also compared to those generated by the EDL in a consistent manner.

Web application. Microreact is a Node.js application written in JavaScript. Locational data are displayed by using the Google Maps Application Programming Interface, and phylogenetic data (in Newick tree format) are displayed by using a custom tree viewer developed for the HTML5 Canvas element (PhyloCanvas). Image files (.png) of annotated trees and subtrees can be saved. Instructions and examples are shown at the microreact.org website. The project URL for exploring the data described in this article is http://www.microreact.org/project/EkUvg9uY?tt=rc.

Nucleotide accession numbers. WGS data for all of the isolates tested in this study have been deposited in the Sequence Read Archive, and accession numbers are included in the metadata available at the project URL.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00444-16/-/DCSupplemental.
We thank the diagnostic laboratories across Europe that contributed their isolates during the structured surveys and whose efforts are crucial to the identification of HRCs and understanding of their spread through European populations. We also thank the sequencing and informatics teams at the Sanger Institute for their assistance and the Wellcome Trust for their support of the Sanger Institute Pathogen Genomics and Biology group.


The members of the European SRL Working Group are Helmut Mittnermayer, Karina Krizwanek, Sabine Stumvoll, Walter Koller, Olivier Deris, Marc Struelens, Dimitr Nashev, Ana Budimir, Smilja Kalenic, Despo Pantosti, Michael Borg, Marga van Santen-Verheuvel, Xander Huijsdens, Gunnsteinn Haraldsson, Karl G. Kristinsson, Monica Monaco, Annalisa Pantosti, Michael Borg, Marga van Santen-Verheuvel, Xander Huijsdens, Lillian Marstein, Trond Jacobsen, Gunnar Skov Simonsen, Marta Airese-Dousa, Hermienel de Lencastre, Agnieszka Luczak-Kadubowska, Waleria Hryniewicz, Monica Straut, Irina Codita, Maria Perez-Vazquez, Jesus Oteo Iglesias, Vesna Cvikovic Spik, Manica Mueller-Premru, Sara Hae ggman, Barbro Olsson-Liljegquist, Mathiell Ellington, and Ankara Kears.

For their affiliations, contact information, and contributions, see Table S2 in the supplemental material.

FUNDING INFORMATION
This work, including the efforts of Matthew Holden, Janina Dordel, Julian Parkhill, and Stephen Bentley, was funded by Wellcome Trust (098051). This work, including the efforts of David M. Aanensen, Corin Yeats, and Artemij Fedosejev, was funded by Wellcome Trust (099202). This work, including the efforts of Brian Spratt, was funded by Wellcome Trust (089472). This work, including the efforts of Santiago Castillo-Ramirez, was funded by Medical Research Council (MRC) (G1000803).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

http://dx.doi.org/10.1002/1521-1878(200012)22:12<1115::AID-BIES9>3.3.CO;2-L


