

RESEARCH

Open Access

# PathoSPOT genomic epidemiology reveals under-the-radar nosocomial outbreaks



Ana Berbel Caban<sup>1†</sup>, Theodore R. Pak<sup>2†</sup>, Ajay Obla<sup>2</sup>, Amy C. Dupper<sup>1</sup>, Kieran I. Chacko<sup>2</sup>, Lindsey Fox<sup>1</sup>, Alexandra Mills<sup>1</sup>, Brianne Ciferri<sup>2</sup>, Irina Oussenko<sup>2</sup>, Colleen Beckford<sup>2</sup>, Marilyn Chung<sup>2</sup>, Robert Sebra<sup>2,3,4,5</sup>, Melissa Smith<sup>2,3</sup>, Sarah Conolly<sup>6</sup>, Gopi Patel<sup>1,6</sup>, Andrew Kasarskis<sup>2,3,7</sup>, Mitchell J. Sullivan<sup>2</sup>, Deena R. Altman<sup>1,2†</sup> and Harm van Bakel<sup>2,3\*†</sup> 

## Abstract

**Background:** Whole-genome sequencing (WGS) is increasingly used to map the spread of bacterial and viral pathogens in nosocomial settings. A limiting factor for more widespread adoption of WGS for hospital infection prevention practices is the availability of standardized tools for genomic epidemiology.

**Methods:** We developed the Pathogen Sequencing Phylogenomic Outbreak Toolkit (PathoSPOT) to automate integration of genomic and medical record data for rapid detection and tracing of nosocomial outbreaks. To demonstrate its capabilities, we applied PathoSPOT to complete genome surveillance data of 197 MRSA bacteremia cases from two hospitals during a 2-year period.

**Results:** PathoSPOT identified 8 clonal clusters encompassing 33 patients (16.8% of cases), none of which had been recognized by standard practices. The largest cluster corresponded to a prolonged outbreak of a hospital-associated MRSA clone among 16 adults, spanning 9 wards over a period of 21 months. Analysis of precise timeline and location data with our toolkit suggested that an initial exposure event in a single ward led to infection and long-term colonization of multiple patients, followed by transmissions to other patients during recurrent hospitalizations.

**Conclusions:** We demonstrate that PathoSPOT genomic surveillance enables the detection of complex transmission chains that are not readily apparent from epidemiological data and that contribute significantly to morbidity and mortality, enabling more effective intervention strategies.

**Keywords:** PathoSPOT, Nosocomial outbreaks, Whole-genome sequencing, Visualization toolkits, MRSA bacteremia

\* Correspondence: [harm.vanbakel@mssm.edu](mailto:harm.vanbakel@mssm.edu)

<sup>†</sup>The authors Ana Berbel Caban and Theodore R. Pak contributed equally to this work.

<sup>†</sup>The authors Deena R. Altman and Harm van Bakel are co-senior authors on this work.

<sup>2</sup>Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York City, NY, USA

<sup>3</sup>Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

The utility of whole-genome sequencing to track transmissions and outbreak events is well-established, in particular for highly clonal pathogens such as methicillin-resistant *Staphylococcus aureus* (MRSA), where classical molecular methods such as multi-locus sequence typing (MLST) and pulsed-field gel electrophoresis (PFGE) do not provide enough resolving power [1–4]. Despite the increased use of WGS, bottlenecks remain that complicate its use in detecting and managing nosocomial outbreaks. Comparative genome analyses often require specialized knowledge and/or selection of appropriate reference sequences. Analysis and visualization frameworks are available to aid genome analysis in global or regional outbreaks [5–7], but these are less suited for nosocomial settings where genomic data need to be integrated with detailed patient histories for contact tracing. This can be time-consuming, especially when relying on manual chart review. Integration with electronic medical record (EMR) systems can aid this process, but tools that combine patient and genomic information in a comprehensive manner are not readily available.

To facilitate detection and mapping of transmission chains in nosocomial settings, we developed the open-source Pathogen Sequencing Phylogenomic Outbreak Toolkit (PathoSPOT), which combines automated comparisons of complete or draft genomes with interactive visualization of clonal clusters. Further integration of epidemiological data enables high-resolution analysis of outbreak phylogenies and contact tracing. We used our toolkit as part of a complete genome surveillance program of MRSA, a common cause of healthcare-associated infections in the USA that pose a fatal threat to patients. PathoSPOT comparisons of MRSA genomes from 197 bacteremic patients identified multiple transmission events and a hospital-wide outbreak encompassing 16 patients that had not been detected by conventional infection prevention strategies. In-depth analysis with PathoSPOT allowed us to reconstruct the outbreak timeline and identify common links among these individuals. Our findings demonstrate the utility of PathoSPOT for precision surveillance in healthcare systems and highlight the role of colonization in long-term nosocomial outbreaks.

## Methods

### Isolate selection, MRSA culturing, DNA extraction, and sequencing

Primary isolates from all MRSA bacteremia cases at two New York City hospitals identified as part of standard clinical testing procedures in the Mount Sinai Health System (MSHS) Clinical Microbiology Laboratory (CML) during a 2-year period were collected and stored in tryptic soy broth (TSB) with 15% glycerol at  $-80^{\circ}\text{C}$ .

Selected isolates were subcultured on tryptic soy agar (TSA) plates with 5% sheep blood (blood agar) (Thermo Fisher Scientific) under nonselective conditions. The Qiagen DNeasy Blood & Tissue Kit (Qiagen) was used for DNA extraction, as previously described [4]. Following quality control and DNA and library preparation, long-read sequencing was performed on the Pacific Biosciences (PacBio) RS-II platform to a depth of  $>200$ -fold.

### Genome assembly

PacBio long-read sequencing data were assembled using a custom genome assembly and finishing pipeline (<https://github.com/powerpak/pathogendb-pipeline>), as previously described [4]. To assess PathoSPOT performance on more fragmented assemblies typically obtained from short-read sequencing data, we sampled simulated paired-end  $2 \times 150$  bp reads from the PacBio assemblies using InSilicoSeq v1.4.6 [8] using the default hiseq model to a depth of 100-fold. Reads were then assembled using the default settings of shovill v1.1.0 (<https://github.com/tseemann/shovill>), a wrapper for SPAdes [9], and annotated with prokka v1.14.6 [10].

### Comparative genome analysis using PathoSPOT-compare

We developed the PathoSPOT-compare pipeline [11] to perform comparative phylogenomic analysis of annotated genome assemblies for the specific purpose of outbreak detection. The pipeline is implemented as a Rakefile (a Makefile for the Ruby language) that calculates dependencies and executes all necessary subtasks to reach desired outputs. PathoSPOT-compare takes FASTA-formatted genome assemblies as input, along with a relational database (SQLite or MySQL) containing metadata for each assembly (including collection time, location, collection method, organism, and patient ID), as well as metadata on patient admission/discharge/transfer (ADT) history (for spatiotemporal analysis).

Genetic distances for outbreak detection are ultimately calculated by counting single nucleotide variant (SNV) differences within core-genome alignments; however, there is a trade-off between aligning increasingly diverse assemblies and a diminishing core-genome size (as more subsequences will fail to align across all assemblies). Therefore, we implemented a hybrid approach, wherein pairwise distances between all assemblies are first estimated using Mash [12], which uses a k-mer-based hashing approach that approximates average nucleotide identity (ANI). Mash distances are used to perform greedy single-linkage hierarchical pre-clustering, with pre-clusters capped at a pre-specified diameter and size. The default parameters, which are also the parameters used for this study, are a maximum Mash pre-cluster diameter of 0.02 (approximating 98% ANI among all

included genomes) and at most 100 genomes per pre-cluster.

Rapid core-genome alignments are then created for each pre-cluster using *parsnp* [13], which is tailored for intraspecific genome analysis and is therefore well-suited for outbreak analysis. Outputted variant call files (VCF) for each pre-cluster are converted to NumPy arrays (NPZ files) for fast loading and subsetting of variant data by *PathoSPOT-visualize*, the downstream visualization web application that can display called variants alongside phylogenies. The primary output for *PathoSPOT-visualize* is a JSON file containing a matrix of pairwise SNV distances for all genomes (with inter-pre-cluster distances left unspecified) and a maximum-likelihood phylogeny for each pre-cluster. Additional optional pipeline tasks export patient location data (as TSV files) and epidemiological data on positive and negative culture results (as JSON files), both of which are automatically utilized and layered onto the comparative genomic analyses within *PathoSPOT-visualize* when available.

#### Interactive detection and visualization of outbreaks with *PathoSPOT-visualize*

To visualize the analyses depicted in this study, we created the *PathoSPOT-visualize* interactive web application [14]. The application uses PHP scripts and AJAX to serve data from the JSON, TSV, and NPZ output files generated by the *PathoSPOT-compare* pipeline, which are then dynamically mapped to interactive HTML5 and scalable vector graphics (SVG) elements using the D3.js (Data-Driven Documents) framework. All views are rendered in the browser, allowing the user to alter settings that trigger live animated transitions and an intuitive sense of how changes propagate between the linked views of data.

There are three main user interfaces, the “heatmap” tool, the “network map” tool, and the “dendro-timeline” tool. Users initially interact with the “heatmap” tool, which starts with the selection of a dataset that can be prefiltered by specimen location, multi-locus sequence type (MLST), and time interval. The user can dynamically adjust the SNV threshold that specifies the genetic distance deemed indicative of transmission. This threshold is used to perform single-linkage hierarchical clustering of genomes within each MASH pre-cluster on the client-side, with the transmission clusters assigned random colors and depicted on a beeswarm timeline plot and a large heatmap of pairwise distances among all selected genomes. The large heatmap can be swapped for the “network map” view, which plots genomes by their collection location in a geospatial layout, overlaid with density plots of overall epidemiological incidence and force-directed network links depicting genetic relationships.

Suitable thresholds for identifying potential transmission events depend on the organism that is being studied, in particular its mutation rate (which determines the number of expected changes during a given period of time) and the extent of genomic diversity among isolates (which determines the size of the core-genome alignment). Sequencing and/or assembly errors can also introduce additional variability that may need to be taken into account depending on the sequencing technology used. If multiple isolates are available per individual, *PathoSPOT* can aid in the threshold selection process by providing a histogram depicting distributions of pairwise SNV distances among same-patient isolates (which are generally expected to be related) and different-patient isolates (which are not, assuming a low level of transmission). As indicated above, the fraction of the genome considered in the core-genome alignment varies per isolate, depending on the diversity and number of sequences in each MASH pre-cluster. Although the default MASH thresholds were selected to yield comparable pre-cluster core-genome coverage (e.g., 74–84% in this study), care should be taken when comparing SNV thresholds between studies and pre-clusters with large differences in coverage.

Epidemiological links within transmission clusters can be further explored in the “dendro-timeline” tool, which combines a traditional phylogenetic dendrogram with a SNV matrix, a mapping of SNV locations onto a reference assembly, and a pannable-zoomable timeline of patient locations over time, with spatiotemporal overlaps highlighted as bright arcs. The phylogeny for the “dendro-timeline” tool is extracted from the larger maximum-likelihood trees built by *parsnp*, based on the SNV threshold and clustering parameters that the user selected in the “heatmap” tool.

#### Case review

We performed a retrospective clinical chart review on all adult (age > 18) subjects identified with MRSA bacteremia. Analyses were performed in SAS (v9.4) [15]. Variables were initially analyzed individually in a univariate logistic regression model. Variables  $p \leq 0.2$  were then placed into a stepwise multivariate logistic regression model, and only those variables significant at  $p \leq 0.05$  were retained in the final multivariate model [16]. An additional in-depth chart review was performed for subjects implicated in transmission events. These details included location (ward, room, bed), all ADT information, procedures, and shared healthcare workers (HCWs). Whole-genome sequencing was performed on the first patient blood isolate positive for MRSA as part of an ongoing genomic surveillance program as previously described [4]. Hand hygiene is monitored by the Infection Prevention and Control department by the use of anonymous observers using the Joint Commission’s Targeted Solutions Tool (TST) [17], which was

implemented in November 2014. This tool allows staff members to document reasons for non-compliance and target areas of interventions. Hand hygiene observations are collected anonymously at entry and exit by trained staff members in each hospital ward. For this study, hand hygiene compliance was calculated monthly by dividing the total number of compliant observations by the total hand hygiene observations for the time period.

**Software**

The PathoSPOT-compare [11] and PathoSPOT-visualize [14] packages developed for this study are both open source. A live demo of all visualizations created for this study, along with documentation on setting up and using the software with example data from this study, can be found at <https://pathospot.org>.

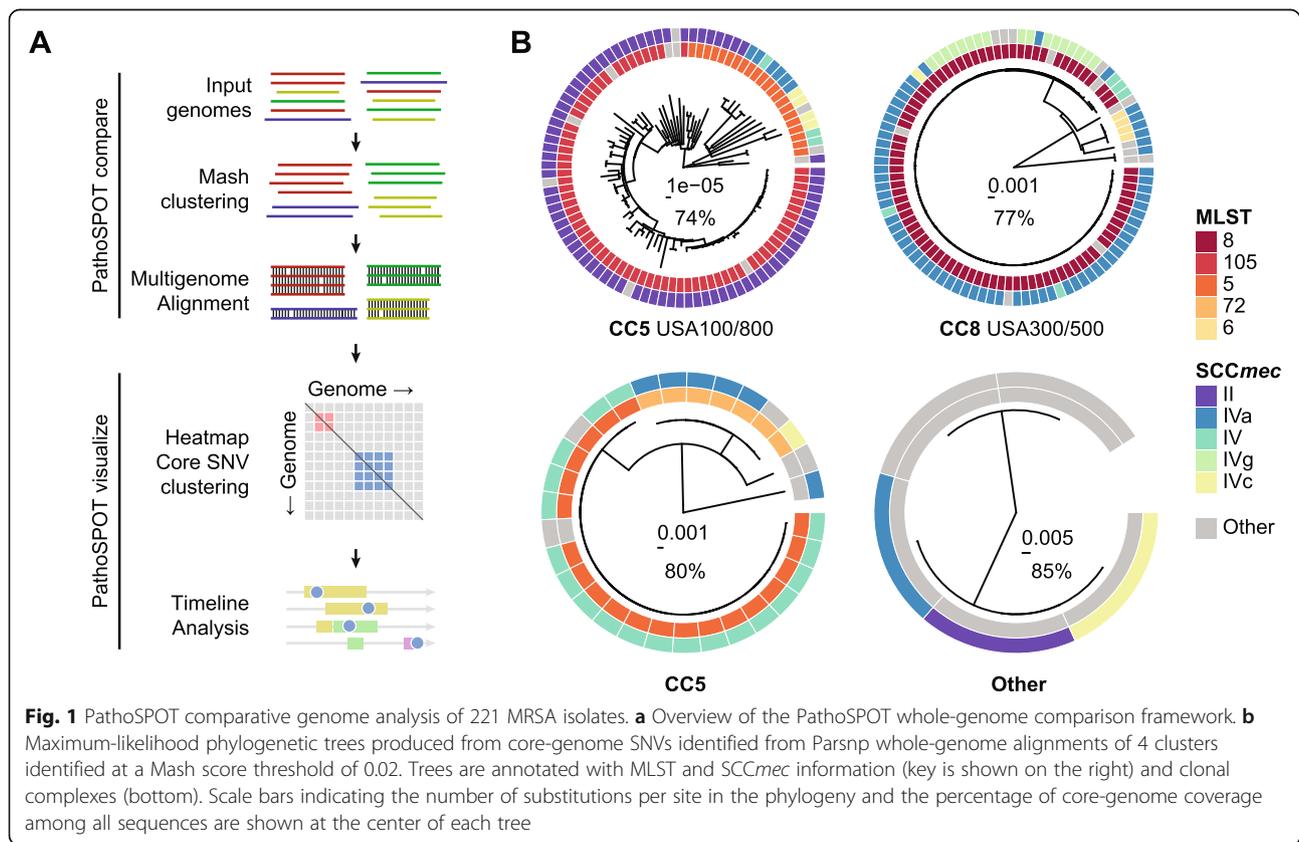
**Results**

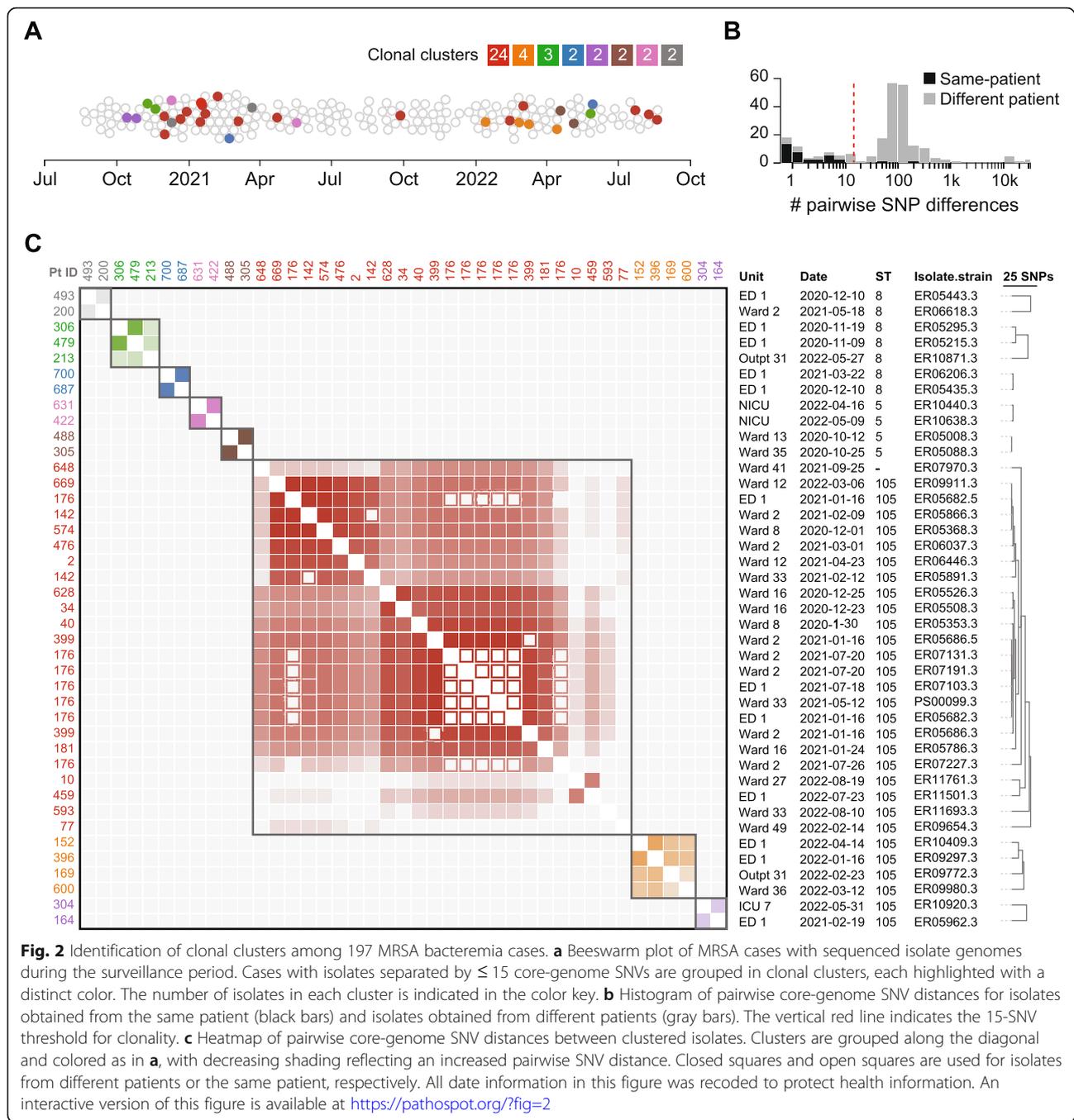
**PathoSPOT surveillance identifies frequent under-the-radar MRSA transmissions**

We developed PathoSPOT to automate comparisons of large numbers of complete or draft microbial genomes, and to rapidly identify closely related isolates indicative of transmission events and map their epidemiological timelines (Fig. 1a). PathoSPOT combines existing tools for whole-genome alignment with custom analysis and

visualization code developed in Ruby, Python, and Javascript (<https://pathospot.org>). To demonstrate its utility, we applied PathoSPOT to MRSA isolates obtained from all bacteremia cases at two hospitals (A and B) during a 2-year period. In total, we sequenced 224 genomes for 221 isolates from 197 patients using PacBio long-read technology and obtained 184 finished-quality and 40 draft genome sequences (Additional file 1: Table S1). In most cases, we only sequenced the primary blood culture, but additional isolates were analyzed for the same patient in cases of prolonged or recurrent infections. We first used the PathoSPOT “compare” pipeline to pre-cluster genomes based on Mash distance [12]. This step groups related genomes prior to multi-genome alignment and avoids the need for manual selection of a reference genome. The Mash distance threshold for MRSA was determined empirically to yield pre-clusters of genomes consistent with known clonal complex assignments based on MLST data derived from each genome [18] and to maximize core-genome alignments (Fig. 1b). Pairwise distances between genomes were then calculated as the number of single nucleotide variants (SNVs) between core-genome alignments in each pre-cluster for further analysis.

To identify transmission events, we used the PathoSPOT “heatmap” visualization (Fig. 2). We set a threshold





of  $\leq 15$  SNVs to identify potential transmissions, based on the extent of intra- and inter-patient variability we previously observed in complete genome analysis of an extended outbreak [4], and considering a core-genome mutation rate of  $\sim 3$  SNVs per Mb per year [2, 19]. The distance threshold can be varied interactively in the heatmap visualization to explore grouping at different levels of relatedness, depending on the pathogen. The linked “network map” provides an accompanying view that plots genomes by their collection location on a building map

(Additional file 2: Fig. S1). At the selected threshold, we identified 8 clonal clusters with a total of 33 patients, implicating 16.8% of surveilled patients in transmission events (Fig. 2c). Most clusters consisted of patient pairs (5/8), but there were 3 with more than two patients. Patients within each cluster typically had overlapping hospital visits (75%) and stayed in the same wards at some point during these visits (63%), but in many cases, MRSA bacteremia was only found after they transferred to different wards. This likely contributed to the fact that none of

the clusters could be recognized epidemiologically. The most striking example of this was a cluster of 24 isolates from 16 patients that were collected over a period of 21 months from 9 different wards.

To demonstrate the ability of PathoSPOT to analyze draft as well as complete genomes, we repeated the same analysis after sampling and assembling a short-read dataset from each PacBio genome. Despite obtaining much more fragmented genomes with an average of 160 contigs and N50 of 235 kb across all isolates (Additional file 2: Fig. S2A-C), PathoSPOT analysis of the short-read assemblies produced identical clonal clusters (Additional file 2: Fig. S2D).

### PathoSPOT timeline highlights the role of colonization in prolonged MRSA outbreaks

The presence of a clonal MRSA cluster among 16 bacteremia cases was consistent with a prolonged “under-the-radar” outbreak. The outbreak strain matched the hospital-associated USA100 lineage (*spa* type t002, MLST 105, staphylococcal cassette chromosome *mec* type II) and was resistant to fluoroquinolones, oxacillin, clindamycin, erythromycin, and gentamicin. We next used the PathoSPOT “dendro-timeline” tool, which combines phylogenetic analysis of outbreak isolates with the ADT history for each patient, to analyze this outbreak in more detail (Fig. 3). The core-genome dendrogram, derived from the multi-genome alignment of the superset of isolates in the same Mash pre-cluster, indicated the presence of subclades with distinct shared variant patterns that were consistent with sub-transmissions within the larger outbreak (Fig. 3a). Isolates from patient 176 (p176), who tested positive for MRSA bacteremia numerous times within a span of 6 months, were represented in distinct subclades, suggesting that the patient carried distinct variants of the outbreak strain at the same time. This intra-host variation was confirmed by sequencing two subclones from p176 isolate ER05682 (Fig. 3a, triangle and rhombus).

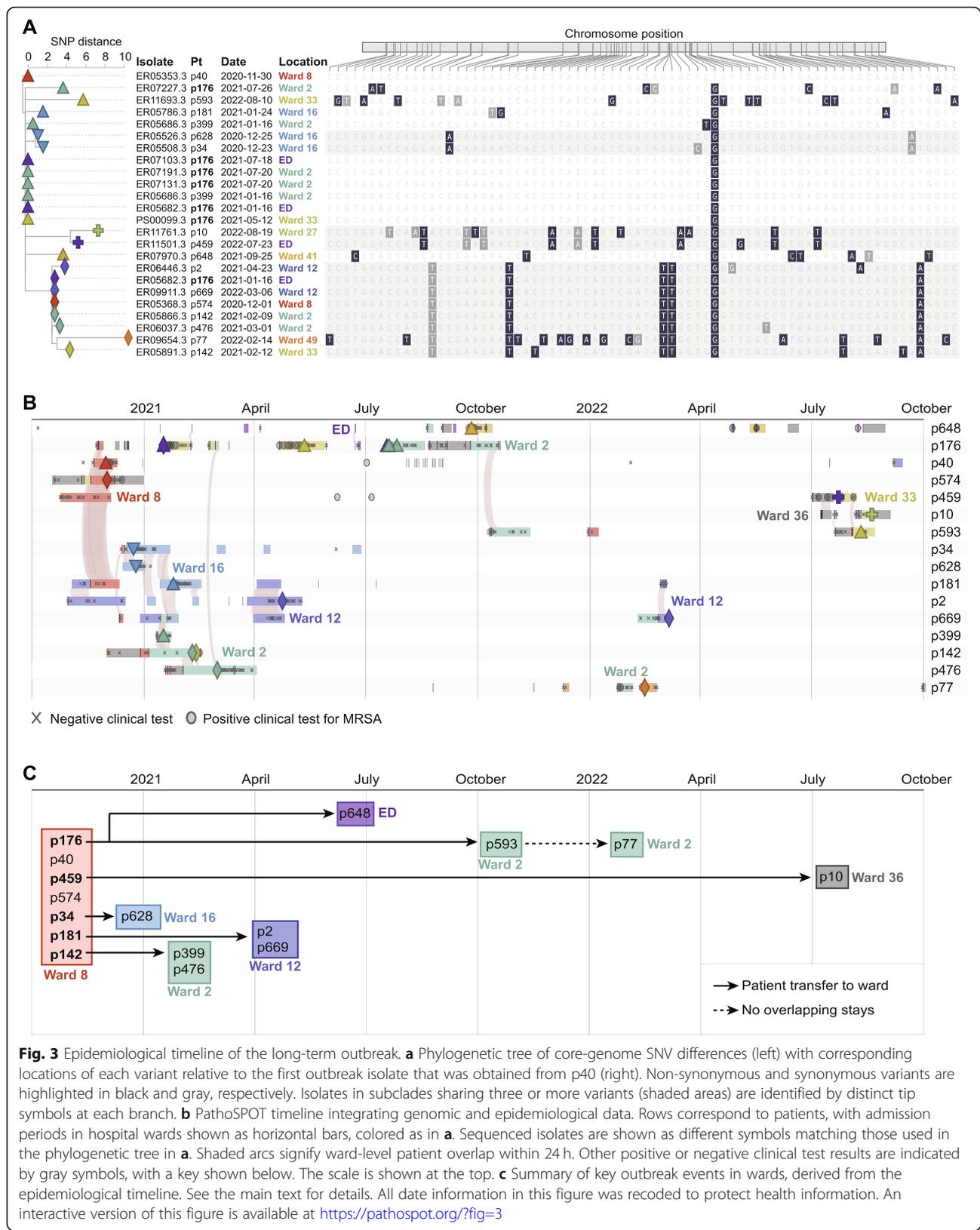
Based on the PathoSPOT timeline of events, we reconstructed patient contacts based on location overlaps (Fig. 3b) and inferred the most likely outbreak scenario (Fig. 3c). The first two patients (p40 and p574) tested positive on ward 8. No other positives were found on this ward, but five other patients had overlapping stays (p176, p459, and p181) or were admitted to the same ward within 4 weeks of the first positive test (p34, p142). All but one patient tested positive for bacteremia within 7 weeks of their stay in ward 8, following transfers or readmissions to other wards. Strikingly, p459, who was discharged from ward 8 four days after the initial positive case, did not present with bacteremia until 20 months later. In the intervening period, this patient had no contact with our health system except for two dermatology office visits where positive wound cultures for

MRSA were obtained. The degree of genetic drift of the p459 isolate (11 SNVs) and pattern of positive wound cultures prior to bacteremia are consistent with long-term colonization after initial exposure in ward 8, although we could not verify this scenario as the wound isolates were not available for sequencing.

Five additional patients tested positive in wards 16, 2, and 12 during the first 6 months of the outbreak (Fig. 3b, c). Each instance was preceded by the transfer of a patient that had previously stayed in ward 8, suggesting that direct or indirect transmissions from these cases propagated the outbreak. Notably, p34 was transferred from ward 8 to ward 16, into a room neighboring p628, who became bacteremic 2 days later. Both their isolates were grouped in the same subclade (Fig. 3a, inverted triangle). Likewise, p142 was transferred from ward 8 to ward 2, where there was overlap with p399 and p476 before all three became bacteremic on this ward. Notably, p142 and p176 overlapped with p476 on two different days in the inpatient hemodialysis unit, providing an alternative acquisition route. Finally, after discharge from ward 8, p181 was readmitted to ward 12, where the patient overlapped stays with patients p2 and p669.

Four late transmission events were identified in months 7 to 21. Two of these events involved patient p176, who tested positive for the outbreak strain on multiple occasions during readmissions. Patient p176 visited the emergency department (ED) on the same day as patient p648 and had an overlapping stay in ward 2 with p593 for at least 5 days, in the months prior to their positive tests. Following readmission after a 20-month hiatus, p459 likely transmitted to p10, based on evidence of an overlapping stay in ward 36 (Fig. 3b) and the high relatedness of their isolates (Fig. 3a, plus sign). Patient p77 was the only person that did not have overlapping stays with other outbreak cases. The patient had a total of two pediatric (ward 49) and one adult (ward 2) admission to the hospital within 21 months. Given that all other outbreak cases were adults, we consider ward 2 the most likely location of MRSA acquisition, where p77 shared HCWs with p593 who was admitted to the same unit 11 weeks before.

Altogether, PathoSPOT analysis suggested that initial exposure in ward 8 resulted in colonization and subsequent clinical infection of 7 patients (44% of the prolonged outbreak cluster), followed by secondary transmissions after ward transfers and/or readmission of these initial cases. An alternative scenario of community transmissions was discounted after mapping of home zip codes, which indicated that 13 of 16 cases lived in geographically distinct neighborhoods. Spatiotemporal analyses of the seven smaller clusters (Additional file 2: Fig. S3) showed that five included direct overlaps, of which three are plausible transmission events, and two



such events occurred months before the clonal blood cultures were obtained.

#### **Hand hygiene compliance and vascular access implicated in under-the-radar outbreak**

As the outbreak extended over multiple wards, we further investigated hand hygiene rates, shared HCWs, patient movements, and clinical characteristics. Average hand hygiene compliance in affected wards ranged between 79 and 83% per month. Compliance in wards 8 and 16 decreased to 70% and 66%, respectively, in the month prior to the first outbreak case, while ward 2 compliance was maintained at 79%. All outbreak patients shared at least one HCW involved in the care of other patients in the cluster. This is consistent with the high degree of overlapping stays in the same ward and suggests that direct and indirect transmissions may have played a role in propagating the outbreak. Although outbreak cases were moved frequently between units based on transfer records, they did not move more frequently than non-outbreak patients.

Chart review of outbreak cases revealed that 69% ( $n = 11$ ) were male, 75% ( $n = 12$ ) had been admitted from home, and 63% ( $n = 10$ ) had a hospital admission in the 90 days prior (Table 1). Seventy-five percent ( $n = 12$ ) were considered hospital-onset (HO)-MRSA as defined by the National Healthcare Safety Network (NHSN) [20], and 88% ( $n = 14$ ) of subjects had an invasive device at the time of infection. Univariate and multivariate analyses of the 16 outbreak cases compared to 34 patients infected with non-outbreak MLST 105 MRSA showed that outbreak cases were significantly associated with HO-MRSA, as well as intravenous chemotherapy prior to bacteremia (Table 1). Five additional variables with  $p \leq 0.2$  were included in the multivariate stepwise regression model, but only variables significant at  $p \leq 0.05$  were retained in the final model. Vascular access, while significant ( $p \leq 0.05$ ) in the univariate model, was not included in the multivariate model due to collinearity issues with receiving cancer treatment. Notable was the presence of active malignancy (57%;  $n = 8$ ) including leukemia ( $n = 5$ ), multiple myeloma ( $n = 1$ ), disseminated Kaposi sarcoma ( $n = 1$ ), and metastatic breast cancer ( $n = 1$ ). Among patients with hematologic malignancies, three had undergone hematopoietic stem cell transplant. Consistent with these findings, the most common presumed source of bacteremia was vascular access ( $n = 9$ ; 56%), followed by skin source ( $n = 4$ ; 25%). The 90-day mortality incidence was 25% ( $n = 4$ ), of which 75% ( $n = 3$ ) was related to bacteremia with the outbreak strain. There were no differences in outcomes between outbreak and non-outbreak patients.

#### **Discussion**

We developed PathoSPOT as a key component of an ongoing genomics-based pathogen surveillance program to facilitate the detection of outbreaks and transmissions. Application of the toolkit to surveillance data from 197 patients with MRSA bacteremia over a 2-year period demonstrates the utility of our toolkit and shows that nosocomial transmissions are important sources of morbidity and mortality. We find that in the absence of genomic surveillance many nosocomial transmissions of MRSA go undetected by standard infection prevention practices, as they only result in clinically apparent infections weeks to months later.

An outbreak among 16 patients from distinct adult medicine wards spanned nearly the entire study period. Reconstruction of the epidemiological timeline with PathoSPOT suggests that this outbreak started with the exposure of 7 patients in a single ward. Subsequent transfers or readmissions of these patients to other wards were a key factor in propagating the outbreak across the hospital. Additional contributing factors may have included shared HCWs and reduced hand hygiene rates surrounding key outbreak events. Frequent room changes within and between wards may also have resulted in contaminated environmental surfaces, which has been shown to play a role in nosocomial transmissions [21, 22]. In the absence of routine patient, HCW, and environmental screening, it was not possible to determine the nature of the initial exposure event.

Our study provides additional support for a role of colonization in the persistence and delayed progression of under-the-radar outbreaks [23, 24]. Skin colonization in particular may have contributed to subsequent infections, as vascular access was significantly associated as the presumed source of bacteremia among outbreak patients. The number of outbreak patients with hematological malignancies and bone marrow suppression was also notable in this respect. These patients are at an increased risk of bacteremia, as central venous catheters remain an essential tool for their treatment, frequently leading to catheter-related infections [25].

The detection of nosocomial transmissions and outbreaks is critical for healthcare organizations, and our findings have important ramifications for increasing the effectiveness of infection prevention programs. It is currently standard practice in most healthcare settings to monitor the rate of positive clinical cultures across sites and wards for changes relative to baseline occurrences. Outbreak investigations are typically only initiated if there is a notable uptick in cases at a particular location within a defined period of time, or when specific concerns are raised by hospital staff. This approach is reactive in nature and in practice means that nosocomial outbreaks are often only recognized after they have

**Table 1** Outbreak patients vs. non-outbreak patients with MLST 105 isolates

Factor	Outbreak patients N = 16 (%)	Non-outbreak patients N = 34 (%)	Univariate analysis		Multivariate analysis	
			OR (95% CI)	p value	OR (95% CI)	p value
Male	11 (69)	21 (62)	1.36 (0.39–4.82)	0.63		
<i>Race/ethnicity</i>						
Non-Hispanic White	5 (31)	14 (41)	Reference			
Non-Hispanic Black	3 (19)	9 (26)	0.93 (0.18–4.90)	0.94		
Hispanic/Latino/Asian	5 (31)	6 (18)	2.33 (0.49–11.17)	0.29		
Unknown	3 (19)	5 (15)	1.68 (0.29–9.75)	0.56		
<i>Age at time of infection</i>						
18–54 years	6 (38)	7 (21)	Reference			
55–69 years	5 (31)	9 (26)	0.65 (0.14–3.04)	0.58		
≥ 70 years	5 (31)	18 (53)	0.32 (0.07–1.41)	0.13		
History of IV drug use	2 (13)	2 (6)	2.29 (0.29–17.90)	0.43		
HIV	1 (6)	3 (9)	0.69 (0.07–7.19)	0.76		
<i>Admission source</i>						
Home	12 (75)	17 (50)	Reference			
NH/Rehab/LTACH	2 (13)	11 (32)	0.26 (0.05–1.38)	0.11		
Other hospitals	2 (13)	6 (18)	0.47 (0.08–2.75)	0.40		
Prior hospital admission (90 days)	10 (63)	23 (68)	0.80 (0.23–2.76)	0.72		
<i>NHSN definitions</i>						
CO-MRSA	4 (25)	24 (71)	Reference		Reference	
HO-MRSA	12 (75)	10 (29)	<b>7.20 (1.87–27.79)</b>	<b>0.004</b>	<b>5.20 (1.04–26.01)</b>	<b>0.04</b>
Presence of invasive device <sup>A</sup>	14 (88)	27 (90)	0.78 (0.12–5.21)	0.80		
Receiving cancer treatment <sup>B</sup>	7 (44)	2 (6)	<b>15.75 (1.75–141.39)</b>	<b>0.01</b>	<b>11.24 (1.72–73.28)</b>	<b>0.01</b>
<i>Charlson Comorbidity Index (CCI)</i>						
0–3	4 (25)	9 (26)	Reference			
≥ 4	12 (75)	25 (74)	1.08 (0.28–4.23)	0.91		
History of MRSA colonization	6 (38)	16 (47)	0.68 (0.20–2.28)	0.53		
<i>Presumed source of MRSA BSI</i>						
Skin source <sup>C</sup>	4 (25)	7 (21)	1.29 (0.32–5.24)	0.73		
Pneumonia	1 (6)	6 (18)	0.31 (0.03–2.83)	0.30		
Vascular access <sup>*†</sup>	9 (56)	9 (26)	<b>3.57 (1.03–12.43)</b>	<b>0.05</b>		
Others/undetermined	2 (13)	12 (35)	0.26 (0.05–1.35)	0.11		
Persistent bacteremia (≥ 5 days)	2 (13)	9 (26)	0.40 (0.08–2.10)	0.28		
ICU admission prior to BSI	4 (25)	3 (9)	3.44 (0.67–17.73)	0.14		
Intubated prior to MRSA BSI	4 (25)	2 (6)	5.33 (0.86–33.00)	0.07		

Bold indicates significance at  $\leq 0.05$

Abbreviations: BSI bloodstream infection, HIV human immunodeficiency virus, ICU intensive care unit, IV intravenous, NHSN National Healthcare Safety Network

<sup>A</sup>Includes devices such as a pacemaker, any vascular access, orthopedic hardware, foley catheter, arteriovenous graft placement, percutaneous endoscopic gastronomy (PEG), ostomy, or any type of urinary collection at the time of first positive bloodstream infection

<sup>B</sup>Includes patients actively receiving cancer treatment through a central venous catheter prior to bacteremia in the outpatient or inpatient setting

<sup>C</sup>Skin source includes skin and soft tissue infections, thrombophlebitis due to peripheral IV catheters

\*Variable not included in the multivariate analysis in order to prevent collinearity between receiving cancer treatment and vascular access

<sup>†</sup>Vascular access devices include a non-tunneled central venous catheter, tunneled catheter (hickman or permacath), implanted port, peripherally inserted central catheter (PICC line), and arteriovenous graft (AVG) and fistula (AVF)

escalated to a sufficiently large group of patients to alter baseline rates. Moreover, as evidenced by our findings, delays between exposure and subsequent clinical infection can obscure even large outbreaks from epidemiological detection.

## Conclusions

Routine health-system-wide monitoring using precise, genomics-based, pathogen surveillance programs supported by rapid analysis frameworks is essential for timely detection of events that are not readily ascertained by conventional epidemiological approaches. Widespread adoption of such programs depends on the availability of accessible tools such as PathoSPOT that can be used by infection prevention staff in near real time without the need for extensive training. The effectiveness of such programs can be further increased when implemented across regional health systems, long-term acute care hospitals, and skilled nursing facilities, to track the dissemination of strains and identify sources and at-risk patients based on contact networks. When combined with timely intervention, these efforts may be of critical importance to help break transmission chains and reduce endemic rates of nosocomial infections.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13073-020-00798-3>.

**Additional file 1.** This file contains Supplementary Table S1, which lists the bacterial genome assembly details.

**Additional file 2: Fig. S1.** Network layout view of clonal clusters. **Fig. S2.** Comparison of PathoSPOT analysis of long-read and short-read assemblies. **Fig. S3.** Other clonal clusters identified by PathoSPOT analysis.

## Acknowledgements

Not applicable.

## Authors' contributions

G.P. provided clinical samples for the study. A.B.C., S.C., G.P., and D.R.A. performed clinical evaluations. C.B., G.P., D.R.A., and H.B. were involved in clinical sample accessioning. C.B. and M.C. performed culturing and DNA extraction. I.O. performed NGS experiments. I.O., R.S., and M.S. provided NGS services. T.R.P., A.O., M.J.S., and H.B. developed the PathoSPOT pipeline and performed the genome assembly. A.B.C., T.R.P., A.C.D., L.F., A.M., B.C., A.K., D.R.A., and H.B. performed analyses of electronic medical records. A.B.C., T.R.P., A.O., A.C.D., K.C., B.C., G.P., A.K., M.J.S., D.R.A., and H.B. analyzed, interpreted, or discussed the data. A.B.C., T.R.P., A.C.D., D.R.A., and H.B. wrote the manuscript. A.B.C., T.R.P., A.K., D.R.A., and H.B. conceived the study. D.R.A. and H.B. supervised the study. T.R.P., A.K., D.R.A., and H.B. raised financial support. All authors read and approved the final manuscript.

## Funding

This research was supported by R01 AI119145 (H.v.B.), the Icahn Institute for Genomics and Data Science (A.K.), the CTSA/NCATS KL2 Program (KL2TR001435, Icahn School of Medicine at Mount Sinai; D.R.A.), and the New York State Department of Health Empire Clinical Research Investigator Program (awarded to Judith A. Aberg, Icahn School of Medicine at Mount Sinai; D.R.A.) and F30 AI122673 (T.R.P.). The research reported in this paper was supported by the Office of Research Infrastructure of the National

Institutes of Health (NIH) under award numbers S10OD018522 and S10OD026880 as well as institutional funds. The funding sources had no influence on the design of the study and collection, analysis and interpretation of data, and writing of the manuscript.

## Availability of data and materials

Genome assemblies have been deposited in Genbank (see Additional file 1: Table S1 for accession numbers). All study data is also available at <https://pathospot.org>. PathoSPOT-compare is available from <https://doi.org/10.5281/ZENODO.4142966> [11] and PathoSPOT-visualize from <https://zenodo.org/record/4149962> [14].

## Ethics approval and consent to participate

The study protocols were reviewed and approved by the Icahn School of Medicine at Mount Sinai Institutional Review Board for the collection and bacterial genome sequencing of discarded clinical bacteremia specimens by the Pathogen Surveillance Program (protocol HS# 13-00981) and chart reviews of outbreak cases and non-outbreak controls (protocol HS# 17-02246), as defined by DHHS regulations. A waiver of authorization for use and disclosure of protected health information (PHI) and a waiver of informed consent were approved for both protocols based on the criteria that the use or disclosure of PHI involved no more than minimal risk to the privacy of individuals and because the research could not practically be conducted without the waiver and without access to and use of the PHI. The research conformed to the principles of the Helsinki Declaration.

## Consent for publication

Not applicable.

## Competing interests

Robert Sebra is VP of Technology Development and a stockholder at Sema4, a Mount Sinai Venture. This work, however, was conducted solely at the Icahn School of Medicine at Mount Sinai. The remaining authors declare that they have no competing interests.

## Author details

<sup>1</sup>Department of Medicine, Division of Infectious Diseases, Icahn School of Medicine at Mount Sinai, New York City, NY, USA. <sup>2</sup>Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York City, NY, USA. <sup>3</sup>Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. <sup>4</sup>Black Family Stem Cell Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. <sup>5</sup>Sema4, a Mount Sinai venture, Stamford, CT 06902, USA. <sup>6</sup>Infection Prevention, The Mount Sinai Hospital, New York City, NY, USA. <sup>7</sup>Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York City, NY, USA.

Received: 4 June 2020 Accepted: 2 November 2020

Published online: 16 November 2020

## References

1. Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, et al. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. *BMJ Open*. 2012; 2 Available from: <https://doi.org/10.1136/bmjopen-2012-001124>.
2. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. *Science*. 2010;327:469–74.
3. Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med*. 2012;366:2267–75.
4. Sullivan MJ, Altman DR, Chacko KI, Ciferri B, Webster E, Pak TR, et al. A complete genome screening program of clinical methicillin-resistant *Staphylococcus aureus* isolates identifies the origin and progression of a neonatal intensive care unit outbreak. *J Clin Microbiol*. 2019;57 Available from: <https://doi.org/10.1128/JCM.01261-19>.
5. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018;34: 4121–3.

6. Zhou Z, Alikhan N-F, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res.* 2018;28:1395–404.
7. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* 2019;29:304–16.
8. Gourelé H, Karlsson-Lindsjö O, Hayer J, Bongcam-Rudloff E. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics.* 2019;35:521–2.
9. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19:455–77.
10. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30:2068–9.
11. Pak T, Webster E, Mjsull, Van Bakel H. powerpak/pathspot-compare: v1.0. Zenodo; 2020. Available from: <https://doi.org/10.5281/ZENODO.4142966>.
12. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 2016;17:132.
13. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 2014;15:524.
14. Pak T. powerpak/pathspot-visualize: v1.0. 2020. Available from: <https://zenodo.org/record/4149962>.
15. SAS Institute. Base SAS 9.4 procedures guide: Statistical Procedures, 2nd Edition. 2013.
16. Dupper AC, Sullivan MJ, Chacko KI, Mishkin A, Ciferri B, Kumaresh A, et al. Blurred molecular epidemiological lines between the two dominant methicillin-resistant *Staphylococcus aureus* clones. *Open Forum Infectious Diseases.* 2019;6. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6735859/>.
17. Anderson R, Rosenberg A, Garg S, Nahass J, Nenos A, Egorova N, et al. Establishing the foundation to support health system quality improvement: using a hand hygiene initiative to define the process. *J Patient Saf.* 2019; Available from: <https://doi.org/10.1097/PTS.0000000000000578>.
18. Jolley KA, Maiden MCJ. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics.* 2010;11:595.
19. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, et al. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc Natl Acad Sci U S A.* 2012;109:4550–5.
20. Center for Disease Control and Prevention. Multidrug-resistant organism & Clostridioides difficile infection (MDRO/CDI) module. 2018.
21. Otter JA, Yezli S, French GL. The role played by contaminated surfaces in the transmission of nosocomial pathogens. *Infect Control Hosp Epidemiol.* 2011;32:687–99.
22. Otter JA, Yezli S, Salkeld JAG, French GL. Evidence that contaminated surfaces contribute to the transmission of hospital pathogens and an overview of strategies to address contaminated surfaces in hospital settings. *Am J Infect Control.* 2013;41:56–11.
23. Senn L, Clerc O, Zanetti G, Basset P, Prod'homme G, Gordon NC, et al. The stealthy superbug: the role of asymptomatic enteric carriage in maintaining a long-term hospital outbreak of ST228 methicillin-resistant *Staphylococcus aureus*. *MBio.* 2016;7:e02039–e02015.
24. Merrer J, Santoli F, Vecchi CA-D, Tran B, De Jonghe B, Outin H. "Colonization pressure" and risk of acquisition of methicillin-resistant *Staphylococcus aureus* in a medical intensive care unit. *Infect Control Hosp Epidemiol.* 2000; 21:718–23.
25. Zakhour R, Chaftari A-M, Raad II. Catheter-related infections in patients with haematological malignancies: novel preventive and therapeutic strategies. *Lancet Infect Dis.* 2016;16:e241–50.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

