Author/s:
Gorrie, CL;Da Silva, AG;Ingle, DJ;Higgs, C;Seemann, T;Stinear, TP;Williamson, DA;Kwong, JC;Grayson, ML;Sherry, NL;Howden, BP

Title:
Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis

Date:
2021-11-01

Citation:
Gorrie, C. L., Da Silva, A. G., Ingle, D. J., Higgs, C., Seemann, T., Stinear, T. P., Williamson, D. A., Kwong, J. C., Grayson, M. L., Sherry, N. L. & Howden, B. P. (2021). Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis. Lancet Microbe, 2 (11), pp.e575-e583. https://doi.org/10.1016/S2666-5247(21)00149-X.

Persistent Link:
https://hdl.handle.net/11343/283344

Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis

Claire L Gorrie, Anders Gonçalves Da Silva, Danielle J Ingle, Charlie Higgs, Torsten Seemann, Timothy P Stinear, Deborah A Williamson, Jason C Kwong, M Lindsay Grayson, Norelle L Sherry, Benjamin P Howden

Summary

Background Pairwise single nucleotide polymorphisms (SNPs) are a cornerstone of genomic approaches to the inference of transmission of multidrug-resistant (MDR) organisms in hospitals. However, the impact of many key analytical approaches on these inferences has not yet been systematically assessed. This study aims to make such a systematic assessment.

Methods We conducted a 15-month prospective study (2-month pilot phase, 13-month implementation phase), across four hospital networks including eight hospitals in Melbourne, VIC, Australia. Patient clinical and screening samples containing one or more isolates of meticillin-resistant *Staphylococcus aureus*, vancomycin-resistant *Enterococcus faecium*, and extended-spectrum β-lactamase-producing *Escherichia coli* and *Klebsiella pneumoniae* were collected and underwent whole genome sequencing. Using the genome data from the top four most numerous sequence types from each species, 16 in total, we systematically assessed the: (1) impact of sample and reference genome diversity through multiple core genome alignments using different data subsets and reference genomes, (2) effect of masking of prophage and regions of recombination in the core genome alignments by assessing SNP distances before and after masking, (3) differences between a cumulative versus a 3-month sliding-window approach to sample genome inclusion in the dataset over time, and (4) the comparative effects each of these approaches had when applying a previously defined SNP threshold for inferring likely transmission.

Findings 2275 samples were collected (397 during the pilot phase from April 4 to June 18, 2017; 1878 during the implementation phase from Oct 30, 2017, to Nov 30, 2018) from 1870 patients. Of these 2275 samples, 1537 were identified as arising from the four most numerous sequence types from each of the four target species of MDR organisms in this dataset (16 sequence types in total: *S aureus* ST5, ST22, ST45, and ST93; *E faecium* ST80, ST203, ST1421, and ST1424; *K pneumoniae* ST15, ST17, ST307, and ST323; and *E coli* ST38, ST131, ST648, and ST1193). Across the species, using a reference genome of the same sequence type provided a greater degree of pairwise SNP resolution, compared with species and outgroup-reference alignments that mostly resulted in inflated SNP distances and the possibility of missed transmission events. Omitting prophage regions had minimal effect; however, omitting recombination regions had a highly variable effect, often inflating the number of closely related pairs. Estimated SNP distances between isolate pairs over time were more consistent using a sliding-window than a cumulative approach.

Interpretation We propose that the use of a closely related reference genome, without masking of prophage or recombination regions, and of a sliding-window approach for isolate inclusion is best for accurate and consistent MDR organism transmission inference, when using core genome alignments and SNP thresholds. These approaches provide increased stability and resolution, so SNP thresholds can be more reliably applied for putative transmission inference among diverse MDR organisms, reducing the chance of incorrectly inferring the presence or absence of close genetic relatedness and, therefore, transmission. The establishment of a broadly applicable and standardised approach, as proposed here, is necessary to implement widespread prospective genomic surveillance for MDR organism transmission.

Funding Melbourne Genomics Health Alliance, and National Health and Medical Research Council of Australia.

Introduction

Antimicrobial-resistant (AMR) pathogens are among the foremost threats to global public health.[1–4] AMR infections, and resulting sequelae, lead to increased morbidity, mortality, and considerable increases in treatment costs and length of hospital stay.[5–7] This situation places an increasing strain on health-care systems as the global burden of AMR pathogens rises.[5,7–9] Among the pathogens of particular concern are multidrug-resistant (MDR) organism species such as

Lancet Microbe 2021

Published Online August 6, 2021 https://doi.org/10.1016/S2666-5247(21)00149-X

See Online/Comment https://doi.org/10.1016/S2666-5247(21)00183-X

Microbiological Diagnostic Unit, Public Health Laboratory (C L Gorrie PhD, A Goncalves Da Silva PhD, T Seemann PhD, Prof D A Williamson PhD, N L Sherry MBBS, Prof B P Howden PhD) **and Department of Microbiology & Immunology** (C L Gorrie, D J Ingle PhD, C Higgs BSc, T Seemann, Prof T P Stinear PhD, Prof D A Williamson, J C Kwong PhD, N L Sherry, Prof B P Howden), **Peter Doherty Institute for Infection & Immunity, University of Melbourne, Melbourne, VIC, Australia; National Centre for Epidemiology & Population Health, Australian National University, Canberra, ACT, Australia** (D J Ingle); **Department of Microbiology, Royal Melbourne Hospital, Melbourne, VIC, Australia** (Prof D A Williamson); **Department of Infectious Diseases and Department of Microbiology, Austin Health, Heidelberg, VIC, Australia** (J C Kwong, M L Grayson MD, N L Sherry, Prof B P Howden); **Department of Medicine, Austin Health, University of Melbourne, Heidelberg, VIC, Australia** (M L Grayson)

Correspondence to: Prof Benjamin Howden, Microbiological Diagnostic Unit, Peter Doherty Institute for Infection & Immunity, University of Melbourne, Melbourne, VIC 3000, Australia **bhowden@unimelb.edu.au**

Research in context

Evidence before this study

We searched PubMed for studies using the search terms: "genomics", "antimicrobial", "resistance", "hospital", and "transmission", combined with each of "*Staphylococcus aureus*" (n=100 articles), "*Enterococcus faecium*" (n=42), "*Klebsiella pneumoniae*" (n=111), and "*Escherichia coli*" (n=123), without date or language restrictions. We considered all studies from this search that used genomics to infer likely transmission of multidrug-resistant (MDR) organisms in the hospital environment, published before Aug 31, 2020. Despite using genomic data and bioinformatic methods to infer transmission, none of these studies systematically and comprehensively assessed the impacts that these approaches would have on genetic relatedness, particularly when considering a real-time and evolving dataset. To successfully implement genomics-based transmission surveillance for these common MDR organisms, standardised approaches must be methodically and systematically tested across multiple species.

Added value of this study

To the best of our knowledge, this study is the first to systematically quantify the effects of multiple genomic analysis approaches on a diverse collection of key MDR organisms to identify the impact on the level of genomic resolution provided, the number of putative transmissions, and the stability of these measures over time and with growing datasets. We assessed these effects for 16 sequence types, four from each of the common MDR organism species: meticillin-resistant *S aureus*, vancomycin-resistant *E faecium*, and extended-spectrum β-lactamase-producing *K pneumoniae* and *E coli*. These pathogens represented the dominant MDR organism genotypes observed in Australian hospitals. We systematically assessed the effects of (1) reference genome selection and overall diversity of sample data, (2) omission of detected prophage and recombination regions, and (3) sample inclusion or exclusion over extended time periods with increasing sample numbers. We quantified the impacts that these methodological choices had on calculating sample relatedness and therefore on transmission inference. We provided recommendations for standardised approaches to transmission surveillance in real time, for a diverse range of species and sequence types.

Implications of all the available evidence

This study addressed the lack of guidelines and standardisation in prospective workflows for whole genome sequencing surveillance of transmission of MDR organisms. Through systematic assessment of approaches used variably in existing analyses we have been able to propose guidelines for future surveillance analyses, as well as to show the effect type and extent for each step of these analyses. Future implementations of genomics for the surveillance of MDR organism transmission in hospitals can now draw from these standardised approaches rather than create novel bespoke analyses with each new investigation or dataset.

the ESKAPE pathogens (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumanii, Pseudomonas aeruginosa,* and *Enterobacter aerogenes*), and *Escherichia coli*.[4,10,11]

WHO recently highlighted the need to invest in resources to enhance the surveillance of AMR,[3,12] which can be facilitated through genomics.[3,13] Although whole genome sequencing (WGS) is increasingly used in investigations of public health outbreaks, these have predominantly focused on retrospective, closed, datasets. In these reports, study-specific analysis approaches have defined single nucleotide polymorphism (SNP) thresholds for ruling isolates as likely or unlikely parts of transmission events, based on a combination of genomic and epidemiological evidence,[14–17] and some have determined thresholds of genomic diversity between sequences that are correlated with epidemiological transmission evidence (eg, SNP distance).[14,15,17] Although these SNP thresholds perform well in a closed dataset, their application to prospective genomic surveillance datasets, with different analysis approaches, needs to be evaluated and developed further, especially when dealing with more genetically complex or temporally diverse datasets.

Many of the MDR organisms posing the greatest health threats exhibit considerable population genomic diversity, prolonged asymptomatic colonisation, horizontal gene transfer, and DNA acquired via homologous recombination. These factors can impact relative genetic relatedness, so the methods and transmission SNP thresholds used must remain robust among such genomic dynamism. A substantial gap remains between bespoke comparative genomics research approaches applied in retrospective studies, and the effective translation of such approaches into real-time surveillance in clinical settings. To address this knowledge gap, we investigated three key approaches to genomics data analysis that could substantially impact the accuracy of surveillance and transmission detection, using a comprehensive genomic and epidemiological dataset for four major hospital-associated MDR organisms. These approaches were: (1) reference genome choice and level of analysis (ie, species *vs* sequence type); (2) omission of DNA regions predicted to be prophage or acquired by recombination; and (3) genome inclusion or exclusion in a growing dataset (a cumulative *vs* sliding-window approach).

Methods

Study design, isolate selection, quality control, typing, and WGS

During our 15-month prospective study (including an 8-week pilot phase [April 4 to June 18, 2017][18] and a 13-month implementation phase [Oct 30, 2017, to Nov 30, 2018]) all positive clinical or screening samples for four dominant health-care-associated MDR organisms

were collected for WGS from eight hospitals in Melbourne (VIC, Australia). The samples included all meticillin-resistant *S aureus,* all *vanA*-positive vancomycin-resistant *E faecium*, all extended-spectrum β-lactamase (ESBL) phenotype *K pneumoniae,* and all ESBL ciprofloxacin-resistant *E coli* . All isolates underwent WGS, on the Illumina NextSeq platform (Illumina, San Diego, CA, USA), and quality control checks as part of standard laboratory workflow (appendix 1 p 2).[18] In-silico species confirmation, multi-locus sequence typing, and acquired AMR-resistant gene detection were also done. To capture diversity within each species, and to focus on the dominant genotypes, we selected all sequences representing the four most common sequence types of each species (appendix 1 p 2). Short read sequence data are available from the Sequence Read Archive under BioProject 565795. This study was approved by the Melbourne Health Human Research Ethics Committee (HREC) and endorsed by the corresponding HREC at each participating site.

Mapping and SNP calling

All mapping and SNP calling analyses were done using snippy (version 4.6.0; applying a minfrac value of 10 and a mincov value of 0.9; appendix 1 p 3).[19] Reference genomes chosen for each sequence type, the same as those used in the pilot and implementation studies, were complete genomes matching the sequence types of interest, either (preferentially) a locally isolated reference genome or a publicly available reference genome.

Pairwise SNP distances and transmission inference thresholds

Pairwise SNPs were calculated in R using harrwietr (version 0.2.3) and the core SNP alignments. Transmission inference thresholds (≤15 SNPs for meticillin-resistant *S aureus*, ≤25 SNPs for other species) were applied, as per the thresholds used to identify genomically closely related isolates, and therefore probably linked by transmission, in the initial pilot study[18] and the subsequent implementation study (unpublished) from which all genome sequence data are drawn. Further details on the rationale for choosing these thresholds are described in appendix 1 (p 3).

Reference genome and sample diversity analysis

Three levels of alignment were undertaken for all sequence types in each species, to investigate the impact of reference genome relatedness and isolate diversity: species level, outgroup-reference level, and sequence type level (appendix 1 p 3). Alignments at the sequence type level, with a reference genome of the same sequence type as the isolates, were used for all subsequent analyses.

Identification and masking of prophage and recombination regions

Previous studies have suggested that prophage and recombination regions result in elevated SNP counts that do not represent the vertical evolution of the population, which might interfere with identifying transmission through evolution.[20–23] All reference chromosomes were screened for prophage using phastaf (version 0.1.0) and the PHASTER database.[24] Gubbins[20] was used to predict regions of recombination in alignments at the sequence type level. The identified prophage and recombination regions were then masked in the alignments, individually and together, using the masking option in snippy.

For **phastaf** see https://github.com/tseemann/phastaf

See **Online** for appendix 1

Cumulative and sliding-window approaches to isolate inclusion

Using the sequence type alignments, without masking for prophage or recombination, two different approaches for isolate inclusion and comparison over time were implemented: a cumulative approach in which all additional isolates were included over time, and a 3-month sliding-window approach. In some cases, isolates potentially arising from the same outbreak are collected over long time periods, and could be key for context and transmission inference. This has been well described for several outbreaks of MDR organisms, such as drug-resistant *K pneumoniae*, in which epidemiologically linked samples have been found over years, in part driven by long-term asymptomatic colonisation.[25] As such, it is important to establish the potential effect of a continually growing and diversifying dataset, as compared to closed short-term datasets, on the outcomes of genomic data analyses.

For **short read sequence data** see https://www.ncbi.nlm.nih.gov/bioproject/PRJNA565795/

For **snippy** see https://github.com/tseemann/snippy

Transmission inference using defined SNP thresholds

Although an SNP threshold is commonly applied to infer likely transmission, the choice of genomic analysis methods can have a large influence on pairwise SNP distance calculations and, therefore, on which isolate pairs fall below the set threshold. Here, we applied SNP thresholds (≤15 SNPs for *S aureus*, ≤25 SNPs for other species, consistent with those in the corresponding pilot study)[18] to classify isolates as being likely or unlikely to be involved in putative transmission events for every approach used in this study. We calculated the overall proportion of isolate pairs that fell below the species' SNP thresholds for likely transmission, and we also identified how many pairs were above the SNP threshold for likely transmission in one, or more, of the alignment approaches, but were below the threshold in another. Specifically, we identified any shift from above to below the SNP threshold noted between the first and last observation of each pair of isolates compared two or more times across the different time windows for inclusion (ie, any instance in which one of these observations was above the SNP threshold and the other fell below; appendix 1 p 3).

For **harrietr** see https://github.com/andersgs/harrietr

Statistical analyses for comparison of reference genome and sample diversity analysis

All statistical analyses to compare the differences in distributions of pairwise SNP distances with each of the

different reference alignment approaches were done in R v3.6.0. Normality was tested in R using a Shapiro-Wilk's test when the data were small enough, and using Q-Q plots for larger datasets. The Kruskal-Wallis test was used for data not normally distributed to assess if differences between any of the three alignment approach groups were significant for each of the sequence types. Subsequently a pairwise Wilcoxon signed-rank test was used to compare each of the alignment approaches individually within each sequence type (ie, species level *vs* outgroup-reference level, species level *vs* sequence type level, and outgroup-reference level *vs* sequence type level), with Benjamini-Hochburg adjustment to adjust for multiple testing, when assessing statistical differences in pairwise SNP distributions between alignment approaches.

All figures were created in R (version 3.6.0; packages used are listed in appendix 1 p 4).

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

During the sampling period, 2275 samples of the four target MDR organisms were collected from 1870 patients (397 during the pilot phase from April 4 to June 18, 2017; 1878 during the implementation phase from Oct 30, 2017, to Nov 30, 2018). Following WGS of these samples, all genomes representing only the four most common sequence types from each species were selected for inclusion in this study (1537 samples from 1299 patients). This dataset included: *S aureus* ST5 (n=61), ST22 (n=222), ST45 (n=158), and ST93 (n=69); *E faecium* ST80 (n=29), ST203 (n=60), ST1421 (n=146), and ST1424 (n=70); *K pneumoniae* ST15 (n=12), ST17 (n=12), ST307 (n=23), and ST323 (n=15); and *E coli* ST38 (n=39), ST131 (n=460), ST648 (n=51), and ST1193 (n=110; appendix 1 pp 19–20; appendix 2).

See Online for appendix 5

See Online for appendix 2

Details on the resulting alignments, including pairwise SNP data, are provided in appendix 3. Phylogenetic trees, to show the relative position of the reference and population structure, are in appendix 1 (pp 6–9). Except for the diverse *E faecium* ST80 and *K pneumoniae* ST17, using a reference of the same sequence type as the isolates usually resulted in maximised core genome size (figure 1A, B) and a finer-scale SNP distance resolution (appendix 1 p 10) that was more representative of the true genetic relatedness (appendix 1 p 4).

See Online for appendix 3

Across all species, masking prophage regions had little to no effect on the core alignment, the core SNP alignment (figure 1C), or pairwise SNP distances (appendix 1 pp 11, 24; appendix 4). Prophage regions often coincided with regions that were already excluded from analysis as they did not form part of the core genome (appendix 1 pp 12–15).

See Online for appendix 4

By contrast, recombination masking showed considerable effects, although the effect size differed among the various species and sequence types (appendix 1 pp 11, 24; appendix 4). The largest differences were among multiple *E faecium* and *E coli* sequence types and *K pneumoniae* ST17, in which recombination masking saw many isolates' pairwise SNP distances fall by hundreds or even thousands of SNPs (figure 1D). The greater the number or size of regions of recombination, particularly relative to the overall genome size, the larger the effect of masking these regions of recombination had on pairwise SNP distances (appendix 1 pp 11–16; appendix 4). For example, some sequence types of *S aureus* (ST5, ST22, and ST93) and *K pneumoniae* (ST15, ST307, and ST323) each had only a few small recombination regions detected (appendix 1 pp 12, 14), and pairwise SNP distances showed minimal changes when this recombination was omitted, whereas many of the other sequence types and species had large areas of genome removed due to recombination masking (figure 1C). In the most extreme cases (*E faecium* ST80 [appendix 1 p 13] and *K pneumoniae* ST17 [appendix 1 p 14]), recombination masking resulted in significant portions of the genome being masked and the average pairwise SNP distances dropping from many thousands of SNPs to hundreds (figure 1D).

The combined masking of both prophage and recombination showed very similar results to those seen when masking for only recombination (appendix 1 p 11), as many prophage regions were encompassed in regions also identified as recombination (appendix 1 pp 12–15).

In the cumulative approach, all new isolates from each sampling month were compared to all previously included isolates. As the total number of isolates increased over time, so did diversity, resulting in a continually diminishing core genome alignment (variant and invariant sites; appendix 1 pp 17, 26; appendix 5; figure 1E). A mean of 17·6% (range 4–57%) of the reference genome length was lost from the sequence type core alignment from the first to last month of sampling. *E coli* ST131 had the greatest loss falling from 91% of the reference genome present in the core alignment in the first sampling month to only 34% in the final month, resulting in an overall loss of 57% of the reference genome. Mean sequence type core genome alignment lengths as a proportion of the reference genome, for each timepoint, are shown for each species individually and for all sequence types combined across species in figure 1E. The core SNP alignment in most sequence types increased over time; although the core genome was shrinking, more of the core sites became variant (ie, SNPs). *E coli* ST131 was an exception, with a steady decrease detected in both the core genome and core SNP alignments (appendix 1 p 17).

The sliding-window approach used a 3-month window, sliding forwards by a single month each time. In this approach, although there were fluctuations in the proportion of the reference in the core genome alignment over time, it did not continually decrease as with the

cumulative approach (figure 1E, appendix 1 pp 18, 27; appendix 6). The mean core alignment size was consistently higher; more potentially informative sites are present at each timepoint, providing finer resolution. For example, although the mean proportion of the reference genome in the core alignment for *E coli* ST131 was reduced to 48% (range 34–91%) in the cumulative approach, it was 68% (range 50–85%) in the sliding-window approach. In providing much larger and more consistently sized core alignments, the proportion of reference genome represented in the core alignment, it is also easier to compare pairwise SNP distances over time.

When assessing the effect of isolate and reference genome diversity we found that the outgroup-reference approach provided the lowest number of likely transmission pairs compared with both the species and sequence type alignments (figure 2; appendix 1 p 22–23; appendix 3). In most sequence types, using the outgroup reference did not identify any pairs falling below the SNP thresholds. Additionally, none of the pairs that experienced a shift below the SNP threshold did so because of the outgroup-reference analysis, except for the *E faecium* ST1424 (appendix 1 p 22); this is possibly because the ST1421 reference genome sits between two subclusters of ST1424 isolates in the outgroup-reference phylogenies, thereby behaving less as an outgroup reference (appendix 1 p 7). The same was calculated for comparing absence of masking of prophage or recombination regions (or both), to the unmasked alignment, including the number of pairs shifting below the threshold following masking of any kind (figure 2; appendix 1 p 24–25; appendix 4). In almost all cases, masking prophage had little effect on reclassifying pairs of isolates to below the SNP thresholds. Conversely, in most species and sequence types in which large amounts of recombination were detected and masked, the number of pairs shifting below the SNP threshold increased by hundreds or, in

See Online for appendix 6

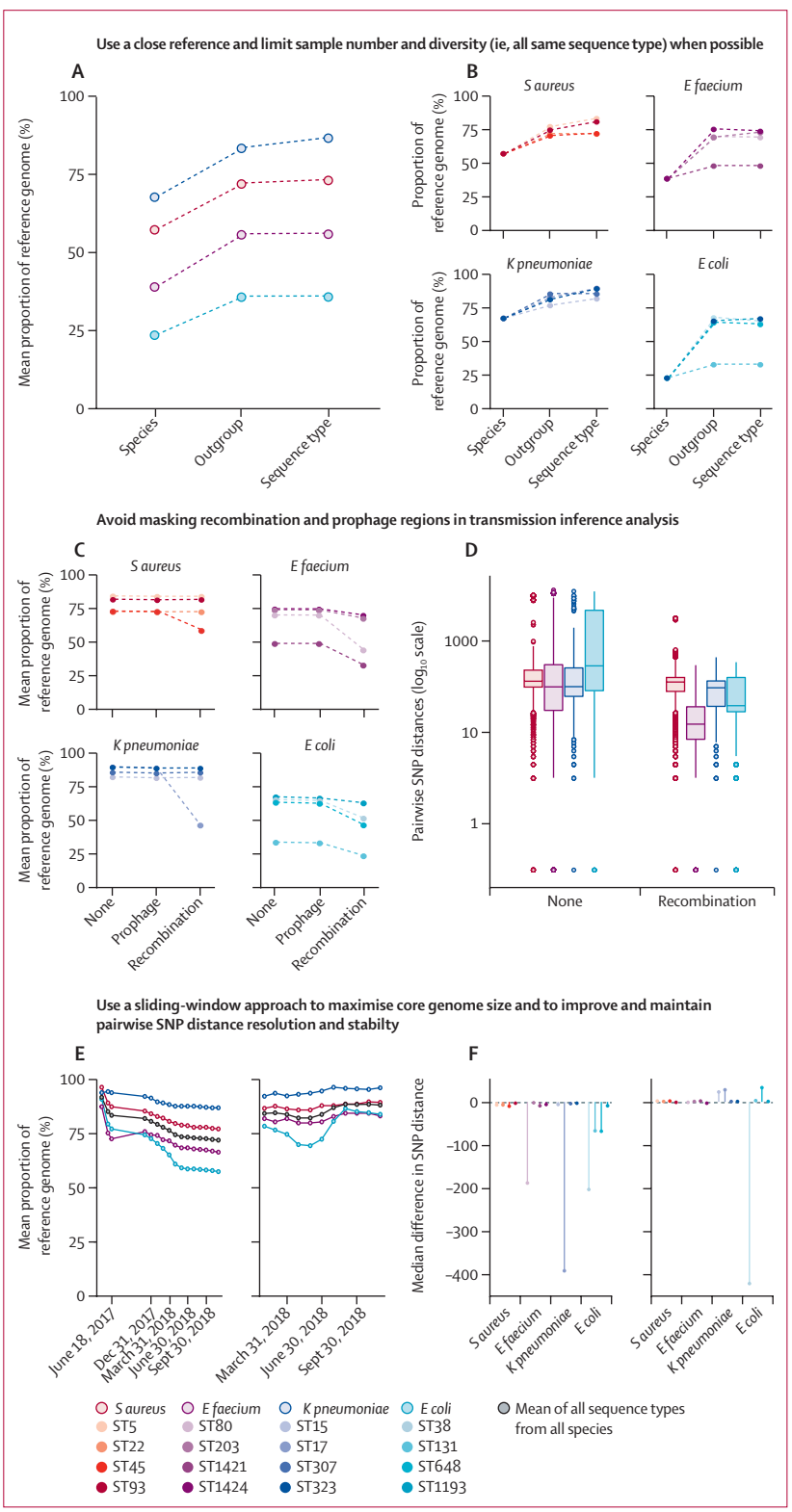

Figure 1: **Framework recommendation and justification for pathogen-specific standardisation for the surveillance of multidrug-resistant organisms using genomics**

Percentage of the reference genome represented in the core genome for each species (A) and sequence type (B), with each of the three different alignment approaches (shown on the x-axis), and for each sequence type (C), with each of the three main approaches to masking regions of horizontal gene transfer (ie, no masking, masking of prophage regions, and masking of recombination regions; shown on the x-axis). (D) Distribution of pairwise SNP distances between all isolate pairs, grouped by species, without any masking and with masking of recombination regions. (E) Proportion of reference genome included in the core alignment of all sequence types, either within each species (mean for all sequence types of a given species; coloured points) or across all species (mean for all sequence types regardless of species; grey points) as a percentage. (F) Median difference between the initial and final pairwise SNP distances, for all pairs compared at least twice, for the cumulative and sliding-window approaches; a negative value shows a median decrease in pairwise SNP distances and therefore a loss of genetic resolution over time as the dataset changes or grows, a positive value shows the opposite. Points and plots are coloured by species and sequence type, and dotted lines are used (in panels A, B, C) for ease of visualising the relationship between discrete approach variables. *E coli*=*Escherichia coli*. *E faecium*=*Enterococcus faecium*. *K pneumoniae*=*Klebsiella pneumoniae*. *S aureus*=*Staphylococcus aureus*. SNP=single nucleotide polymorphism.

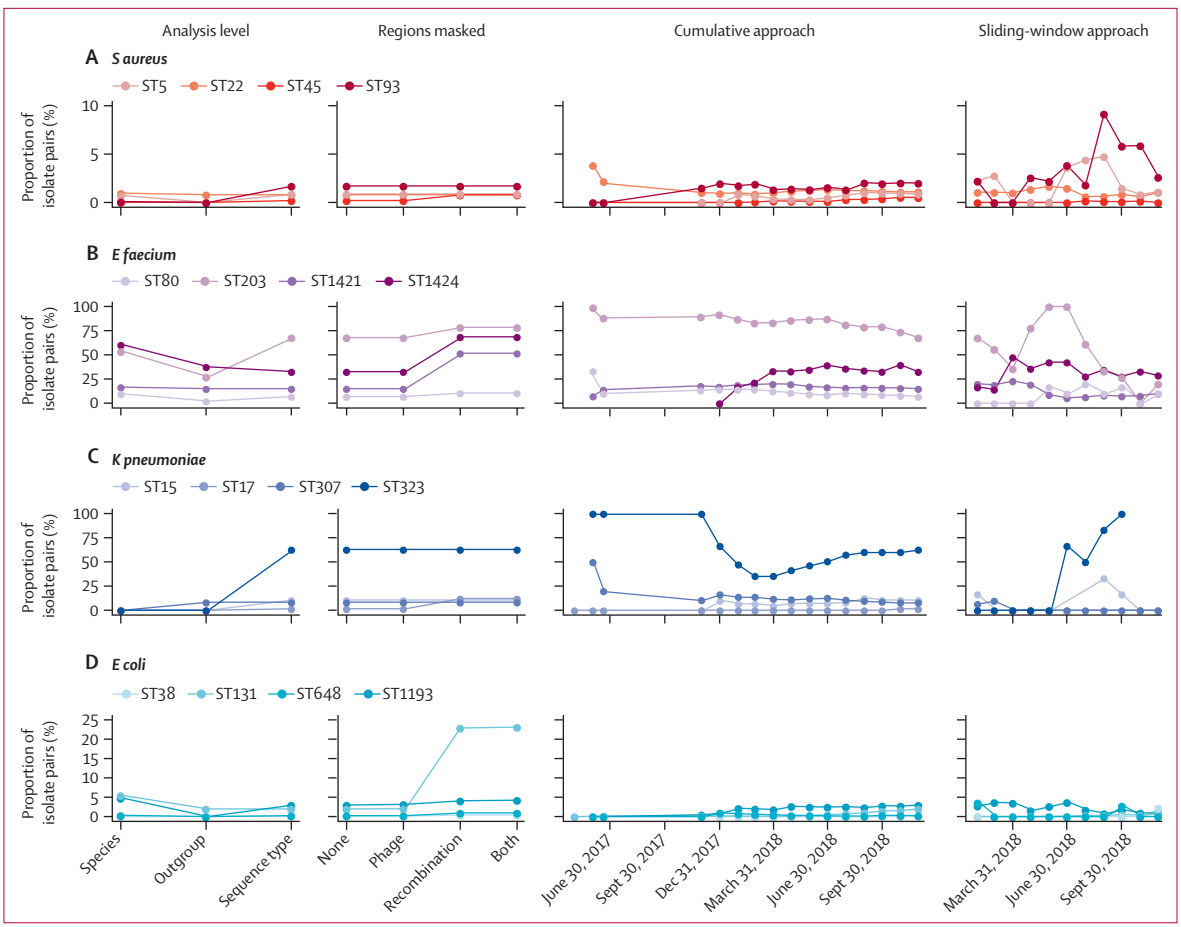

Figure 2: **Effects of analysis level, masking of prophage or recombination regions (or both), and different approaches to sample inclusion over time, on the proportion of isolate pairs falling under the SNP threshold for putative transmission**

The y-axis shows the percentage of pairs of isolates between which the genetic distance is equal to or smaller than the SNP threshold for putative transmission for each species; for *S aureus* (A) the threshold is 15 SNPs, for all other species (B–D) the threshold is 25 SNPs. The y-axis maximum value differs by species but is consistent across all plots for the species. All plots are coloured by sequence type. *E coli*=*Escherichia coli*. *E faecium*=*Enterococcus faecium*. *K pneumoniae*=*Klebsiella pneumoniae*. *S aureus*=*Staphylococcus aureus*. SNP=single nucleotide polymorphism.

the case of *E faecium* ST1421 and *E coli* ST131, by thousands. Finally, we considered the effect of the cumulative and sliding-window approaches to sample inclusion (figure 2; appendix 1 p 26–27; appendices 5, 6), identifying many cases of pairs of isolates shifting below the SNP threshold over time (appendix 1 p 28).

Discussion

The findings of our study show that the best generalisable approach for optimising genomic resolution and pairwise SNP distance stability was to use a closely related reference genome, without omitting prophage or recombination regions, and using a sliding window for sample inclusion. These methods provide finer scale resolution and greater consistency and accuracy in pairwise SNP distances for inferring isolate relatedness, making the application of a single SNP threshold to define transmission more appropriate than other approaches. These findings provide the basis for a framework for pathogen-specific standardisation for the surveillance of MDR organisms using genomics.

Prospective WGS of hospital-associated MDR organisms will enhance identification of real-time transmission, leading to optimised infection prevention and control and thus limiting further spread; however, methods need to be standardised. Previous studies are often retrospective and ad hoc, frequently tailored to a specific, narrow dataset, such as closely related isolates from a single pathogen sequence type or a rare AMR-resistant phenotype.[17,25–27] This is not the reality of prospective hospital or jurisdictional wide surveillance, in which multiple pathogens and sequence types are detected over time.[16] As such, the results, methods, and thresholds that have been used are not necessarily broadly applicable for prospective surveillance where the dataset continues to expand over time. Here, we utilised a multi-institutional dataset of MDR organisms to systematically investigate a range of approaches on the

outcome of potential transmission analyses, providing recommendations for future implementation (figure 1).

Using a more distant reference genome inflates pairwise SNP distances, increases ancestral SNPs, and decreases the number of SNPs that have arisen more recently, hence losing the fine-scale resolution required for transmission inference. Given the increased core genome size and fine-scale resolution among more closely related isolate pairs offered when using a closely related reference genome, we recommend using this approach whenever possible, although the increased accuracy will be reduced when isolates from within a sequence type are highly diverse (figure 1A, B).

Prophage masking had little effect in this dataset, primarily because the prophage sites corresponded to regions that were already absent from the core genome alignment. In datasets where this is not the case, the effect might change but should be assessed. Masking recombination had varying effects, heavily dependent on the individual sequence type datasets. In cases where isolates were closely related before masking, there was little effect, with the opposite seen in more diverse sequence types. Isolate pairs that have large pairwise SNP distances, but which have many of these SNPs masked as regions of recombination, can then erroneously appear to be closely related and could incorrectly be inferred as likely transmission. In these cases, masking recombination when inferring transmission would be inappropriate and misleading as they are not truly genetically close. The number and size of recombination regions, as well as the extent of the effect of masking, should be carefully considered; a pair of isolates that have a small number of SNPs after masking but had a hundred regions masked spanning thousands of SNPs, should not be considered as closely related as a pair that had a small number of pairwise SNPs both before and after masking. When isolates are already closely related (ie, separated by small SNP distances), masking prophage or recombination (or both) makes minimal, if any, difference in pairwise SNP distances—meaning that transmission inference is unaffected. Although masking prophage and recombination might be applicable when studying long-term evolution in diverse populations, it should not be routinely applied for the species discussed here when assessing potential transmission. In summary, masking of prophage has minimal effects but increases time and effort required, and masking of recombination regions has the potential to inappropriately reduce the number of SNPs between truly distant isolates (figure 1C, D).

Finally, determining putative transmission often revolves around ruling isolates in or out of a particular genomic cluster, on the basis of set genomic thresholds and supported by epidemiological analyses. In a truly real-time dataset, new isolates will be continually added over time. The four species in this study can reside as asymptomatic commensal organisms, remaining undetected unless carriage screening is undertaken, and during this time can undergo diversifying evolution within the host. Given the shrinking core genome and core SNPs, it is also possible that isolate pairs that are distantly related at an initial timepoint could lose much of that measurable genetic distance by the final timepoint (figure 1F). This presents at least two serious issues in determining genetic relatedness. If using a threshold to rule transmission in or out, this isolate pair would be initially ruled out and subsequently ruled in as putative transmission. Scaling the numbers of SNPs such that they are proportionate to the amount of core genome or to the entire reference genome might lessen these effects, but if the parts that are lost from the core over time are the more diverse, these scaled or adjusted numbers will still fall short of the true diversity. Many of these problems are true of the cumulative approach to sample inclusion but are reduced when using a sliding-window approach and as such we recommend this approach. Core genome and SNP alignments, and relative pairwise SNP distances, remain more stable over time, making it easier to standardise or draw comparisons between pairs of isolates over time. This approach is also less computationally intensive, given the smaller number of isolates at each timepoint.

Ultimately, in the context of determining putative transmission it is probable that an SNP threshold will be implemented to rule isolates in or out of transmission events. However, although the threshold might be set, we have shown that changes in analysis or isolates included can see pairs of isolates shifting from above the SNP threshold (and, therefore, ruled out of transmission events) to below the threshold (and thus ruled in). The most dramatic influences on this shifting from above to below the SNP threshold were observed when masking regions of recombination, followed by reference genome choice, and when isolate diversity in the alignment is high. Of note, despite the shrinking core alignments observed over time in the cumulative isolate inclusion approach, compared with the sliding-window approach, we saw relatively small numbers of isolates switching from above to below the SNP threshold. However, a larger influence was seen among the more genetically diverse sequence types (*E faecium* ST1421 and *E coli* ST131).

Although this novel study uses a comprehensive and diverse dataset to systematically address many key factors that can considerably influence transmission inference, several limitations and factors still need to be considered. Closer reference genomes offer finer resolution SNP distances, more representative of the true genetic relatedness; however, for some diverse or polyphyletic sequence types, it appears to make little difference whether an outgroup-reference genome or one of the same sequence types is used. Although the outcome for these diverse sequence types is not worse with a closer reference, there is minimal benefit observed to using one. Another potentially problematic scenario is when no close reference genome is available, such as when

assessing rare or novel sequence types. Next, although we do not recommend routinely masking prophage and recombination regions there could be exceptions. For example, if prophage regions are conserved across all isolates then SNPs in these conserved regions might result from vertical evolution and be informative for SNP distances. Alternatively, when recombination is limited to only few regions, rather than many regions throughout, this could indicate limited, recent recombination events and consequently few evolutionary steps; in these cases, masking these few regions could be appropriate. When considering a cumulative or sliding-window approach to isolate inclusion, it should be noted that links between genetically close but temporally distant isolates could be missed with the sliding-window approach. However, using approaches such as single-linkage methods (to identify relatedness between windows) could be used to highlight persistent transmission chains.[28] For example, if in time period one hypothetical isolates A and B are closely related, and in time period two isolates B and C are closely related, although isolate A is not within time period two, we can infer that although separated by time, A is related to C through B. Last, although our results show that different methods can have a potentially large impact on which pairs of isolates fall above or below a set SNP threshold for transmission inference, given the variability observed for some of the species and sequence types, a single SNP threshold might not be applicable across multiple distinct strains. Further detailed investigation is warranted, to identify the extent, impact, and potential solutions to this issue.

Although there are limitations to this study, as described, it remains the first systematic analysis to assess these factors across such a large and diverse dataset. In summation, when implementing WGS for transmission surveillance of common MDR organisms we recommend using a closely related genome, without masking of prophage or recombination regions, and a sliding-window approach (figure 1). These approaches all contribute to maximising SNP distance resolution and stability in an evolving, real-time dataset. These findings help fill the knowledge gap that has hindered the effective implementation of real-time genomic surveillance of MDR organisms in health-care facilities. The standardised and generalisable approaches presented here for implementation are an important step in facilitating widespread genomic analysis and prospective surveillance of MDR organisms in clinical settings worldwide.

Contributors

BPH and MLG designed and managed the Controlling Superbugs Study. BPH, CLG, and NLS designed this project and verified the underlying data of the study. CLG conducted all genomic, bioinformatic and statistical analyses, and produced the manuscript and all accompanying figures and tables. AGDS was part of the Controlling Superbugs Study Group for the initial project, provided guidance and insights, and proofread and edited the manuscript in full. DJI provided ongoing input and discussion and edited the manuscript at various stages. CH helped with quality control for both sequence and epidemiological data, did the long read sequencing and assembly of the *E faecium* ST1424 reference genome, and proofread and edited the manuscript. TS wrote a python code to calculate which sites in the reference genome were categorised as core sites and offered bioinformatic advice. DAW and TPS provided guidance and feedback both during the study and for the final manuscript. JCK was part of the Controlling Superbugs Study Group for the initial project. NLS was part of the Controlling Superbugs Study Group for the initial project, helped with data collection and quality control, provided guidance and input throughout, and edited the manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

We declare no competing interests.

Data sharing

Raw sequence data has been uploaded to the Sequence Read Archive under BioProject PRJNA565795.

Acknowledgments

The authors would like to acknowledge the other members of the Controlling Superbugs Study Group that includes Robyn Lee (previously MDU, Toronto, ON, Canada; currently University of Toronto, Toronto, ON, Canada), Rhonda Stuart (Infectious Diseases, Monash Health; Medicine, Monash University, Clayton, VIC, Australia), Tony Korman (Infectious Diseases and Microbiology, Monash Health; Medicine, Monash University, Clayton, VIC, Australia), Caroline Marshall (VIDS, Melbourne Health, Melbourne, VIC, Australia; Peter Doherty Institute, Melbourne, VIC, Australia), Hiu Tat (Mark) Chan (Microbiology, Melbourne Health, Melbourne, VIC, Australia), Maryza Graham (Infectious Diseases and Microbiology, Monash Health; Medicine, Monash University, Clayton, VIC, Australia), Marcel Leroi (Microbiology, Austin Health, Heidelberg, VIC, Australia), Caroline Reed (Microbiology, Melbourne Health, Melbourne, VIC, Australia; Peter MacCallum Cancer Centre, Melbourne, VIC, Australia), Michael Richards (VIDS, Melbourne Health, Melbourne, VIC, Australia; Peter Doherty Institute, Melbourne, VIC, Australia), Monica Slavin and Leon Worth (Infectious Diseases, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia; National Centre for Infections in Cancer Melbourne, VIC, Australia; University of Melbourne, Melbourne, VIC, Australia), Elizabeth Grabsch (Microbiology, Austin Health, Heidelberg, VIC, Australia), Joanna Price and Carolyn Tullett (Infection Control, Austin Health), Despina Kotsanas (Microbiology, Monash Health, Clayton, VIC, Australia), Louise Wright (Infection Control, Monash Health, Clayton, VIC, Australia), Suraya Hanim Abdullah Hashim and Jennifer Mitchell (Infectious Diseases, Melbourne Health, Melbourne, VIC, Australia), Olivia Smibert (Infectious Diseases, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia), and Carol Wedge (Data Entry, Austin Health, Heidelberg, VIC, Australia). This work was supported by the Melbourne Genomics Health Alliance (funded by the State Government of Victoria [Australia], Department of Health and Human Services, and the ten member organisations); a National Health and Medical Research Council (Australia) Partnership grant (GNT1149991) and individual grants from National Health and Medical Research Council (Australia) to NLS (GNT1093468), JCK (GNT1142613) and BPH (GNT1105905).

References

1 WHO. Antimicrobial resistance: global report on surveillance. Geneva: World Health Organisation, 2014.
2 WHO. Global antimicrobial resistance surveillance system (GLASS) report: early implementation 2017–2018. Geneva: World Health Organization, 2018.
3 WHO. Global antimicrobial resistance and use surveillance system (GLASS): whole-genome sequencing for surveillance of antimicrobial resistance. Geneva: World Health Organization, 2020.
4 Centres for Disease Control and Prevention. Antibiotic resistance threats in the United States, 2019. Atlanta, GA: US Department of Health and Human Services, Centres for Disease Control and Prevention, 2019.
5 Cosgrove SE. The relationship between antimicrobial resistance and patient outcomes: mortality, length of hospital stay, and health care costs. *Clin Infect Dis* 2006; **42** (suppl 2): S82–89.

6 Schulgen G, Kropec A, Kappstein I, Daschner F, Schumacher M. Estimation of extra hospital stay attributable to nosocomial infections: heterogeneity and timing of events. *J Clin Epidemiol* 2000; **53:** 409–17.

7 Arefian H, Hagel S, Heublein S, et al. Extra length of stay and costs because of health care-associated infections at a German university hospital. *Am J Infect Control* 2016; **44:** 160–66.

8 Dramowski A, Whitelaw A, Cotton MF. Burden, spectrum, and impact of healthcare-associated infection at a South African children's hospital. *J Hosp Infect* 2016; **94:** 364–72.

9 Maragakis LL, Perencevich EN, Cosgrove SE. Clinical and economic burden of antimicrobial resistance. *Expert Rev Anti Infect Ther* 2014; **6:** 751–63.

10 Boucher HW, Talbot GH, Bradley JS, et al. Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. *Clin Infect Dis* 2009; **48:** 1–12.

11 Pendleton JN, Gorman SP, Gilmore BF. Clinical relevance of the ESKAPE pathogens. *Expert Rev Anti Infect Ther* 2013; **11:** 297–308.

12 WHO. Global antimicrobial resistance surveillance system (GLASS) report. Geneva: World Health Organisation, 2019.

13 Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF. Using genomics to track global antimicrobial resistance. *Front Public Health* 2019; **7:** 242.

14 Sherry NL, Lane CR, Kwong JC, et al. Genomics for molecular epidemiology and detecting transmission of carbapenemase-producing Enterobacterales in Victoria, Australia, 2012 to 2016. *J Clin Microbiol* 2019; **57:** e00573–19.

15 Gorrie CL, Mirčeta M, Wick RR, et al. Gastrointestinal carriage is a major reservoir of *Klebsiella pneumoniae* infection in intensive care patients. *Clin Infect Dis* 2017; **65:** 208–15.

16 Raven KE, Gouliouris T, Brodrick H, et al. Complex routes of nosocomial vancomycin-resistant *Enterococcus faecium* transmission revealed by genome sequencing. *Clin Infect Dis* 2017; **64:** 886–93.

17 Harris SR, Cartwright EJ, Török ME, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect Dis* 2013; **13:** 130–36.

18 Sherry NL, Lee RS, Gorrie CL, et al. Pilot study of a combined genomic and epidemiologic surveillance program for hospital-acquired multidrug-resistant pathogens across multiple hospital networks in Australia. *Infect Control Hosp Epidemiol* 2020; **42:** 573–81.

19 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015; **32:** 268–74.

20 Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015; **43:** e15.

21 Schierup MH, Hein J. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 2000; **156:** 879–91.

22 Posada D, Crandall KA. The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol* 2002; **54:** 396–402.

23 Didelot X, Maiden MCJ. Impact of recombination on bacterial evolution. *Trends Microbiol* 2010; **18:** 315–22.

24 Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016; **44:** W16–21.

25 Kwong JC, Lane CR, Romanes F, et al. Translating genomics into practice for real-time surveillance and response to carbapenemase-producing Enterobacteriaceae: evidence from a complex multi-institutional KPC outbreak. *PeerJ* 2018; **6:** e4210.

26 Snitkin ES, Zelazny AM, Thomas PJ, et al. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci Transl Med* 2012; **4:** 148ra116–6.

27 Witney AA, Gould KA, Pope CF, et al. Genome sequencing and characterization of an extensively drug-resistant sequence type 111 serotype O12 hospital outbreak strain of *Pseudomonas aeruginosa*. *Clin Microbiol Infect* 2014; **20:** O609–18.

28 Williamson DA, Chow EPF, Gorrie CL, et al. Bridging of *Neisseria gonorrhoeae* lineages across sexual networks in the HIV pre-exposure prophylaxis era. *Nat Commun* 2019; **10:** 3988.