Pan-pathogen deep sequencing of nosocomial bacterial pathogens in Italy in spring 2020: a prospective cohort study

Harry A Thorpe, Maiju Pesonen, Marta Corbella*, Henri Pesonen, Stefano Gaiarsa, Christine J Boinett, Gerry Tonkin-Hill, Tommi Mäklin, Anna K Pöntinen, Neil MacAlasdair, Rebecca A Gladstone, Sergio Arredondo-Alonso, Teemu Kallonen, Dorota Jamrozy, Stephanie W Lo, Chrispin Chaguza, Grace A Blackwell, Antti Honkela, Anita C Schürch, Rob J L Willems, Cristina Merla, Greta Petazzoni, Edward J Feil, Patrizia Cambieri, Nicholas R Thomson†, Stephen D Bentley†, Davide Sassera†, Jukka Corander†*

Summary

Background Nosocomial infections pose a considerable risk to patients who are susceptible, and this is particularly acute in intensive care units when hospital-associated bacteria are endemic. During the first wave of the COVID-19 pandemic, the surge of patients presented a significant obstacle to the effectiveness of infection control measures. We aimed to assess the risks and extent of nosocomial pathogen transmission under a high patient burden by designing a novel bacterial pan-pathogen deep-sequencing approach that could be integrated with standard clinical surveillance and diagnostics workflows.

Methods We did a prospective cohort study in a region of northern Italy that was severely affected by the first wave of the COVID-19 pandemic. Inpatients on both ordinary and intensive care unit (ICU) wards at the San Matteo hospital, Pavia were sampled on multiple occasions to identify bacterial pathogens from respiratory, nasal, and rectal samples. Diagnostic samples collected between April 7 and May 10, 2020 were cultured on six different selective media designed to enrich for *Acinetobacter baumannii*, *Escherichia coli*, *Enterococcus faecium*, *Enterococcus faecalis*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Streptococcus pneumoniae*, and DNA from each plate with positive growth was deep sequenced en masse. We used mSWEEP and mGEMS to bin sequencing reads by sequence cluster for each species, followed by mapping with snippy to generate high quality alignments. Antimicrobial resistance genes were detected by use of ARIBA and CARD. Estimates of hospital transmission were obtained from pairwise bacterial single nucleotide polymorphism distances, partitioned by within-patient and between-patient samples. Finally, we compared the accuracy of our binned *Acinetobacter baumannii* genomes with those obtained by single colony whole-genome sequencing of isolates from the same hospital.

Findings We recruited patients from March 1 to May 7, 2020. The pathogen population among the patients was large and diverse, with 2148 species detections overall among the 2418 sequenced samples from the 256 patients. In total, 55 sequence clusters from key pathogen species were detected at least five times. The antimicrobial resistance gene prevalence was correspondingly high, with key carbapenemase and extended spectrum ß-lactamase genes detected in at least 50 (40%) of 125 patients in ICUs. Using high-resolution mapping to infer transmission, we established that hospital transmission was likely to be a significant mode of acquisition for each of the pathogen species. Finally, comparison with single colony *Acinetobacter baumannii* genomes showed that the resolution offered by deep sequencing was equivalent to single-colony sequencing, with the additional benefit of detection of co-colonisation of highly similar strains.

Interpretation Our study shows that a culture-based deep-sequencing approach is a possible route towards improving future pathogen surveillance and infection control at hospitals. Future studies should be designed to directly compare the accuracy, cost, and feasibility of culture-based deep sequencing with single colony whole-genome sequencing on a range of bacterial species.

Funding Wellcome Trust, European Research Council, Academy of Finland Flagship program, Trond Mohn Foundation, and Research Council of Norway.

Introduction

Nosocomial infections caused by multidrug-resistant bacterial pathogens are a major cause of mortality and morbidity among patients in hospital and add a significant economic and management burden to health-care systems. According to estimates of infection prevalence in Europe from 2016 to 2017, 5·9% of all patients and 19·2% of patients on intensive care units (ICUs) had a nosocomial infection.[1] Another European study, which used data from 2015, estimated that 63·5% of all infections with antimicrobial resistant (AMR) bacteria were associated with health care.[2] This threat is potentially elevated under circumstances where the inflow of critically ill patients is rapidly increased, such as during the COVID-19 pandemic. Pathogen

Lancet Microbe 2024

Published Online https://doi.org/10.1016/S2666-5247(24)00113-7

*Contributed equally
†Contributed equally

Department of Biostatistics, Faculty of Medicine, University of Oslo, Oslo, Norway (H A Thorpe PhD, M Pesonen PhD, H Pesonen DSc, G Tonkin-Hill PhD, A K Pöntinen PhD, N MacAlasdair PhD, R A Gladstone PhD, S Arredondo-Alonso PhD, Prof J Corander PhD); **Microbiology and Virology Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy** (M Corbella MSc, S Gaiarsa PhD, C Merla PhD, G Petazzoni MSc, P Cambieri MD); **Parasites and Microbes, Wellcome Sanger Institute, Cambridge, UK** (C J Boinett PhD, G Tonkin-Hill, N MacAlasdair, D Jamrozy PhD, S W Lo PhD, C Chaguza PhD, G A Blackwell PhD, Prof N R Thomson PhD, Prof S D Bentley PhD, Prof J Corander); **Department of Computer Science, University of Helsinki, Helsinki, Finland** (T Mäklin PhD, Prof A Honkela PhD); **Institute of Biomedicine, University of Turku, Turku, Finland** (T Kallonen PhD); **Department of Medical Microbiology, Universitair Medisch Centrum Utrecht, Utrecht, Netherlands** (A C Schürch PhD, Prof R J L Willems); **Department of Medical, Surgical, Diagnostic and Pediatric Sciences, University of Pavia, Pavia, Italy** (G Petazzoni); **Milner Centre for Evolution, University of Bath, Claverton Down, Bath, UK** (Prof E J Feil PhD); **Department of Biology and Biotechnology, University of Pavia, Pavia, Italy** (D Sassera PhD); **Fondazione IRCCS Policlinico San Matteo, Pavia, Italy** (D Sassera);

Helsinki Institute for Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
(Prof J Corander)

Correspondence to:
Dr Harry A Thorpe, Department of Biostatistics, Faculty of Medicine, University of Oslo, 0313 Oslo, Norway
harry.thorpe@medisin.uio.no

or

Prof Jukka Corander, Department of Biostatistics, Faculty of Medicine, University of Oslo, 0313 Oslo, Norway
jukka.corander@medisin.uio.no

Research in context

Evidence before this study

Soon after the emergence of SARS-CoV-2, multiple clinical studies established that mortality from severe infection was highest among older, immunocompromised individuals and patients with severe comorbidities. Several follow-up studies investigating the association between bacterial co-infections and COVID-19 mortality concluded that the effect of co-infections tends to be minimal. However, it is expected that in regions with a high burden of endemic nosocomial pathogens, there is an elevated risk of hospital transmission of virulent and multidrug-resistant bacterial strains during invasive procedures, such as ventilator treatment. We first searched PubMed from database inception to Aug 7, 2023, with the search terms "SARS-CoV-2" AND "bacterial" AND "co-infections", which resulted in 636 publications, including reviews and meta-analyses. We then searched for "SARS-CoV-2" AND "co-infection" AND "whole-genome sequencing", which returned only seven publications, five of which focused on SARS-CoV-2. This indicates a paucity of studies relying on whole-genome sequencing of bacterial pathogens that were circulating in health-care settings during the pandemic-driven disruption of infection control. Both searches were restricted to articles published in English.

Added value of this study

Our study presents the first attempt to recover bacterial strains at single nucleotide polymorphism-level resolution, while simultaneously capturing most clinically relevant major bacterial pathogens present in the gut, upper airways, or lungs of the patients. The deep-sequencing approach used in the study enabled quantification of antimicrobial resistance prevalence and estimation of hospital transmission. By sampling patients within a region severely affected by both the COVID-19 pandemic and endemic circulation of nosocomial bacterial pathogens, we were able to assess transmission in a high-risk setting.

Implications of all the available evidence

Our study serves as a proof-of-concept of how to use the power of deep sequencing to investigate multiple relevant pathogenic organisms simultaneously instead of a siloed approach based on whole-genome sequencing of single isolates of individual species. The approach we describe could facilitate the development of improved future guidelines for assessing and managing the risk of nosocomial infections for different patient populations, particularly on intensive care units.

surveillance is key to managing the burden of nosocomial infections, and in the last 15 years genomics has become a powerful tool for tracking pathogen evolution and transmission.[3,4]

Whole-genome sequencing (WGS) of DNA from a pure colony identified after culturing is considered the gold standard for pathogen surveillance.[4–6] However, this approach has a major limitation: it only provides a partial snapshot of the bacterial diversity that can be present in a clinical sample. This limitation can, in principle, be remedied by picking and sequencing multiple colonies per sample, but this is, in practice, both expensive and time-consuming. Shotgun metagenomics has the advantage of being able to capture all microbial diversity present in a sample and has been applied clinically, for example to establish the aetiology of conditions with unknown cause, but with variable success.[7–9] However, this approach is not generally suitable for genomic epidemiology and transmission analysis which require strain-level resolution, owing to the high sample biomass and sequencing depth required.

These issues have motivated the development of culture-based metagenomic profiling, whereby bacteria of interest are first enriched by culturing on selective media, followed by deep sequencing of the total DNA present.[10] This approach maintains the genetic diversity present within the taxa of interest, while eliminating the problems of low biomass samples, contamination with host DNA, or low abundance of taxa of interest. Because the selected taxa are typically closely related, potentially including several closely related strains, sequence types (STs), or sequence clusters (SCs) from the same species, standard metagenomics methods do not provide sufficient resolution to separate strains, especially from short-read sequencing data. To overcome this limitation, we developed a reference-based deconvolution algorithm (implemented in mSWEEP[11] and mGEMS[12]) that can accurately reconstruct individual bacterial genomes from metagenomic sequencing data.[13,14] This culture-based, deep-sequencing approach has been used to study the major community-acquired pathogen *Streptococcus pneumoniae*, showing that the richer data facilitated finer-scale resolution in both transmission and evolutionary analyses than by using a single colony-based approach.[14]

In this study, we used deep sequencing after culturing, combined with ST-level deconvolution with mSWEEP and mGEMS, to characterise the bacterial pathogen population within a single hospital in Pavia, northern Italy, during the first wave of the COVID-19 pandemic. Northern Italy has a high burden of AMR infections and was the first region in Europe to be severely affected by COVID-19; such conditions presented serious challenges to infection control, risking increased nosocomial transmission of bacterial pathogens. We focused on *Acinetobacter baumannii*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, *Enterococcus faecium*, *Enterococcus faecalis*, *Escherichia coli*, and *Staphylococcus aureus* as these species contribute the most to the overall burden of AMR and nosocomial infections according to recent estimates.[2] The aims of the study were to identify the bacterial pathogens and AMR-associated

genes present in the hospital, assess their distribution according to ward and sample type, characterise the pathogen genetic diversity, and infer transmission, and compare the deep-sequencing data with existing single-colony whole-genome sequence data.

Methods

Study design and participants

This prospective cohort study was done in an 800-bed teaching hospital serving the Lombardy region in northern Italy (Fondazione IRCCS Policlinico San Matteo, Pavia, Italy) during the first wave of the COVID-19 pandemic in the spring of 2020. Nasal, rectal, and respiratory samples collected according to the routine screening and diagnostic protocols at the hospital were used, and no exclusion was applied on the basis of hospital ward, diagnosis or patient characteristics. For each patient, we recorded the dates of admission and discharge, hospital ward, age, and sex (appendix 1 p 1).

The study was designed and conducted in accordance with the Helsinki Declaration and approved by the Ethics Committee of Fondazione IRCCS Policlinico San Matteo in Pavia, Italy (2023-3.11/105). The work described herein is a prospective study done on bacterial isolates from human samples that were obtained as part of routine hospital care and hence the requirement for patient informed consent was waived.

Culturing

Samples were cultured on six different media to select for pathogen species of interest. Chocolate agar plus PolyViteX (PVX), and Columbia agar plus 5% sheep blood, (COS), were used to select for all bacteria and fungi. Columbia Colistin-Nalidixic Acid agar plus 5% sheep blood (CNA), was used to select for Gram-positive bacteria. McConkey agar (MCK), was used to select for Gram-negative bacteria. Mannitol Salt Agar (MSA), was used to select for *Staphylococcus* spp. CHROMID Candida Agar (CAN), was used to select for *Candida* spp. All agar was purchased from Biomerieux, Marcy l'Etoile, France. Nasal swabs were plated on PVX, CNA, and CAN; rectal swabs were plated on CNA, CAN, and MCK; and respiratory samples were plated on PVX, COS, MCK, MSA, and CAN.

All samples were cultured for 48 h (at 37°C, PVX at 5% CO_2), and plates were swept to collect all microorganisms, and material was dissolved in saline buffer, pelleted, and frozen (appendix 1 pp 1–2, 7).

DNA extraction, library preparation, and sequencing

DNA was extracted by use of the QIAGEN DNeasy 96 Power Soil Kit (QIAGEN, Hilden, Germany) according to the manufacturer's instructions. DNA sequencing library preparation was done with the QIAseq FX DNA Library Kit (QIAGEN, Hilden, Germany). WGS was done on the Illumina Novaseq 6000 platform (Illumina, San Diego, CA, USA) with 150 bp paired-end reads (appendix 1 p 2).

Reference database construction

We built two sets of reference databases for use with the mSWEEP–mGEMS[11,12] pipeline; the first contained a broad set of species, and the second consisted of detailed databases of the pathogen species of interest. To build the broad reference, we used Kraken version 1.1.1[15] to identify the species present within the sequencing read sets (appendix 1 p 3). We identified 101 species and downloaded their genomes from the National Center for Biotechnology Information and subjected them to quality control (appendix 1 pp 20–21). For each pathogen species of interest we used high-quality, detailed sets of genome assemblies from curated collections to enable SC-level identification.[16–20] Each of these reference sets was clustered into SCs by use of PopPUNK.[21] Each SC was additionally labelled by the most common ST contained within it, as these are widely recognised groups (for example *K pneumoniae* SC1_ST307). The two sets of reference genomes were combined to give a database of 10 993 genomes representing 101 species (appendix 1 pp 2–3, 21–29).

See Online for appendix 1

Species and sequence cluster detection

We used a reference-based, hierarchical approach to detect species and SCs accurately (appendix 1 p 10). Themisto version 0.2.0[22] was used to pseudoalign the sequence reads to the combined reference database, mSWEEP version 1.4.0[11] was used to estimate species abundances, and mGEMS version 1.0.0[12] was used to bin the reads into species bins. For each pathogen species of interest, the species-binned reads were used as input to a second round of the pipeline, with detailed species-specific reference databases, to estimate SC abundances and bin the reads into SC bins. To ensure that species and SC assignments were accurate, we used Mash version 2.3[23] to check that the genetic distances were within the expected range for each cluster (appendix 1 pp 3–4).

The code for our demix checking pipeline is available at https://github.com/harry-thorpe/demix_check

Antimicrobial resistance gene detection

We used ARIBA version 2.14.6[24] with the CARD version 3.2.8[25] database to identify AMR genes in the sequence read sets (appendix 1 p 4).

Mapping and single nucleotide polymorphism calling

Mapping was done with Snippy version 4.6, and Snippy-core was used to generate whole-genome and core-genome alignments. We then used snp-dists version 0.7 to calculate pairwise single nucleotide polymorphism (SNP) distances from these alignments. For the analysis of *A baumannii* SC1_ST2, we followed the same procedure, but we also incorporated the single-colony whole-genome sequence reads from Petazzoni and colleagues.[26] We did phylogenetic analysis by use of FastTree version 2.1.10 Double precision[27] (using the generalised time-reversible plus CAT model; appendix 1 pp 4–5).

Code available at https://github.com/tseemann/snippy

Code available at https://github.com/tseemann/snp-dists

Transmission analysis

We used the within-patient binned mapped genome (BMG) SNP distance distributions to identify the between-patient BMG pairs that were likely to be the result of transmission, assuming that pairs of isolates resulting from either recent transmission or repeated sampling of the same patient would be indistinguishable. For each SC, we used the 95th percentile within-patient BMG SNP distance as a threshold, and then selected the between-patient BMG pairs with SNP distance of no more than the threshold as putative transmission events (appendix 1 p 5).

See Online for appendix 3

Statistical analysis

The presence of pathogen species in ICUs versus ordinary wards was compared by use of Fisher's exact tests, whereas the presence of species versus sample types was compared by use of Chi squared tests. A two-sample *t* test was used to compare the mean number of pathogen species carried by patients between ICUs and ordinary wards. All tests were done by use of R version 4.3.3. Significant difference was considered at $p<0{\cdot}05$, and p values were adjusted for multiple comparisons with Bonferroni correction where appropriate.

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

We applied a novel culture-based deep-sequencing approach to a set of 1130 clinical samples (497 rectal swabs, 439 nasal swabs, and 194 respiratory samples), which were collected for diagnostic purposes and processed in parallel to do this study (appendix 1 p 7). The enrolled cohort consisted of 256 patients who were admitted to the San Matteo hospital between March 1 and May 7, 2020 (169 men and 87 women; median age 63 years [IQR 51–72]), of whom 143 were on ordinary wards and 125 on ICUs (12 patients were present on both at some point). 45 (18%) patients died during the course of the study. The number of admissions per day, age distribution by sex, hospital stay duration, and sampling times per patient are shown in appendix 1 (p 8). The 1130 clinical samples were cultured on six different media types to select for pathogens of interest (appendix 1 p 7), resulting in growth on 2681 of 3786 plates (median positive samples per patient 2 [IQR 1–6]).

See Online for appendix 5

See Online for appendix 6

We sequenced the total DNA from 2418 plate sweeps to high depth (mean 30 million reads; appendix 4), and used mSWEEP and mGEMS, with a broad reference database of 10 993 genomes (appendix 1 pp 10, 20–21; appendix 2), chosen after an initial scan of the reads with Kraken (appendix 1 p 9), to identify species and bin reads into species bins. These steps resulted in 3233 high-confidence assignments to 52 different species (appendix 1 pp 11–12; appendix 4), of which 2148 (66%) were assigned to seven of the eight pathogen species of interest: *A baumannii, E coli, E faecium, E faecalis, K pneumoniae, P aeruginosa,* or *S aureus* (appendix 1 pp 11–12). For these pathogen species, we used the species-binned reads as input to a second round of the binning pipeline, with detailed species-specific reference databases (appendix 1 pp 10, 22–29; appendix 3), resulting in 2238 high-confidence assignments to 132 different SCs (figure 1; appendix 4).

See Online for appendix 4

See Online for appendix 2

The number of patients carrying each SC is shown in figure 1. All *A baumannii* assignments were identified as ST2 (global clone 2),[28] in agreement with single colony whole-genome sequences collected from the same hospital during a nosocomial outbreak that overlapped the time period covered by this study.[26] In total, 15 *E coli* SCs were identified, of which the most common (SC2_ST131) corresponded to ST131, an ST which has been frequently reported in screening of nosocomial colonisation.[29] 22 *E faecium* and 39 *E faecalis* SCs were identified, including the hospital-associated *E faecium* ST117[17] and *E faecalis* ST6 and ST28.[18] The most common of the 19 *K pneumoniae* SCs corresponded to the hospital-associated clones ST307, ST392, and ST45.[19,30] Of the 23 *P aeruginosa* SCs detected, the most common corresponded to ST253, also known as PA14, which is the most common clone in *Pseudomonas* group 2.[31] For *S aureus*, the most common of the 15 detected SCs corresponded to the hospital-associated ST22[32] and the livestock-associated ST398.[33,34] Aside from these pathogen species of main clinical interest, the most common species identified from the samples were other *Staphylococcus, Streptococcus,* and *Enterococcus* spp (appendix 1 p 13).

The number of species detected in an individual patient was significantly higher in ICUs (mean 3·8 [SD 1·8]) than in ordinary wards (mean 1·6 [SD 1·2]) (*t* test, $p<0{\cdot}0001$). We then tested the association between each pathogen species and ward type, revealing that the presence of *A baumannii, K pneumoniae, P aeruginosa,* and *E faecium* was significantly associated with ICUs for both nasal and rectal samples and *E faecalis* was associated with ICUs for nasal samples, whereas *E coli* from rectal samples was associated with ordinary wards (figure 2A and appendix 5). When this analysis was done at the level of individual SCs (but without considering sample types separately), nine SCs were significantly associated with ICUs after multiple correction (appendix 6).

We screened for AMR-associated genes within the read sets using ARIBA with the CARD database, focusing on carbapenemase and extended spectrum ß-lactamase (ESBL) genes, and genes conferring vancomycin resistance (figure 2B). The most common carbapenemase genes were $bla_{\text{OXA-23}}$, $bla_{\text{KPC-2}}$, and the $bla_{\text{OXA-51}}$-like $bla_{\text{OXA-66}}$; these genes were detected in 77 (62%) of 125, 76 (61%), and 62 (50%) patients on ICUs, respectively, and in lower numbers on ordinary wards 19 (14%) of 138 patients, 31 (23%), and 12 (9%), respectively. We detected a large number of ESBL genes, of which the most common were $bla_{\text{OXA-1}}$, $bla_{\text{CTX-M-15}}$, $bla_{\text{TEM-187}}$, and $bla_{\text{CTX-M-14}}$; these were present in 81 (65%) of 125, 77 (62%), 64 (51%), and 47 (38%) patients on ICUs, respectively. We detected genes from both

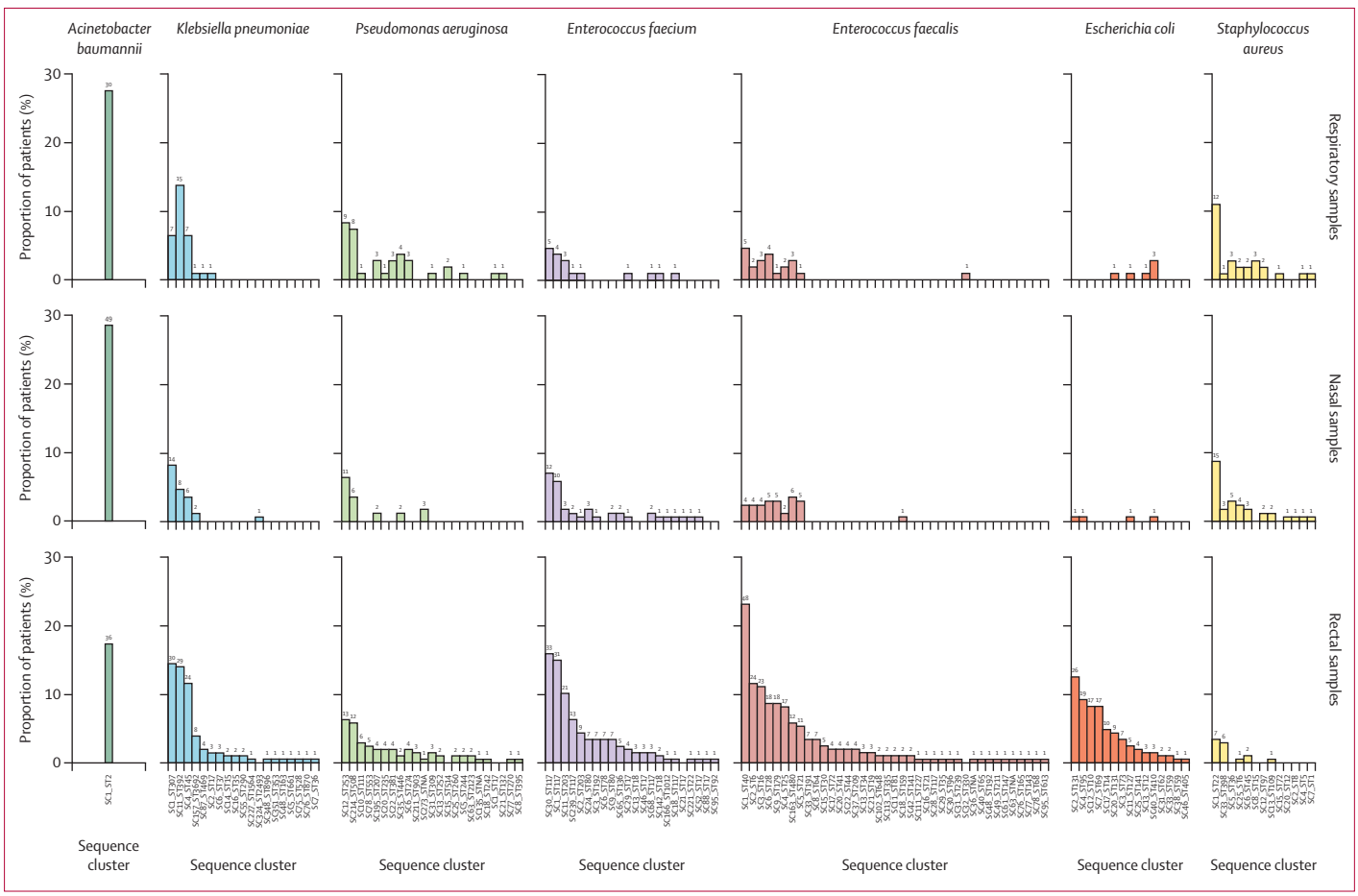

Figure 1: **The prevalence of pathogen species and SCs among the patients**
The percentage of patients who were colonised by each SC is shown, calculated separately for each sample type. The numbers above each bar indicate the corresponding numbers of patients with positive samples. Within each species, the SCs are ordered by overall prevalence. SC=sequence cluster. ST=sequence type.

the *vanA* and *vanB* vancomycin resistance operons, with the *vanA* operon being more common (50 [40%] patients on ICUs) than the *vanB* operon (nine [7%] patients on ICUs).

Deep sequencing and accurate read binning enabled us to do whole-genome mapping and SNP calling, but to differentiate our samples from traditional single colony samples we refer to ours as BMGs; mapping quality shown in appendix 1 (pp 14–15). We first investigated within-patient diversity for each SC that was detected at least five times (n=55) (figure 3; appendix 1 p 16). The core-genome SNP diversity was extremely low; 46 (84%) of 55 SCs had a median SNP distance of 0, and 49 (89%) contained no BMG pairs with a distance of ten or more SNPs (figure 3). More diversity was observed when whole-genome SNPs were considered, but this was still relatively low; 40 (73%) of 55 SCs had a median SNP distance of no more than three, and 50 (91%) had a median SNP distance of no more than eight (appendix 1 p 16). This suggests that the patients generally carried only one major lineage from each SC (along with some minor strain variation), even if they carried multiple SCs from the same species. Notable exceptions to this observation were found for *A baumannii* SC1_ST2, whereby one patient carried two sublineages of this SC, with BMGs differing by 510–13 core-genome SNPs (807–94 whole-genome SNPs), and *S aureus* SC33_ST398, whereby one patient carried two sublineages with BMGs differing by 224–26 core-genome SNPs (262–331 whole-genome SNPs).

We then investigated between-patient diversity in the same SCs (figure 3; appendix 1 p 16), which was overall much more extensive than the diversity within patients but with considerable differences between the species in their patterns of diversity. For the notorious hospital-associated species and lineages, a low amount of diversity was observed even between patients. *A baumannii* SC1_ST2 BMGs had a median pairwise distance of two core-genome SNPs (IQR 0–2) and five whole-genome SNPs (3–9), with 25% of pairs being identical according to the core-genome SNPs. There were also several much more divergent pairs, separated by 509–16 core-genome SNPs (626–954 whole-genome SNPs), and these were explained by the two separate sublineages of this SC described above.

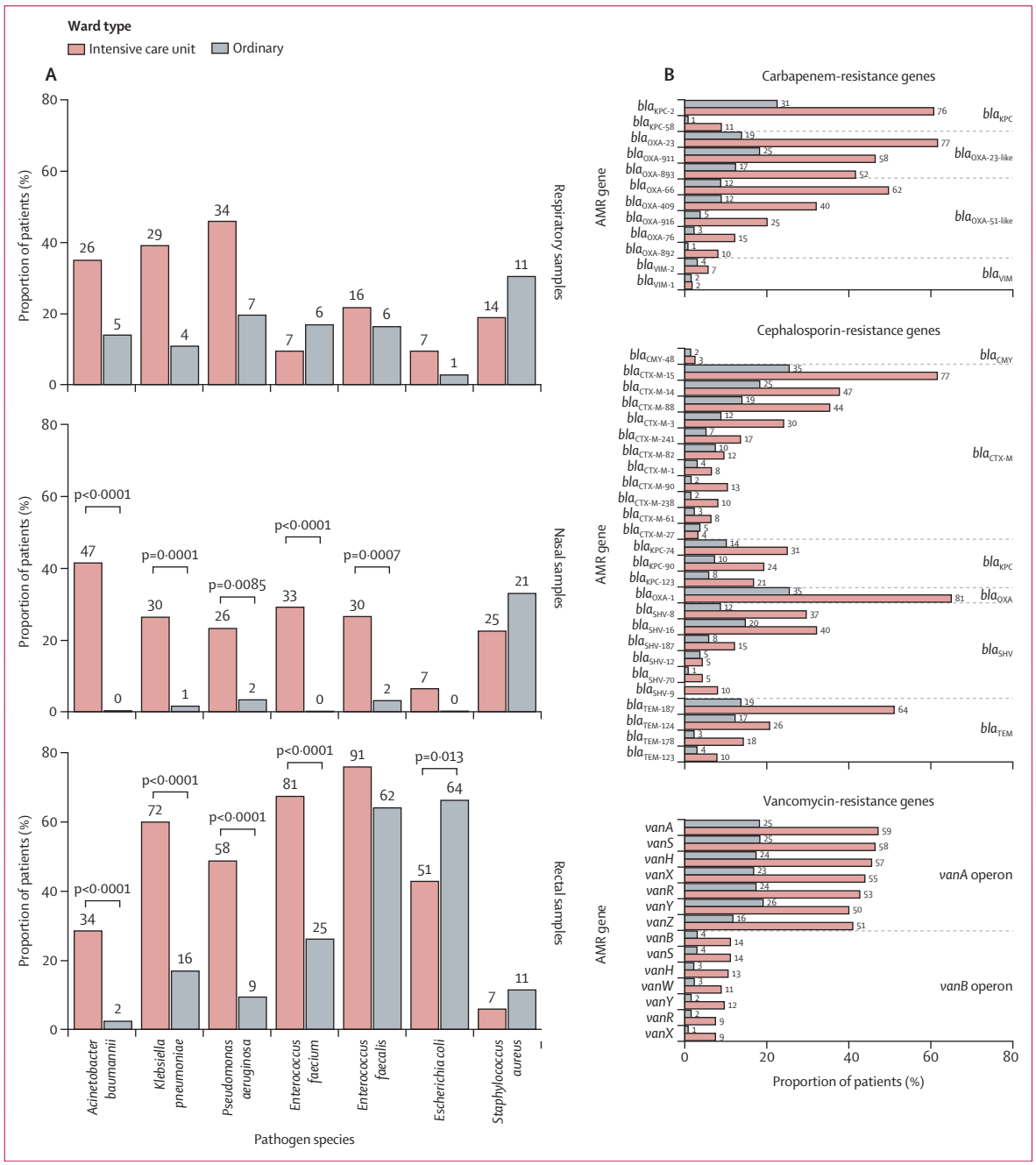

Figure 2: **Pathogen and AMR gene distribution among patients and within the hospital**

(A) Pathogen prevalence and distribution. The percentage of patients who were colonised by each pathogen species is shown, calculated separately for different ward and sample types. The numbers above each bar indicate the corresponding numbers of patients with positive samples, and the p values show the results of a Fisher's exact test of association between species prevalence and ward type. (B) AMR gene prevalence and distribution. AMR genes were detected in the raw deep sequencing reads, and the percentage of patients carrying each gene on ICU and ordinary wards is shown. The numbers to the right of each bar indicate the corresponding numbers of patients with positive samples. The three most important resistance classes among the pathogen species of interest are shown (genes conferring resistance to carbapenems, cephalosporins, and vancomycin). For carbapenem and cephalosporin resistance genes, the different *bla* types are separated by dashed lines, and ordered by prevalence within each group. For vancomycin resistance genes, the *vanA* and *vanB* operons are separated by a dashed line. AMR=antimicrobial resistance. ICU=intensive care unit.

K pneumoniae also showed minimal diversity, with four of five SCs having a median distance of 0 core-genome SNPs, corresponding to no more than ten whole-genome SNPs. For *E faecium*, seven of 12 SCs had a median pairwise distance of 0 core-genome SNPs, and three of 12 had distances between one and five SNPs. When whole-genome SNPs were considered, the distances were considerably greater, suggesting that the number of non-core SNPs was

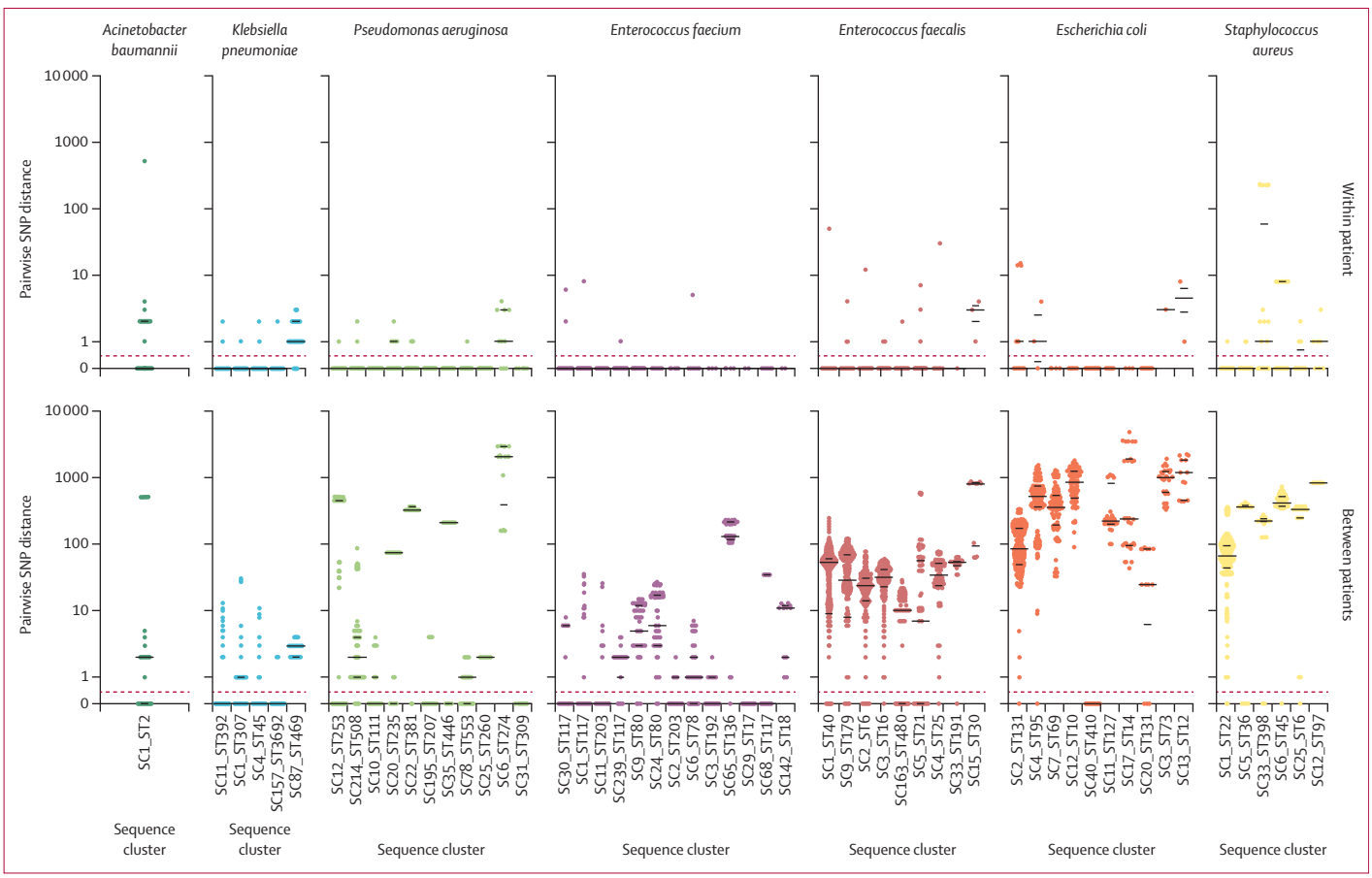

Figure 3: **Pairwise core-genome SNP distances among the SCs**

Core-genome pairwise SNP distances were calculated from mapping-based alignments for each SC. Each point shows a comparison between a pair of binned mapped genomes, and the comparisons are separated into those from the same patient (top), and different patients (bottom). The distributions are shown by violin scatter plots, with additional information given by the horizontal black bars: long bars depict medians, short bars depict 25th and 75th percentiles. SNP distances (y-axis) are displayed on a $\log_{10}$ scale; 0 has been included to ensure that identical genomes are present, with the red dashed lines indicating the break in the scale. Only SCs that were detected in at least five patients are shown. SC=sequence cluster. SNP=single nucleotide polymorphism. ST=sequence type.

significant. The SNP distances observed among *P aeruginosa* SCs were more variable, with six of 11 SCs having median distances of no more than two core-genome SNPs (≤6 whole-genome SNPs), and four of 11 having distances of more than 90 SNPs (for both core and whole genome).

For *E coli* SCs, the median SNP distance was more than 90 SNPs for eight of ten SCs in both core and whole genome, with the majority of SCs having SNP distances in the hundreds, indicating extensive between-patient diversity (figure 3; appendix 1 p 16). The exceptions were identical isolates, according to core-genome SNPs, of SC2_ST131 (carried by four patients), SC20_ST131 (two patients), SC40_ST410 (three patients), and SC4_ST95 (three patients). ST131 is a known health-care-associated clone,[35–37] and ST410 has previously been described as an emerging problem in hospitals.[38] *S aureus* also showed high diversity, with median distances of more than 100 SNPs for five of six SCs (both core-genome and whole-genome SNPs). However, there were a small number of identical BMGs (core-genome SNPs) shared between 24 patients in these SCs. *E faecalis* similarly showed considerable, albeit lower, diversity, with median distances of more than 32 core-genome SNPs for seven of nine SCs (>108 whole-genome SNPs). Identical BMGs were present in eight of nine SCs from this species according to both core-genome and whole-genome SNPs, and in two SCs these made up 25% of the pairs (core-genome SNPs).

The very low BMG SNP distances observed among the pathogen species strongly indicated hospital transmission, so we used the sequence data to estimate this burden for each pathogen. For each SC, we identified patients who were linked by at least one pair of BMGs, and then partitioned these occurrences into those that were putatively transmission-linked and those that were not. Putative transmission was highest within and between ICUs, followed by ICU–ordinary ward links (appendix 1 pp 17–18). The species with the highest percentage of transmission-linked occurrences were *A baumannii, K pneumoniae, P aeruginosa,* and *E faecium,* with more than 75% occurrences being putatively transmission-linked when calculated from core-genome SNPs (>81% from whole-genome SNPs; figure 4;

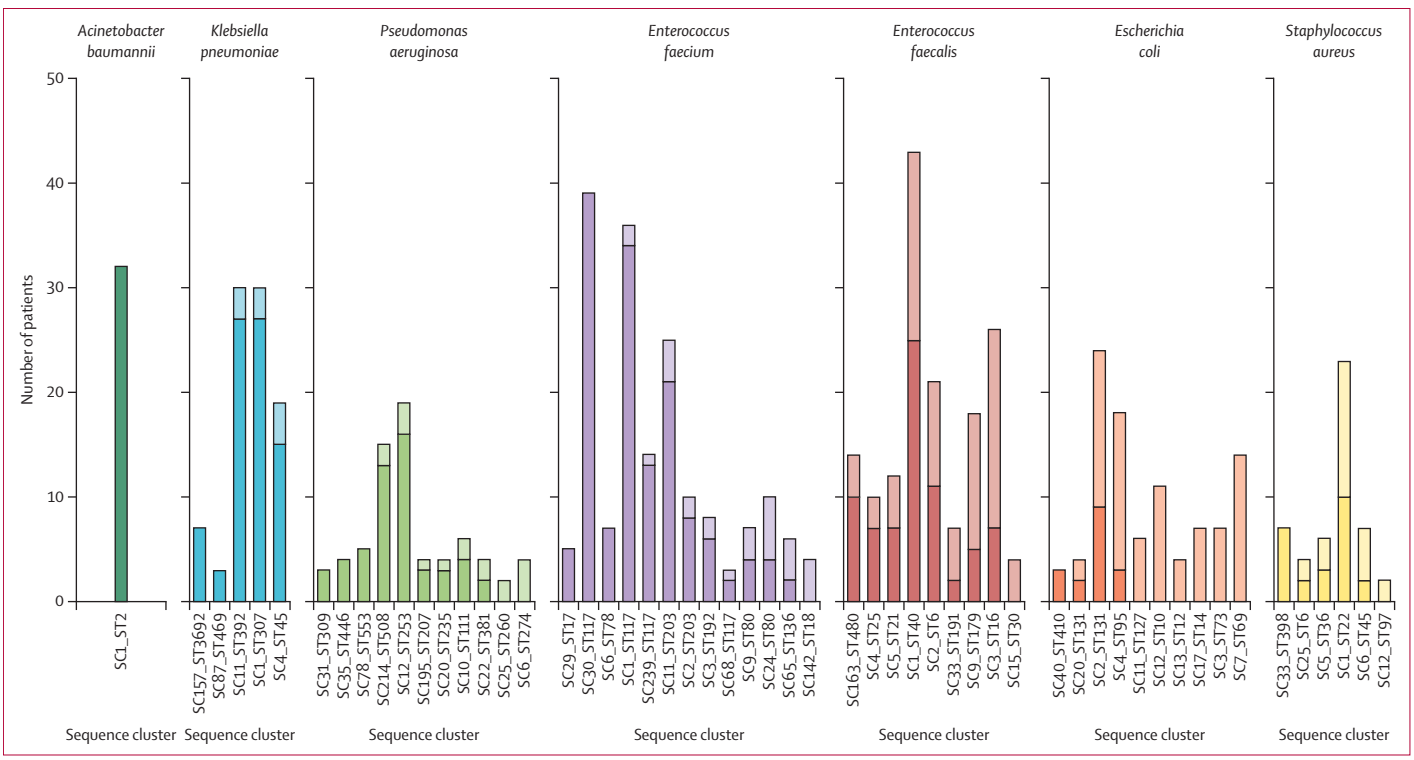

Figure 4: **The burden of hospital transmission based on estimates from core-genome SNP distances**
Core-genome SNP distances were used to infer networks of patients who carried binned mapped genomes that were closely related enough to be linked by transmission, according to SC-specific thresholds obtained from within-patient samples. For each SC, the occurrences of the binned mapped genomes among patients were partitioned into those that were transmission-linked (eg, those that were closely related enough to another binned mapped genome from a different patient; dark colour) and those that were not (light colour). As the height of the bar indicates the total number of patients from whom high-quality binned mapped genomes were obtained, the dark portion of the bar shows the proportion of these occurrences that were probably the result of hospital transmission. Only SCs that were detected in at least five patients are shown. SNP=single nucleotide polymorphism. ST=sequence type.

appendix 1 p 19). Within these species, the dominant SCs were *E faecium* SC1_ST117 and SC30_ST117, *A baumannii* SC1_ST2, and *K pneumoniae* SC1_ST307 and SC11_ST392, each with at least 25 linked patients. Approximately half of occurrences in *E faecalis* and *S aureus* were putatively transmission-linked (figure 4; appendix 1 p 19). *E coli* was least associated with transmission, with a few scattered examples of linked patients contributing to 17 (17%) of 98 occurrences (both core-genome and whole-genome SNPs). The results were highly concordant when calculated from core-genome and whole-genome SNPs, indicating that both measures captured similar information.

Motivated by the pattern of within-patient and between-patient diversity of *A baumannii* SC1_ST2, we compared all the *A baumannii* BMGs identified by mSWEEP-mGEMS with the reads from single-colony *A baumannii* isolates obtained previously from the same hospital and period[26] by combining them in a core-genome SNP phylogeny (figure 5A; microreact project). This showed two clearly separated clades of nearly identical genomes, belonging to the two SC1_ST2 subclones previously identified by Petazzoni and colleagues,[26] and both clades included samples from the two approaches (single colony and deep sequencing).

For more on the **microreact project** see https://microreact.org/project/7kCEtVo1po9Mb2vTxfX4uM-abau-demixed-single-colony-reads-core-sites-fasttree

Additionally, deep sequencing yielded reads from both subclones in samples from a single patient, whereas single colony whole-genome sequencing only identified one of the two, although colonies from multiple clinical samples were sequenced from this same patient. We therefore did a third round of the binning pipeline with a reference database consisting of the 91 single colony genomes belonging to SC1_ST2, labelled by subclone (SC1_1, n=68; SC1_2, n=23; figure 5B). This confirmed that both subclones were carried by the patient.

Discussion

Applying our deep-sequencing after culturing method to a cohort of hospital patients during the first COVID-19 wave in spring 2020, we discovered and characterised a substantial pathogen burden in the patient population, with 55 of 132 SCs being detected at least five times. The pathogen burden was highest in the ICUs, with *A baumannii*, *K pneumoniae*, *P aeruginosa*, and *E faecium* being significantly associated with ICUs. These pathogens were found to be colonising multiple body sites in a typical ICU patient; this is probably due to the high hospital prevalence of the pathogens and severely ill state of the patients, but it might also have been a consequence of the difficult situation regarding infection control during the

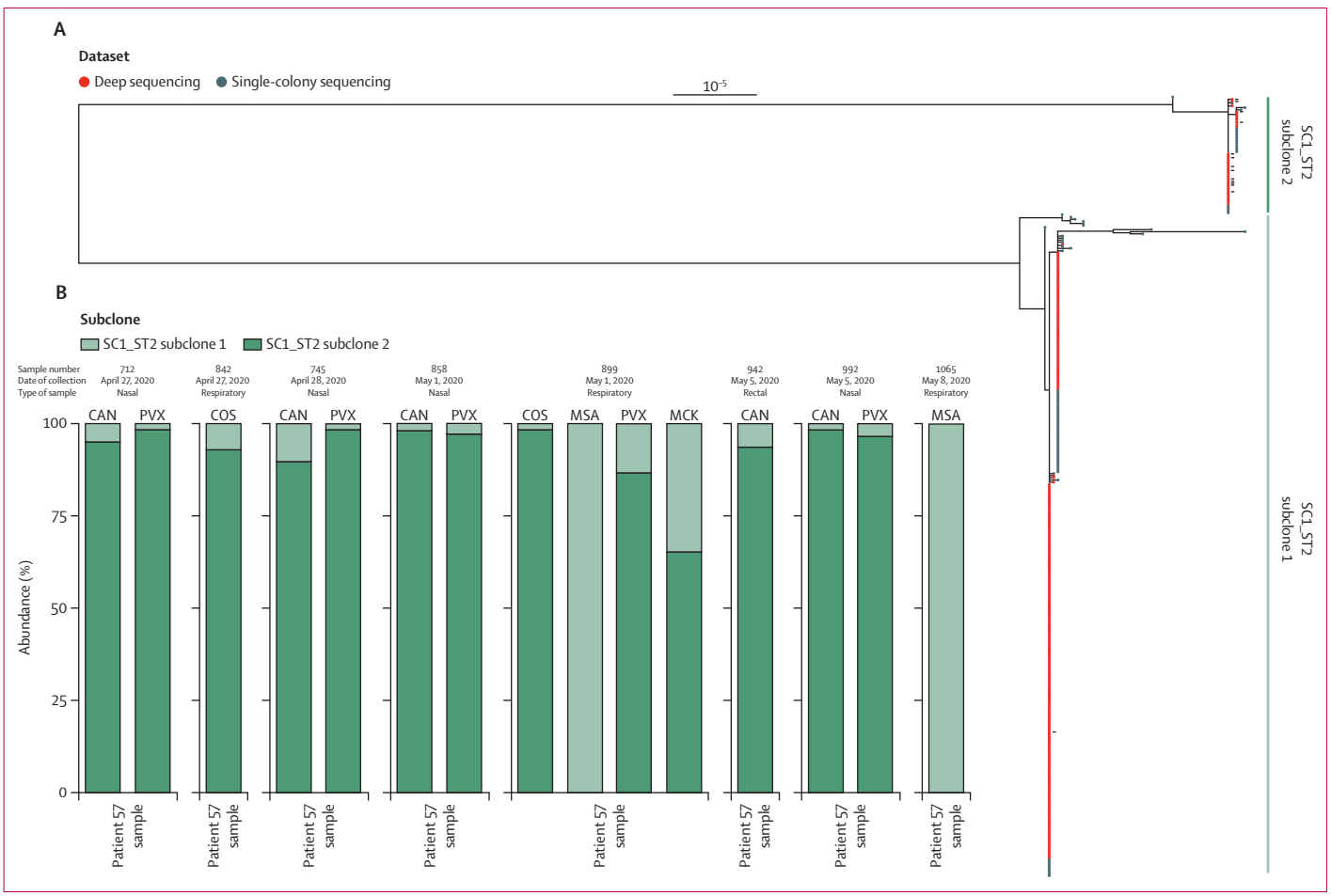

Figure 5: **Fine scale analysis of the *Acinetobacter baumannii* SC1_ST2 subclones**

(A) Phylogenetic tree of the two SC1_ST2 subclones. Single-colony whole-genome sequence data corresponding to *Acinetobater baumannii* from the same hospital as the current study (grey) was combined with the demixed deep sequencing reads from the current study (red). Reads were mapped to a common reference (SAMN07258611) by use of Snippy, and FastTree was used to generate a phylogenetic tree from the core-genome alignment (generalised time-reversible plus CAT model). The two SC1_ST2 subclades are labelled, and the scale bar indicates SNPs per site. (B) Reanalysis of the *A baumannii* SC1_ST2 subclones from a single patient who was found to carry both subclones. Each bar shows a single sequencing sample, and the bars are coloured by the subclone detected. The samples are grouped by the original swab sample (as one swab sample resulted in multiple sets of sequencing reads) and ordered by date of sample collection. This information, along with the sample type and culture media is shown above the bars. CAN=CHROMID Candida Agar. COS=Columbia agar + 5% sheep blood. MCK=McConkey agar. MSA=Mannitol Salt Agar. PVX=Chocolate agar + PolyViteX. SC=sequence cluster. ST=sequence type.

early pandemic. AMR gene prevalence was also very high, with several carbapenemase and ESBL genes being present in over 40% of patients on ICUs. Previous research from the same hospital with single colony whole-genome sequencing identified bla_{KPC-2} and bla_{OXA-23} as the dominant carbapenemase genes carried by *K pneumoniae* and *A baumannii*, respectively.[19,26] Such systemic presence of AMR bacteria highlights their ability to sustain themselves successfully in various niches in the hospital, as has been shown by other studies where sampling of the hospital environment was done.[4] We observed considerable differences in the genetic diversity among the pathogen species: *A baumannii* and *K pneumoniae* were dominated by very closely related BMGs; *E coli, E faecalis,* and *S aureus* were dominated by distantly related BMGs; and *P aeruginosa* and *E faecium* exhibited a mixture of both. Using these distances to infer transmission, we estimated that hospital transmission was likely to be a significant mode of acquisition for each of the pathogen species. For *A baumannii*, *K pneumoniae*, *P aeruginosa*, and *E faecium*, this was particularly significant with more than 75% of occurrences being potentially linked by transmission. These results emphasise the need for more comprehensive approaches to tracking the dissemination and evolution of these pathogens.

By comparing our results for *A baumannii* with the single colony-based whole-genome sequencing surveillance done during the same period at the study hospital,[26] we showed that our approach is able to detect and distinguish highly related strains with the sensitivity of the single-colony approach, and has the important added benefit of being able to delineate co-infections or co-colonisation within single patients.

Our study has some important limitations. Sequencing was done only on samples available through standard

diagnostic procedures, thus providing uneven sampling. Approximately one-third of the patients had already been admitted to the hospital when the sampling started, patients were not always sampled on admission, and we did not have access to samples taken before the start of the pandemic. Additionally, we did not perform power calculations before starting the sampling due to uncertainties about sample and personnel availability as a result of the COVID-19 pandemic. Another limitation is that although we detected AMR genes, we did not attempt to link them to their host species or strains. This was because many of these genes are known to be carried on plasmids and are present in multiple species, for example the $bla_{CTX\text{-}M\text{-}15}$ ESBL gene is common in both *K pneumoniae* and *E coli*. It would be attractive to extend the methodology to work with both long-read and short-read data in the future, as this would enable mobile genomic elements such as plasmids to be tracked. It is also possible that the thresholds used to validate the SC-level assignments by mSWEEP were too conservative, leading to missed assignments. However, we maintain that this step was key to ensure that incorrect assignments were not included.

The wealth of data generated by deep sequencing is an increasingly accessible way to explore bacterial variation in several ecological niches. Our study shows that this approach is also feasible for scrutiny of clinically relevant nosocomial bacteria, serving as a first step towards its further development and use as a research tool in clinical microbiology.

Contributors

JC, DS, SDB, and NRT conceptualised the study and obtained funding. JC, DS, and CJB administered the project. JC, DS, HAT, MC, SG, CJB, GT-H, TM, AKP, NM, RAG, SA-A, TK, DJ, SWL, CC, GAB, AH, ACS, RJLW, CM, GP, EJF, PC, NRT, and SDB obtained resources. HAT, MP, TM, DS, and JC developed the methodology. MC, SG, CM, GP, PC, and DS did laboratory investigations. HAT, MP, HP, GT-H, TM, SG, and JC did computational investigations. HAT, MC, SG, CJB, GT-H, TM, AKP, NM, RAG, SA-A, TK, DJ, SWL, CC, GAB, CM, GP, PC, DS, and JC curated the data. MP did the statistical analysis. HAT and JC did the bioinformatic analysis. HAT visualised the data. JC and GT-H accessed and verified the underlying data. HAT, JC, and DS wrote the manuscript. All authors reviewed and edited the manuscript, had full access to all the data in the study, and accept responsibility for the decision to submit for publication.

Declaration of interests

SWL received Robert Austrian Research Award sponsored by Pfizer, outside of the scope of this study. All other authors declare no competing interests.

Data sharing

Deep sequence data generated for this study are available at the ENA under project PRJEB39567. Sequence data of *E faecium* clade B (*Enterococcus lactis*) isolates are available at the ENA under project PRJEB28495.

Acknowledgments

This research was funded in part by the Wellcome Trust Grant number 206194. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. Additional funding was obtained from the European Research Council (grant 742158, JC, HAT, HP), Academy of Finland Flagship programme (JC, AH), Trond Mohn Foundation (BATTALION grant, JC, AKP, RAG, NM), Research Council of Norway (grant 2999131 JC, GT-H). The authors thank Pasquale Piemontese, Debora De Vitis, Chiara Rebuffa, Vincenzo Brunco, Marco Ardizzone, and Alessia Girello, the laboratory technicians of the Microbiology and Virology Unit at Fondazione IRCCS Policlinico San Matteo (the study hospital) for helping with the microbiology work.

References

1 Suetens C, Latour K, Kärki T, et al. Prevalence of healthcare-associated infections, estimated incidence and composite antimicrobial resistance index in acute care hospitals and long-term care facilities: results from two European point prevalence surveys, 2016 to 2017. *Euro Surveill* 2018; **23:** 1800516.

2 Cassini A, Högberg LD, Plachouras D, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect Dis* 2019; **19:** 56–66.

3 Harris SR, Feil EJ, Holden MTG, et al. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 2010; **327:** 469–74.

4 Doughty EL, Liu H, Moran RA, et al. Endemicity and diversification of carbapenem-resistant *Acinetobacter baumannii* in an intensive care unit. *Lancet Reg Health West Pac* 2023; **37:** 100780.

5 Köser CU, Holden MTG, Ellington MJ, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 2012; **366:** 2267–75.

6 Tong SYC, Holden MTG, Nickerson EK, et al. Genome sequencing defines phylogeny and spread of methicillin-resistant *Staphylococcus aureus* in a high transmission setting. *Genome Res* 2015; **25:** 111–18.

7 Qi C, Hountras P, Pickens CO, et al. Detection of respiratory pathogens in clinical samples using metagenomic shotgun sequencing. *J Med Microbiol* 2019; **68:** 996–1002.

8 Rodino KG, Toledano M, Norgan AP, et al. Retrospective review of clinical utility of shotgun metagenomic sequencing testing of cerebrospinal fluid from a US tertiary care medical center. *J Clin Microbiol* 2020; **58:** e01729-20.

9 Oechslin CP, Lenz N, Liechti N, et al. Limited correlation of shotgun metagenomics following host depletion and routine diagnostics for viruses and bacteria in low concentrated surrogate and clinical samples. *Front Cell Infect Microbiol* 2018; **8:** 375.

10 Whelan FJ, Waddell B, Syed SA, et al. Culture-enriched metagenomic sequencing enables in-depth profiling of the cystic fibrosis lung microbiota. *Nat Microbiol* 2020; **5:** 379–90.

11 Mäklin T, Kallonen T, David S, et al. High-resolution sweep metagenomics using fast probabilistic inference. *Wellcome Open Res* 2021; **5:** 14.

12 Mäklin T, Kallonen T, Alanko J, et al. Bacterial genomic epidemiology with mixed samples. *Microb Genom* 2021; **7:** 000691.

13 Mäklin T, Thorpe HA, Pöntinen AK, et al. Strong pathogen competition in neonatal gut colonisation. *Nat Commun* 2022; **13:** 7417.

14 Tonkin-Hill G, Ling C, Chaguza C, et al. Pneumococcal within-host diversity during colonization, transmission and treatment. *Nat Microbiol* 2022; **7:** 1791–804.

15 Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014; **15:** R46.

16 Horesh G, Blackwell GA, Tonkin-Hill G, Corander J, Heinz E, Thomson NR. A comprehensive and high-quality collection of *Escherichia coli* genomes and their genes. *Microb Genom* 2021; **7:** 000499.

17 Arredondo-Alonso S, Top J, McNally A, et al. Plasmids shaped the recent emergence of the major nosocomial pathogen *Enterococcus faecium*. *MBio* 2020; **11:** e03284-19.

18 Pöntinen AK, Top J, Arredondo-Alonso S, et al. Apparent nosocomial adaptation of *Enterococcus faecalis* predates the modern hospital era. *Nat Commun* 2021; **12:** 1523.

19 Thorpe HA, Booton R, Kallonen T, et al. A large-scale genomic snapshot of *Klebsiella* spp isolates in northern Italy reveals limited transmission between clinical and non-clinical settings. *Nat Microbiol* 2022; **7:** 2054–67.

20 Gladstone RA, Lo SW, Lees JA, et al. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine* 2019; **43:** 338–46.

21 Lees JA, Harris SR, Tonkin-Hill G, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res* 2019; **29:** 304–16.

22 Alanko JN, Vuohtoniemi J, Mäklin T, Puglisi SJ. Themisto: a scalable colored *k*-mer index for sensitive pseudoalignment against hundreds of thousands of bacterial genomes. *Bioinformatics* 2023; **39:** i260–69.

23 Ondov BD, Treangen TJ, Melsted P, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016; **17:** 132.

24 Hunt M, Mather AE, Sánchez-Busó L, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom* 2017; **3:** e000131.

25 Alcock BP, Huynh W, Chalil R, et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res* 2023; **51:** D690–99.

26 Petazzoni G, Bellinzona G, Merla C, et al. The COVID-19 Pandemic sparked off a large-scale outbreak of carbapenem-resistant *Acinetobacter baumannii* from the endemic strains at an Italian hospital. *Microbiol Spectr* 2023; **11:** e0450522.

27 Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 2010; **5:** e9490.

28 Hamidian M, Nigro SJ. Emergence, molecular mechanisms and global spread of carbapenem-resistant *Acinetobacter baumannii*. *Microb Genom* 2019; **5:** e000306.

29 Nicolas-Chanoine M-H, Bertrand X, Madec J-Y. *Escherichia coli* ST131, an intriguing clonal group. *Clin Microbiol Rev* 2014; **27:** 543–74.

30 David S, Reuter S, Harris SR, et al. Epidemic of carbapenem-resistant *Klebsiella pneumoniae* in Europe is driven by nosocomial spread. *Nat Microbiol* 2019; **4:** 1919–29.

31 Wiehlmann L, Cramer N, Tümmler B. Habitat-associated skew of clone abundance in the *Pseudomonas aeruginosa* population. *Environ Microbiol Rep* 2015; **7:** 955–60.

32 Hsu L-Y, Harris SR, Chlebowicz MA, et al. Evolutionary dynamics of methicillin-resistant *Staphylococcus aureus* within a healthcare system. *Genome Biol* 2015; **16:** 81.

33 Kinross P, Petersen A, Skov R, et al. Livestock-associated meticillin-resistant *Staphylococcus aureus* (MRSA) among human MRSA isolates, European Union/European Economic Area countries, 2013. *Euro Surveill* 2017; **22:** 16-00696.

34 Merla C, Kuka A, Petazzoni G, et al. Livestock-associated methicillin-resistant *Staphylococcus aureus* in inpatients: a snapshot from an Italian hospital. *J Glob Antimicrob Resist* 2022; **30:** 10–15.

35 Brodrick HJ, Raven KE, Kallonen T, et al. Longitudinal genomic surveillance of multidrug-resistant *Escherichia coli* carriage in a long-term care facility in the United Kingdom. *Genome Med* 2017; **9:** 70.

36 Mills EG, Martin MJ, Luo TL, et al. A one-year genomic investigation of *Escherichia coli* epidemiology and nosocomial spread at a large US healthcare network. *Genome Med* 2022; **14:** 147.

37 Roberts LW, Hoi LT, Khokhar FA, et al. Genomic characterisation of multidrug-resistant *Escherichia coli, Klebsiella pneumoniae*, and *Acinetobacter baumannii* in two intensive care units in Hanoi, Viet Nam: a prospective observational cohort study. *Lancet Microbe* 2022; **3:** e857–66.

38 Roer L, Overballe-Petersen S, Hansen F, et al. *Escherichia coli* sequence type 410 is causing new international high-risk clones. *MSphere* 2018; **3:** e00337-18.