

# The interplay between community and hospital *Enterococcus faecium* clones within health-care settings: a genomic analysis

Sebastiaan J van Hal, Rob J L Willems, Theodore Gouliouris, Susan A Ballard, Teresa M Coque, Anette M Hammerum, Kristin Hegstad, Mette Pinholt, Benjamin P Howden, Surbhi Malhotra-Kumar, Guido Werner, Katsunori Yanagihara, Ashlee M Earl, Katherine E Raven, Jukka Corander, Rory Bowden, the Enterococcal Study Group\*



## Summary

**Background** The genomic relationships among *Enterococcus faecium* isolates are the subject of ongoing research that seeks to clarify the origins of observed lineages and the extent of horizontal gene transfer between them, and to robustly identify links between genotypes and phenotypes. *E faecium* is considered to form distinct groups—A and B—corresponding to isolates derived from patients who were hospitalised (A) and isolates from humans in the community (B). The additional separation of A into the so-called clades A1 and A2 remains an area of uncertainty. We aimed to investigate the relationships between A1 and non-A1 groups and explore the potential role of non-A1 isolates in shaping the population structure of hospital *E faecium*.

**Methods** We collected short-read sequence data from invited groups that had previously published *E faecium* genome data. This hospital-based isolate collection could be separated into three groups (or clades, A1, A2, and B) by augmenting the study genomes with published sequences derived from human samples representing the previously defined genomic clusters. We performed phylogenetic analyses, by constructing maximum-likelihood phylogenetic trees, and identified historical recombination events. We assessed the pan-genome, did resistome analysis, and examined the genomic data to identify mobile genetic elements. Each genome underwent chromosome painting by use of ChromoPainter within FineSTRUCTURE software to assess ancestry and identify hybrid groups. We further assessed highly admixed regions to infer recombination directionality.

**Findings** We assembled a collection of 1095 hospital *E faecium* sequences from 34 countries, further augmented by 33 published sequences. 997 (88%) of 1128 genomes clustered as A1, 92 (8%) as A2, and 39 (4%) as B. We showed that A1 probably emerged as a clone from within A2 and that, because of ongoing gene flow, hospital isolates currently identified as A2 represent a genetic continuum between A1 and community *E faecium*. This interchange of genetic material between isolates from different groups results in the emergence of hybrid genomes between clusters. Of the 1128 genomes, 49 (4%) hybrid genomes were identified: 33 previously labelled as A2 and 16 previously labelled as A1. These interactions were fuelled by a directional pattern of recombination mediated by mobile genetic elements. By contrast, the contribution of B group genetic material to A1 was limited to a few small regions of the genome and appeared to be driven by genomic sweep events.

**Interpretation** A2 and B isolates coming into the hospital form an important reservoir for ongoing A1 adaptation, suggesting that effective long-term control of the effect of *E faecium* could benefit from strategies to reduce these genomic interactions, such as a focus on reducing the acquisition of hospital A1 strains by patients entering the hospital.

**Funding** Wellcome Trust.

**Copyright** © 2022 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

## Introduction

*Enterococcus faecium* is a pathogen of global significance, causing serious hospital-associated infections. The impact of these infections is augmented by the high rates of acquired resistance, including to first-line antibiotics such as vancomycin.<sup>1,2</sup> The control of hospital-acquired *E faecium* infection is challenging because routes of transmission are often unclear, and the pathogen tends to persist in the hospital environment despite infection

control campaigns that have successfully reduced the spread of other harmful nosocomial pathogens.

*E faecium* genomes are considered to form two distinct groups, A and B, corresponding to isolates originating from humans in a hospital setting (A) and those in a community setting (B).<sup>3,4</sup> Cluster A can be further split into two subgroups referred to as clades A1 and A2, on the basis of a detailed analysis in which A2 isolates were mostly collected from animals.<sup>4</sup> Subsequent

*Lancet Microbe* 2022;  
3: e133–41

Published Online  
January 18, 2021  
[https://doi.org/10.1016/S2666-5247\(21\)00236-6](https://doi.org/10.1016/S2666-5247(21)00236-6)

\*Collaborators listed at the end of the Article

Department of Infectious Diseases and Microbiology, Royal Prince Alfred Hospital, Sydney, NSW, Australia (S J van Hal PhD); Central Clinical School, University of Sydney, Sydney, NSW, Australia (S J van Hal); Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, Netherlands (R J L Willems PhD); Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK (T Gouliouris PhD); Microbiological Diagnostic Unit Public Health Laboratory, The University of Melbourne at The Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia (S A Ballard, B P Howden PhD); Department of Microbiology, Ramón y Cajal University Hospital and Ramón y Cajal Health Research Institute, Madrid, Spain (T M Coque PhD); Network Research Centre for Epidemiology and Public Health, Madrid, Spain (T M Coque); Statens Serum Institut, Copenhagen, Denmark (A M Hammerum PhD); Norwegian National Advisory Unit on Detection of Antimicrobial Resistance, University Hospital of North-Norway, Department of Microbiology and Infection Control, Tromsø, Norway (K Hegstad PhD); Department of Clinical Microbiology, Hvidovre University Hospital, Hvidovre, Denmark (M Pinholt PhD); Laboratory of Medical Microbiology, Vaccine & Infectious Disease Institute,

Universiteit Antwerpen,  
Wilrijk, Belgium  
(S Malhotra-Kumar PhD);  
National Reference Centre for  
Staphylococci and Enterococci,  
Division of Nosocomial  
Pathogens and Antibiotic  
Resistances, Department of  
Infectious Diseases, Robert  
Koch Institute, Wernigerode  
Branch, Wernigerode, Germany  
(G Werner PhD); Department of  
Laboratory Medicine, Nagasaki  
University Graduate School of  
Biomedical Sciences, Nagasaki,  
Japan (K Yanagihara PhD);  
Infectious Disease &  
Microbiome Program, Broad  
Institute, Cambridge, MA, USA  
(A M Earl PhD); Department of  
Medicine, University of  
Cambridge, Cambridge, UK  
(K E Raven PhD); Department of  
Biostatistics, University of  
Oslo, Oslo, Norway  
(J Corander PhD); Parasites and  
Microbes, Wellcome Sanger  
Institute, Saffron Walden, UK  
(J Corander); The Walter and  
Eliza Hall Institute of Medical  
Research, Parkville, VIC,  
Australia (R Bowden PhD);  
Department of Medical  
Biology, University of  
Melbourne, Melbourne, VIC,  
Australia (R Bowden); Wellcome  
Centre for Human Genetics,  
University of Oxford, Oxford,  
UK (R Bowden)

Correspondence to:  
Sebastiaan J van Hal, Royal  
Prince Alfred Hospital,  
Sydney, NSW 2050, Australia  
sebastiaan.vanhal@health.  
nsw.gov.au

### Research in context

#### Evidence before this study

We searched PubMed for studies in English published between Jan 1, 1995, and Dec 31, 2019, using the terms “*Enterococcus faecium*”, and “sequencing” or “population structure”. We filtered the 124 and 76 studies identified including only those that used whole genome sequencing in datasets of more than 50 isolates. Several additional studies were identified by sifting through manuscript references. Following the initial description of *Enterococcus faecium* groupings A1, A2, and B, current studies have questioned the position or existence of A2 within this overall population structure.

#### Added value of this study

We showed that A1 emerged as a clone from within A2 and that, because of ongoing gene flow, hospital isolates currently identified as A2 represent a genetic continuum between A1 and community *E faecium*. By contrast, the contribution of B group genetic material to A1 is limited to a few small

regions of the genome and appears to be driven by genomic sweep events. A2 isolates entering the hospital form an important reservoir for ongoing A1 adaptation, fuelled by a directional pattern of recombination mediated by mobile genetic elements.

#### Implications of all the available evidence

New hospital *E faecium* clones will continue to emerge, drawing genetic material—including antimicrobial resistance genes—from community clones entering the health-care system. Control of vancomycin-resistant *E faecium* in the longer term will thus remain a challenge, especially if infection control interventions are limited to stopping patient-to-patient transmission of strains that are already present in the hospital, rather than reducing opportunities for new and potentially adaptively favoured forms to arise, which can then acquire vancomycin resistance and present a direct threat.

investigations of larger sample sets have confirmed the expansion of a tightly clustered group A (A1), comprising the majority of hospital isolates. However, these data did not support a single distinct A2 clade, but rather numerous A2 subgroups that appear to be largely host-specific, including a human A2 or non-A1 clade.<sup>5–8</sup> Controversy remains on whether A2 constitutes a single clade, paraphyletic with A1, or rather a series of earlier-branching lineages of the A group. Irrespective of this distinction, just as for group B, human A2 carriage is mostly seen in the gastrointestinal tract of individuals from the community, rather than those admitted to hospital. At the level of individuals, carriage is dynamic: strains are often observed to be replaced by the predominant hospital A1 clade on entering the health-care system or replaced by community strains upon leaving it.

Considering that *E faecium* is highly recombinogenic, genetic interactions between distinct *E faecium* groups might be important in the continuing evolution of hospital strains. However, this interaction remains largely undefined. In this study, we aimed to analyse a large global sample of hospital-isolated *E faecium* to resolve the interplay between A1 and non-A1 groups and explore the potential role of non-A1 isolates in shaping the population structure of hospital *E faecium*.

## Methods

### Isolate collection and data availability

We invited groups that had previously published *E faecium* genome data to contribute short-read sequence data. Participants were asked to select both vancomycin-susceptible and vancomycin-resistant *E faecium* sequence data from patients who were hospitalised and to ensure that isolates from the same hospital were not enriched by

known outbreaks, on the basis of local epidemiological data.

### Classification of sequences

We trimmed short-read sequence data passing quality metrics using trimmomatic (version 0.38) before mapping against reference Aus0004 (GenBank CP003351, *vanB*-positive Australian, hospital, clade A1 isolate) using bwa (version 0.7.17-r1188).<sup>9,10</sup> We identified single nucleotide polymorphisms (SNPs) using FreeBayes (version 1.3.2-dirty) with variants considered present if the proportion of reads supporting the allele was greater than 90%.<sup>11</sup> All SNPs were subsequently included, provided that more than 75% of isolates had data across that site. Insertions and deletions were excluded from the analysis.

We used informative SNPs to cluster isolates using the Hierarchical Bayesian Analysis of Population Structure (hierBAPS) software.<sup>12</sup> We separated the hospital-based isolate collection into three groups, or so-called clades, known as A1, A2, and B, by augmenting the study genomes with 33 published sequences derived from human samples representing the previously defined genomic clusters clade B (n=8), A2 (n=11), and A1 (n=14).<sup>4</sup> Subsequent clade assignment of study genomes was based on group comembership. In multiple runs of hierBAPS, groups A and B were recovered when the maximum level of clusters was set at two. All subsequent runs with increasing maximum level of clusters from two to 20 converged on the same posterior clustering estimate, in which group A split into two subgroups corresponding to the published A1 and A2 lineages. To exclude a possible misclassification error based on the species of origin, the analysis was repeated including an additional 19 non-human (animal and environmental) genomes, obtaining concordant

results, in which human and non-human A2 isolates joined a single group.<sup>4</sup>

### Phylogenetic analysis and detection of recombination

We identified historic recombination events using ClonalFrameML (version 1.12) with an iterative approach.<sup>2,13</sup> Identified recombination events on each branch were masked before rebuilding the tree for phylogenetic reconstructions against the original dataset. We also assessed recombination using Gubbins (version 2.4.1), with all maximum likelihood phylogenetic trees constructed with a GTR+ model implemented in RaxML (version 8.1.3).<sup>14,15</sup> Bootstrap analysis was implemented on the final phylogenetic trees with 100 independent maximum likelihood runs, with branch and node support inferred after 1000 bootstrap replicates.

### Pan-genome, resistome analysis, and identification of mobile genetic elements

We determined the pan-genome using contigs assembled with SPAdes, version 3.13.1, under the careful option.<sup>16</sup> Contigs smaller than 2000 bp were removed and discarded from individual assemblies, which were annotated using prokka, version 1.13, before pan-genome discovery with Roary.<sup>17,18</sup>

We identified the in-silico resistome using the de-novo assemblies, requiring at least 80% coverage and 95% sequence identity to a specific gene, using AMRFinder specifying *E faecium*.<sup>15</sup>

We examined the genomic data for mobile genetic elements (MGEs) using two approaches. The first identified the locations of chromosomal MGEs of the reference with use of the online tools ISFinder and RAST.<sup>19,20</sup> Mapped sequences were subsequently interrogated with a similar MGE considered present, provided that more than 85% coverage and more than 90% similarity was obtained across the query region. The second approach used the prokka annotations and Roary output to define MGEs within the assembled contigs.

All genome-wide analyses were undertaken on the same set of 10 kb windows relative to the reference sequence. Regions associated with the *vanB* transposon (2835430–2869240) were excluded.

### Chromosome painting and FineSTRUCTURE analyses

The ChromoPainter tool embedded within FineSTRUCTURE (version 2.0.3) was used to paint each genome, setting a starting mean recombination rate of  $2.3 \times 10^{-7}$ .<sup>21</sup> This method assigns every section of an isolate's genome as a succession of haplotype fragments originating from the closest matching sequence in the collection, or donor haplotypes with donor switches occurring at recombination breakpoints. Each painted genome is subsequently represented as a probability matrix of probable donors per 10 kb window. These painted genomes were aggregated by defined clades—A1, A2, and B—with similar admixture patterns across most of

their genome representing a continuous line of ancestry. Hybrids were defined when the dominant proportion of painted loci arose from outside their assigned clade. Different admixture cutoffs were used to define a hybrid group, and a value higher than 30% was found to be the most informative when mapping back to the phylogeny and determining intergroup relationships.

### Recombination directionality

We further assessed highly admixed regions using fastGEAR, which identifies recombination events by looking for similar genomic segments between diverse clusters of data.<sup>22</sup> Using a hidden Markov model, fastGEAR assigns the origin of DNA fragments relative to other sequences through identification of lineages. Recent recombination events do not affect all genomes, and thus directionality can be inferred. By contrast, so-called remote events reveal shared ancestry between genomes, consistent with segments arising from outside the sample set—for which the directionality of transfer cannot be inferred.

### Statistical analysis

We used the Kruskal-Wallis test for between-genomic group comparisons, with a p value lower than 0.05 considered significant. These comparisons were done with the R stats package, version 3.6.3.

### Role of the funding source

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

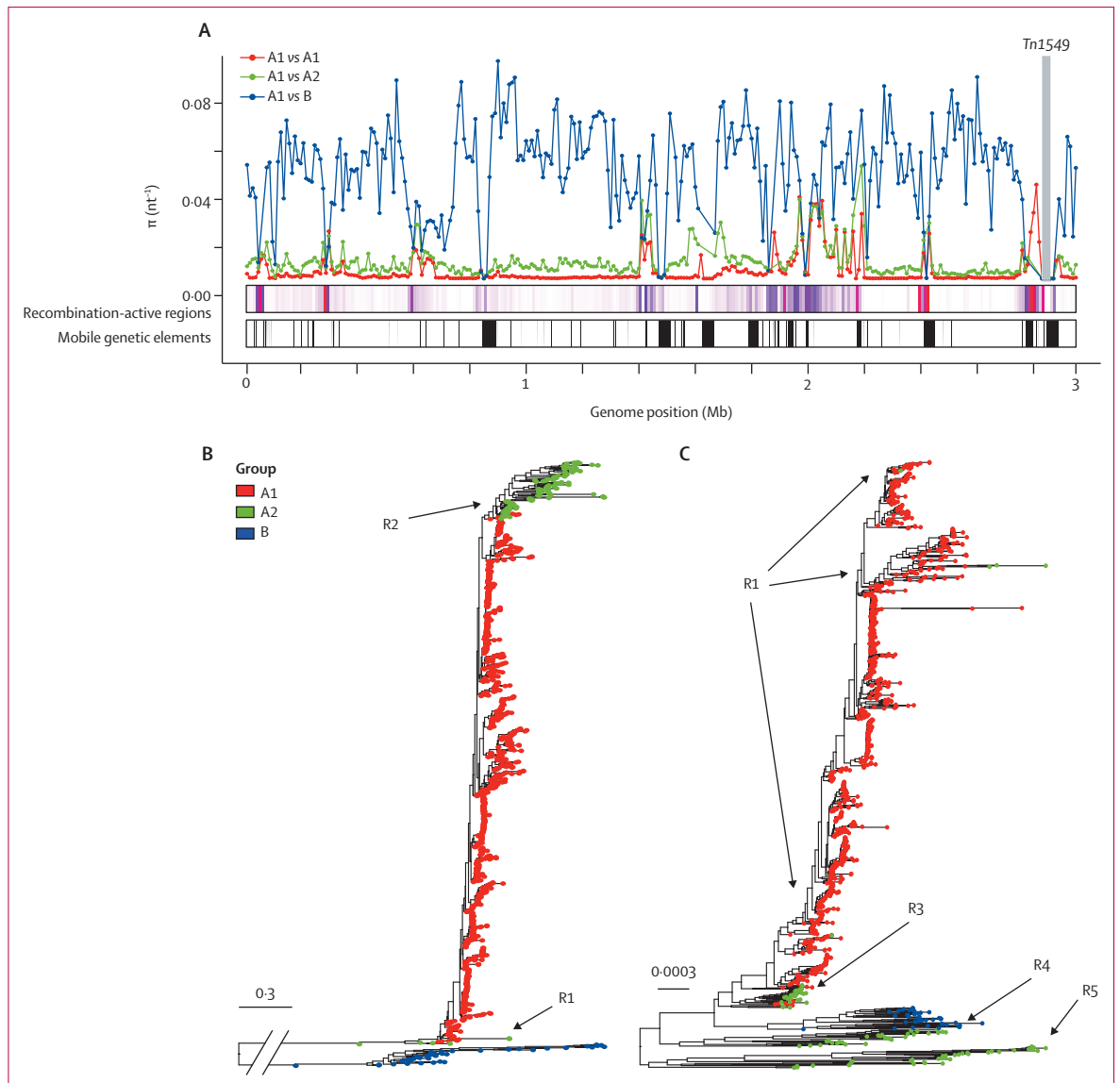
### Results

We assembled a collection of 1095 isolate sequences from 34 countries, from 1957 to 2016, of which 321 were newly sequenced and uploaded (PRJNA636894), with the remaining sequences (774) downloaded from the National Center for Biotechnology Information. An additional 52 sequences were downloaded and represented the previous genomic basis for *E faecium* groupings (appendix pp 3–45). Most isolates (770 [70%] of 1095) came from clinical infections, with the remainder split between colonisation (143 [13%]) and unknown (182 [17%]) sources.

With the Aus0004 genome as a reference to identify 261 543 variant positions across the 2.96 Mb genome, we grouped the genomes using hierBAPs.<sup>12</sup> In terms of previously defined genomic groups, 997 (88%) of 1128 genomes clustered as A1, 92 (8%) as A2, and 39 (4%) as B. Despite all study genomes originating from patients who were hospitalised, previous well established associations would indicate that A2 and B genomes represent community *E faecium* isolates entering the hospital.<sup>4,7</sup>

Plotting molecular diversity in 10 kb windows along the genome supported the distinct identities of the

See Online for appendix



**Figure 1: *Enterococcus faecium* population structure**

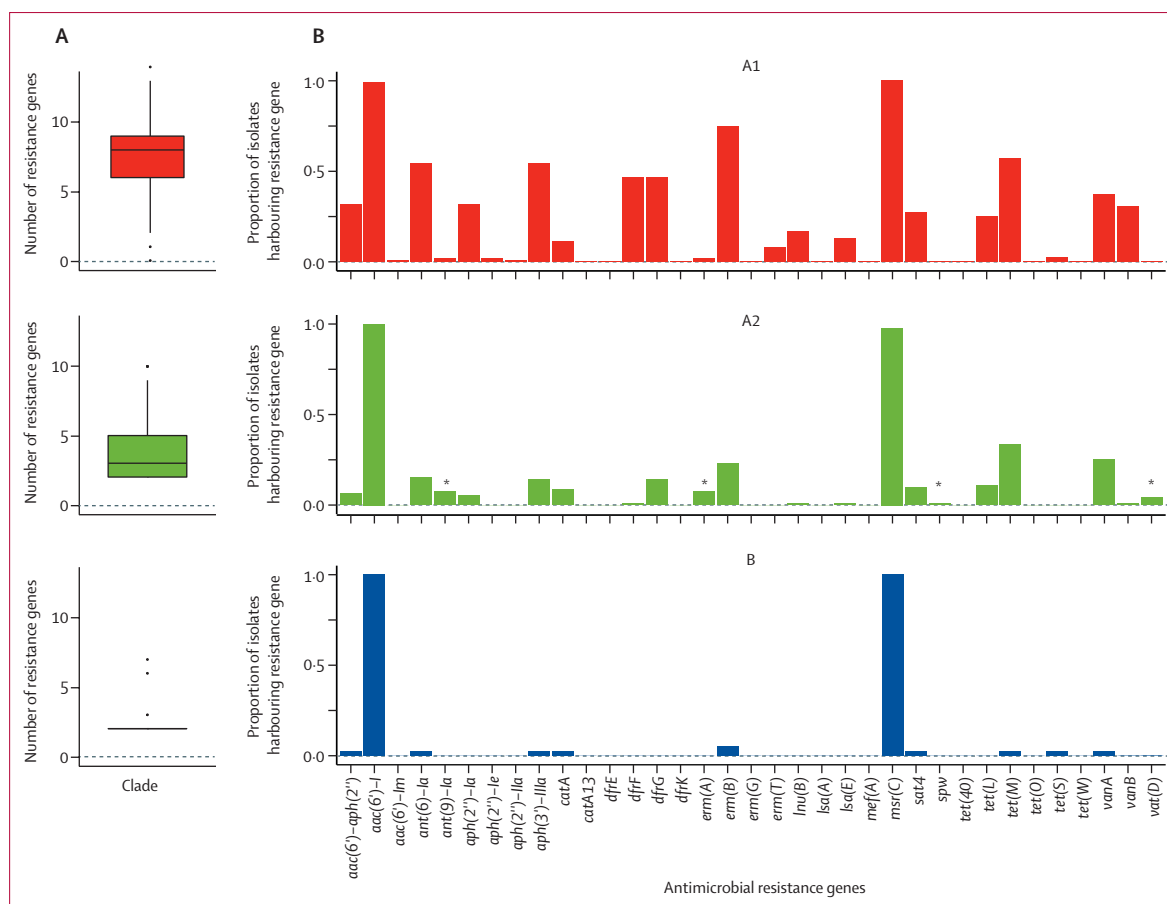
(A) Mean pairwise diversity ( $\pi$ ,  $\text{nt}^{-1}$ ) within (A1 vs A1) and between (A1 vs A2 and A1 vs B) groups, plotted in 10 kb windows across the genome; *Tn1549*, carrying the *vanB* gene cluster, was excluded from the analysis and is depicted by the labelled grey box; regions where between-group diversity converges coincide in many instances with recombination-active regions and mobile genetic elements. (B, C) Maximum-likelihood phylogenies of 1128 *E. faecium* genomes, before (B) and after (C) masking of recombination events identified using ClonalFrameML, with tips coloured by group assigned using Bayesian Analysis of Population Structure; subgroups of A2 isolates are labelled R1 to R5 according to their tree position; isolates labelled R2 in (B) split into subgroups R3, R4, and R5 in (C).

3 groups, with mean between-group diversity approximately five times (A1 vs A2) and 40 times (A1 vs B) that within A1, across the majority of the genome, whereas the mean diversity was 28 times in A2 versus B. However, in several regions of the genome, within-group and between-group diversity converged, corresponding to regions of recombination, coinciding with and mediated by MGEs (figure 1A).

To investigate relationships between the three groups, we constructed a maximum-likelihood phylogeny, which confirmed the deep split between groups A

and B. A2 isolates separated into two subgroups, labelled R1 and R2, occupying distinct branches from A1 (figure 1B).

We observed a relative rate of recombination to mutation greater than 4 across the genome, consistent with high rates of recombination (appendix p 2).<sup>23</sup> Irrespective of method used (Gubbins vs ClonalFrameML), we identified more historic recombination events in comparisons of B with A1 genomes than in comparisons of A1 with A2 genomes. The masking of identified recombination events excluded on average 2.5 Mb, or 85% of B genomes,



**Figure 2: Antimicrobial resistance**

(A) Box and whisker plot of the number of resistance genes carried by each of the *E faecium* clades A1, A2, and B. (B) The proportion of isolates from each clade harbouring any specific resistance gene (labelled on the x-axis). \*Indicates a gene that is preferentially carried by A2 clade compared with A1.

distorting the inferred relationships between B and A genomes. Because of the shorter time depth, this effect was less marked when comparing A1 with A2 isolates alone (with an average of 0.9 Mb or 33% of genome masked as a result of recombination), with the R2 subgroup split into distinct subgroups (ancestral to A1) and labelled R3, R4, and R5 (figure 1C).

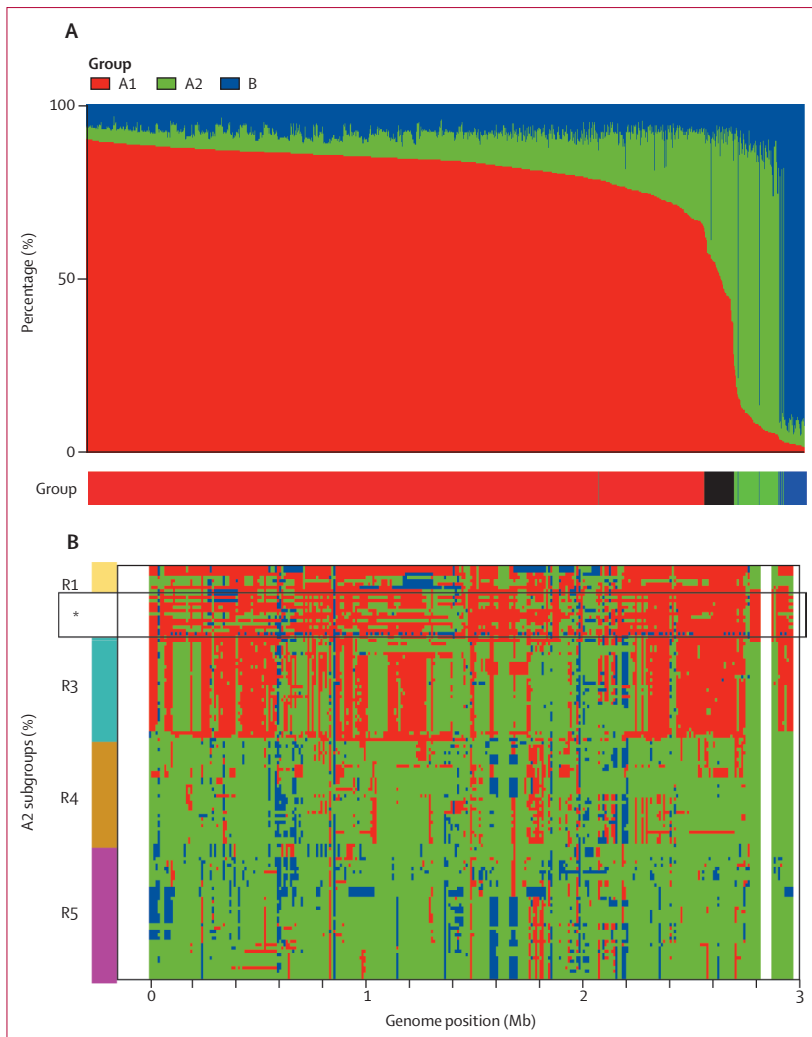
A2 hospital *E faecium* genomes were identified in 19 (56%) of 34 sampled countries, indicating that observations about the distribution of A2 were not driven by a single region of the world. Similarly, hospital B isolates were not restricted to any one country.

A pan-genome analysis identified 1854 core genes (each found in more than 95% of all isolates) from a total of 18468 detected genes. Several genes (299 in B and 23 in A2) were restricted (found in fewer than 5% of A1 isolates and more than 50% of either A2 or B) to non-A1 isolates (appendix pp 46–57). Most of such genes were annotated as hypothetical proteins. The remaining genes were found only in B isolates and included genes associated with carbohydrate metabolism, phosphotransferase systems, and iron transporters, recapitulating previous associations.<sup>4</sup>

Notably, genomes belonging to A2 and B contained genes associated with MGEs not found in A1 isolates. Non-A1 isolates harboured significantly fewer resistance genes ( $p < 0.001$ ) than hospital isolates; all isolates harboured an aminoglycoside resistance (*aac(6)-I*) gene and an efflux pump (*msrC*) encoding macrolide resistance. We observed an incremental increase in the median number of resistance genes, from two in B and three in A2, to eight in A1 (figure 2A). Vancomycin resistance was seen in 19 (21%) of A2 isolates and in one (3%) of 39 B isolates, consistent with gene movement from A1 to non-A1 isolates from patients entering the health-care system. Nevertheless, non-A1 isolates were a small but important contribution to the overall hospital antimicrobial resistance reservoir with the detection of *qacC*, encoding quaternary ammonium compound resistance, found only in two A2 genomes, and several other resistance genes preferentially carried by A2 genomes (figure 2B).

Using ChromoPainter, we painted the 1128 genomes, clustering isolates by their hierBAPs assignment within the same 10 kb windows.<sup>21</sup> Across most of the genome, the A1, A2, and B groups maintained their own identities





**Figure 3: Chromosome painting assessing admixture in *Enterococcus faecium***

(A) Total ancestry contribution from each of the A1, A2, and B groups by isolate, arranged in decreasing order of A1 ancestry; the legend beneath shows Bayesian Analysis of Population Structure group assignment, with hybrids denoted in black. (B) Chromosome painting of fragments along the genome for A2 isolates arranged by position on phylogeny (figure 1C), with subgroups shown by the coloured bar; R1 subgroup includes isolates assigned to A2 that have little estimated A2 ancestry; R4 and R5 are considered pure A2 isolates, whereas isolates within the R3 subgroup represent a probable clonal hybrid between A1 and A2. \*This panel contains A1-labelled hybrid isolates falling within R3 on the phylogeny, which share a pattern of co-ancestry, distinct from other R3 isolates.

despite significant recombination that can shuffle phylogenetic relationships. However, within our collection, we found a small proportion of isolates (49 [4.3%] of 1128) in which more than 30% of the genomic content could be traced to an alternative group. The 33 previously A2-labelled and 15 A1-labelled isolates (48 [4.3%] of 1128 sampled isolates) represent hybrid genomes that reflect a continuum of ancestry between A1 and A2 (figure 3A). We detected a single A1-labelled hybrid with B, suggesting that a similar continuum exists between A1 and B. We found no A2–B hybrids in our hospital-derived dataset, possibly reflecting the relatively reduced opportunities for these groups to interact within the hospital.

We noted no distinctive clinical or genomic characteristics among the 49 hybrid isolates, of which 31 (63%) represented infection isolates and ten (21%) harboured the vancomycin-resistance gene cluster.

Patterns of admixture between groups also drive overall phylogenetic relationships, with the position of A2 subgroups completely explained by their distinct genomic compositions and differing contributions of A1 ancestry. The R1 and R3 phylogenetic subgroups comprised hybrids with varying estimated proportions of A1 and A2 ancestry, whereas R4 and R5 accounted for non-hybrid isolates representing some of the diversity of the A2 group (figure 3B). By contrast with R1 isolates, the tight grouping of R3 genomes on the phylogeny suggests a single hybridisation event followed by clonal expansion. Accordingly, the A1 isolates that sit phylogenetically within the R3 genomes (figure 3B) are A1-labelled hybrids with distinct co-ancestry patterns from A2-labelled R3 isolates, across a minority of the genome.

Chromosome painting also revealed subsequent diversification of hybrid clones through the import of genetic material, including material representing group B sequences (figure 3B). These patterns confirmed that large-scale recombination was responsible for establishing distinct new genetic identities in A1 *E faecium*. Together with the presumption that A1 isolates primarily circulate in hospitals, A1–A2 hybrids continue to arise within and form part of the hospital-resident microbiota, and thus provide potential raw material for future forms of hospital *E faecium*, distinct from the more transient (unrecombined) A2 and B lineages.

Several short segments of the genome contained sequences from a different group, including several regions of high admixture (more than 75% of isolates in any one group contained imported sequences from another group; figure 4A). These regions were found in all groups, corresponding to genomic segments of converging diversity and were associated with MGEs (figure 1A).

Using fastGEAR, we examined the direction of horizontal transfer in two representative highly admixed regions (figure 4B).<sup>22</sup> Focusing on so-called remote events (those that are mapped back to the root of the tree) in region 1, a single haplotype was observed across all hospital isolates. Genes of interest in this region included the *liaS*, *liaF*, and *liaR* gene cluster that forms part of a three-component regulatory system involved in cell envelope stress responses, and the polysaccharide biosynthesis protein (CpsL) associated with virulence in streptococcal species (figure 4C). Within A2 isolates, we observed a distinct pattern of co-ancestry with group B in region 1 and remote events appearing to originate from A1, including a class C  $\beta$ -lactamase that has been associated with ampicillin resistance in *E faecium*.<sup>24</sup>

Focusing on recent events in region 2, we observed several distinct patterns of transfer from A2 and B into A1. Despite the number of probable events observed

and their association with MGEs, genomic diversity within groups remained relatively low across this region. Evidence of restriction and dissemination of a single haplotype was seen in the *ccpA* gene (encoding the catabolite control protein A), a crucial regulator of carbon metabolism in Gram-positive bacteria.

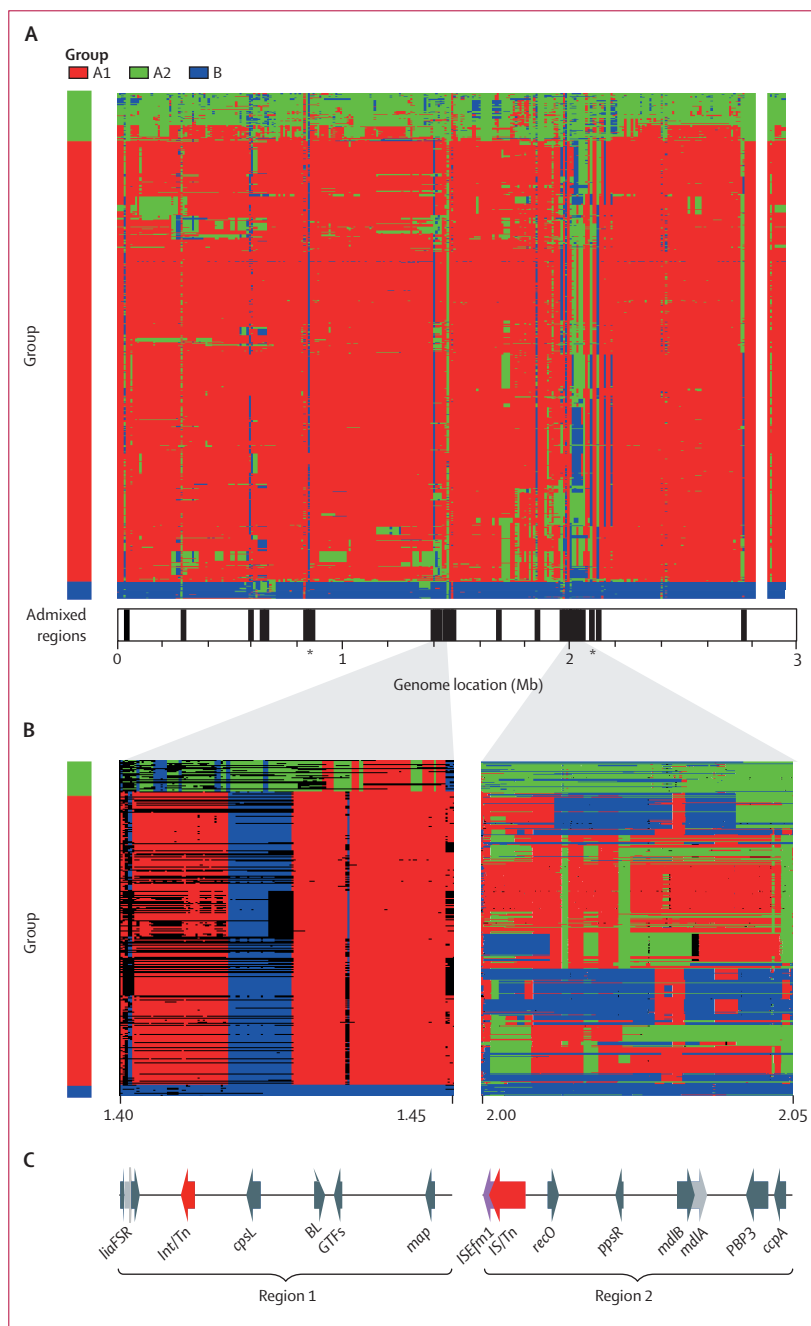
Exploring the other highly admixed regions revealed convergent evolution leading to a single genomic form. Two genes of interest were noted. One, *lgt* (located at 2.06 Mb), encoding a prolipoprotein diacylglycerol transferase, plays a crucial role in membrane anchoring, growth, virulence, and stress responses, especially under oxidative stress.<sup>25</sup> A second gene with a similar pattern of variation, the amidase gene (located at 0.85 Mb) encodes the N-acetylmuramoyl-L-alanine amidase enzyme, involved in biosynthesis of peptidoglycan integral to *E faecium* cell walls.<sup>26</sup>

## Discussion

In this study, we contribute to resolving the uncertainties about the *E faecium* global population structure by assembling the largest-to-date collection of genomes from patients who were hospitalised, using a sampling strategy designed to maximise the representation of both colonising and infection isolates. Benefiting from both the broad-based sample set and an appropriate analysis framework, our data supports the distinctiveness of separate A and B groups.

Additionally, we showed that group A isolates do split into two distinguishable subgroups, approximately equivalent to previous A1 and A2 clades. Some of the previous controversy about the status of A2 is thus probably explained by sampling strategies that might have left the small hospital A2 pool unsampled. After masking of identified recombination events, A2 isolates could be distinguished into subgroups seemingly ancestral to the lower-diversity A1 subgroup. Moreover, our observations suggest that A1 initially emerged as a clone from A2, evidenced by A1's low divergence from A2 compared with B. This emergence probably occurred in the hospital, an environment in which high rates of recombination continue to occur. These events are facilitated by MGEs within and between groups, as evidenced by regions of sequence convergence across *E faecium* populations.

Subgroup relationships were examined in more detail by use of a co-ancestry approach. The advantage of this approach is that identified relationships are not dependent on incomplete assessments of recombination through phylogenetic reconstructions and subsequent exclusion of substantial amounts of genomic data leading to loss of inferences as seen with B genomes. By contrast, chromosome painting provides a coherent view of genetic exchange and showed a continuum of co-ancestry patterns linking groups because of the interplay between the different *E faecium* groups found in hospitals. New genomic forms arise from this interplay, presumably



**Figure 4: Extensive admixture and directionality of recombination in *Enterococcus faecium***

(A) Chromosome painting along the genome for all 1128 isolates (rows) presented by group in order of the recombination-masked tree (figure 1C) and coloured by the inferred group of origin of each segment; legend indicates the Bayesian Analysis of Population Structure group assignment with the same colour key; highly admixed regions (more than 75% of isolates in any one group containing material from another group) are shown by the black lines in the panel underneath by genome position; \*location of the *amidase* (left) and *lgt* (right) genes. (B) Recombination events inferred by fastGEAR of two representative 50 kb segments with the origin and the directionality of the sequences inferred by group colour within the panels; both remote (coloured by group) and recent (depicted by black lines) recombination events are shown within panel corresponding to region 1; only recent (coloured by group) events are depicted in the second panel corresponding to region 2. (C) Annotations of genes of interest are shown across the two regions; only genes with a known function in *E faecium* and those associated with mobile genetic elements are shown; region 1 genes are the *LiaFSR* gene cluster, *Int/Tn*, *cpsL*, *BL*, *GTFs*, and *map*; region 2 genes are *ISEfm1*, *IS/Tn*, *recO*, *ppsR*, *mdlA* and *mdlB*, *PBP3*, and *ccpA*.

within the gastrointestinal tracts of patients who are hospitalised. These forms have the potential to become established as either novel A1 sublineages, with further differentiation through recurrent recombination (as we have showed in other work)<sup>27</sup> or as more distinct new lineages that would parallel the origins of A1.

Together these data emphasise the genomic distinctions between the three groups and the high rates of recombination between resident hospital-adapted A1 isolates and enterococcal A2 or B (ie, community) isolates entering the hospital. The observed direction of exchange in sampled isolates is preferentially from A2 or B to A1, and it probably includes some ascertainment bias (a limitation of our study of hospital isolates). Nonetheless we showed the flux of genetic variation into the hospital, enhancing opportunities to form new lineages by recombination with already-adapted lineages. Therefore, the population structure of A1 continues to evolve by introgression of genes and through the formation of hybrid genomes. Higher rates of introduction and mixing of genomically diverse lineages can only make the formation of new, more highly adapted and potentially difficult to control clones of *E faecium* more likely.

Strictly, our analysis applies only to *E faecium* obtained from humans who were hospitalised. It remains unknown whether our findings reflect *E faecium* interactions between isolates sampled from other settings, such as environmental or animal sources. Nevertheless, it is clear (from other work) that A2-like isolates are seen in animals and the environment, and similar genomes are detected in humans and companion animals, thus the rate of genomic interchange is likely to reflect opportunities for interaction.<sup>7</sup>

The best terminology to describe A1's relationship to the other diverse A subgroups is likely to remain uncertain until further research confirms the current findings, establishes the full set of such groups in different environments and hosts, and perhaps revisits the taxonomic relationship between the A and B groups of *E faecium*.

Most of the recombination in this study appears to affect a minority of the genome, corresponding to highly admixed regions and so-called remote events, in which sequences have been imported from outside our sampled *E faecium* collection. Although the overall effect of these regions on A1 adaptation remains unknown, such segments tend to be within previously documented hotspots for further recombination.<sup>2</sup> One such region, the *LiaFSR* gene cluster associated with stress responses, carries among A1 isolates a limited range of alleles apparently originating through remote recombination events from B genomes and is suggestive of directional selection with subsequent clonal expansion implying possible selective sweep events. Subsequent imports of A2-like sequences appear to have supplied a distinct new haplotype, the only one on which daptomycin resistance has been documented.<sup>27</sup> Therefore, although remote

recombination seems crucial to the genomic diversity of this gene cluster, group A2 clearly represents the more important recent and ongoing source of genetic material for A1.<sup>28</sup> Exploring the other highly admixed regions (carrying the *lgt* and *amidase* genes) revealed patterns of convergent evolution leading to a single genomic form suggestive of inter-lineage recombination in the face of selective pressure.

MGEs play a crucial role in recombination in *E faecium*, with allele replacement events across A1, generally associated with genes involved in environmental adaptation and stress responses.<sup>7</sup> Although the adaptive consequences of such genomic sweeps remain unclear, it is tantalising to speculate that these might be the events that explain the establishment and expansion of A1 in the health-care sector.

Antimicrobial resistance does not feature heavily in our analysis, despite the association of many resistance phenotypes in *E faecium* mediated by MGEs. Possible reasons include the relatively small non-A1 resistance gene pool and the dominance of A1 isolates within the hospital, such that resistance mobilisation most likely represents within-group recombination. This characteristic of resistance dissemination separate from genome adaptation might explain why, as others have observed, antimicrobial resistance remains disconnected from A1 clonal success and expansion.<sup>6,29</sup>

Our study is limited by the small number of genomes originating from several known high burden *E faecium* countries such as the USA. Sparse sampling from Asia and Africa might also have influenced our conclusions.

In conclusion, we provide a detailed analysis based on the patterns of co-ancestry of segments of the *E faecium* genome using chromosome painting. This has allowed us to elucidate key aspects of the population structure and dynamics of hospital-associated *E faecium* on a global scale, resolving long-standing questions about the relationships and interplay between *E faecium* genomes.

#### Contributors

SJvH, RJLW, AMH, SAB, TMC, KH, HTW, BPH, SM-K, GW, KY, AME, and KER provided sequence data. SJvH, RJLW, JC, and RB did the experimental design. SJvH, JC, and RB did the analyses. SJvH and RB prepared the manuscript and verified the data. All authors read and approved the final manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

#### Enterococcal Study Group

Hendrik T Westh (MRSA Knowledge Center, Department of Clinical Microbiology 445, Hvidovre Hospital, Hvidovre, Denmark); Katherine Loens (National Reference Centre for Enterococci, University Hospital Antwerp, Edegem, Belgium); Basil B Xavier (Laboratory of Medical Microbiology, Vaccine & Infectious Disease Institute, Universiteit Antwerpen, Wilrijk, Belgium); Veerle Matheeußen (National Reference Centre for Enterococci, University Hospital Antwerp, Edegem, Belgium); Herman Goossens (National Reference Centre for Enterococci, University Hospital Antwerp, Edegem, and Laboratory of Medical Microbiology, Vaccine & Infectious Disease Institute, Universiteit Antwerpen, Wilrijk, Belgium)

#### Declaration of interests

We declare no competing interests.



**Data sharing**

All new short-read sequence data (n=321) was uploaded to the National Center for Biotechnology Information (NCBI) under project number PRJNA63689. The remaining sequences (n=807) were downloaded from NCBI with project numbers and associated metadata provided in the appendix (pp 3–45).

**Acknowledgments**

We wish to acknowledge those responsible for all aspects of isolate curation, data collection, typing, and sequencing at the different locations: A P Tedim from Spain, P Worning from Denmark, J K Bender from Germany, and A Beukers from Australia. This research was funded in whole, or in part, by the Wellcome Trust (203141/Z/16/Z). For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. RJLW was supported by the Joint Programming Initiative in Antimicrobial Resistance (JPIAMR Third call, STARCS, JPIAMR2016-AC16/00039). TMC was supported by the European Commission (JPIAMR - STARCS project), InGEMICS-C (S2017/BMD-3691) funded by Comunidad de Madrid (Spain), CIBER in Epidemiology and Public Health (CB06/02/0053), and grants AC16/00039 and P118/01942, integrated in the Spanish 2013–16 and 2017–20 R+D+I State Plans and co-funded by Instituto de Salud Carlos III and the European Regional Development Fund. SM-K acknowledges funding as part of the Methusalem-Excellence consortium VAX-IDEA and JPIAMR-STARCS (JPIAMR2016-AC16/00039). AME was supported in part by federal funding from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract HHSN272200900018C and grant U19AI110818 to the Broad Institute. JC was supported by the European Research Commission grant 742158. RB benefited from Victorian State Government Operational Infrastructure Support and Australian Government's Independent Research Institutes Infrastructure Support Scheme (National Health and Medical Research Council) to the Walter and Eliza Hall Institute of Medical Research and core funding to the Wellcome Centre for Human Genetics provided by the Wellcome Trust (award 203141/Z/16/Z).

**References**

- Arias CA, Murray BE. The rise of the Enterococcus: beyond vancomycin resistance. *Nat Rev Microbiol* 2012; **10**: 266–78.
- van Hal SJ, Ip CLC, Ansari MA, et al. Evolutionary dynamics of *Enterococcus faecium* reveals complex genomic relationships between isolates with independent emergence of vancomycin resistance. *Microb Genom* 2016; **2**: e000048.
- Willems RJL, Top J, van Schaik W, et al. Restricted gene flow among hospital subpopulations of *Enterococcus faecium*. *MBio* 2012; **3**: e00151–12.
- Lebreton F, van Schaik W, McGuire AM, et al. Emergence of epidemic multidrug-resistant *Enterococcus faecium* from animal and commensal strains. *MBio* 2013; **4**: e00534–13.
- Gouliouris T, Raven KE, Ludden C, et al. Genomic surveillance of *Enterococcus faecium* reveals limited sharing of strains and resistance genes between livestock and humans in the United Kingdom. *MBio* 2018; **9**: e01780–18.
- 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.
- 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.
- Rios R, Reyes J, Carvajal LP, et al. Genomic epidemiology of vancomycin-resistant *Enterococcus faecium* (VREfm) in Latin America: revisiting the global VRE population structure. *Sci Rep* 2020; **10**: 5636.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; **30**: 2114–20.
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754–60.
- Garrison EMG. Haplotype-based variant detection from short-read sequencing. *arXiv* 2012; published online July 20. <https://arxiv.org/abs/1207.3907> (preprint).
- Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol* 2013; **30**: 1224–28.
- Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* 2015; **11**: e1004041.
- Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014; **30**: 1312–13.
- Feldgarden M, Brover V, Haft DH, et al. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob Agents Chemother* 2019; **63**: e00483–19.
- Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012; **19**: 455–77.
- Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; **30**: 2068–69.
- Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015; **31**: 3691–93.
- Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* 2006; **34**: D32–36.
- Aziz RK, Bartels D, Best AA, et al. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008; **9**: 75.
- Lawson DJ, Henthall G, Myers S, Falush D. Inference of population structure using dense haplotype data. *PLoS Genet* 2012; **8**: e1002453.
- Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Martinen P. Efficient inference of recent and ancestral recombination within bacterial populations. *Mol Biol Evol* 2017; **34**: 1167–82.
- Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J* 2009; **3**: 199–208.
- Miller WR, Munita JM, Arias CA. Mechanisms of antibiotic resistance in enterococci. *Expert Rev Anti Infect Ther* 2014; **12**: 1221–36.
- Reffuveille F, Serror P, Chevalier S, et al. The prolipoprotein diacylglycerol transferase (Lgt) of *Enterococcus faecalis* contributes to virulence. *Microbiology (Reading)* 2012; **158**: 816–25.
- Vermassen A, Leroy S, Talon R, Provot C, Popowska M, Desvaux M. Cell wall hydrolases in bacteria: insight on the diversity of cell wall amidases, glycosidases and peptidases toward peptidoglycan. *Front Microbiol* 2019; **10**: 331.
- van Hal SJ, Willems RJL, Gouliouris T, et al. The global dissemination of hospital clones of *Enterococcus faecium*. *Genome Med* 2021; **13**: 52.
- de Been M, van Schaik W, Cheng L, Corander J, Willems RJ. Recent recombination events in the core genome are associated with adaptive evolution in *Enterococcus faecium*. *Genome Biol Evol* 2013; **5**: 1524–35.
- Raven KE, Reuter S, Reynolds R, et al. A decade of genomic history for healthcare-associated *Enterococcus faecium* in the United Kingdom and Ireland. *Genome Res* 2016; **26**: 1388–96.