Genome-based characterization of hospital-adapted *Enterococcus faecalis* lineages

Kathy E. Raven,[1*] Sandra Reuter,[1] Theodore Gouliouris,[1,2,3] Rosy Reynolds[4,5], Julie E Russell[6], Nicholas M. Brown[2,4], M. Estée Török,[1,2,3] Julian Parkhill,[7] Sharon J. Peacock.[1,3,7,8]

[1] University of Cambridge, Department of Medicine, Box 157 Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, United Kingdom

[2] Public Health England, Clinical Microbiology and Public Health Laboratory, Box 236, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, United Kingdom

[3] Cambridge University Hospitals NHS Foundation Trust, Hills Road, Cambridge CB2 0QQ, United Kingdom

[4] British Society for Antimicrobial Chemotherapy, Griffin House, 53 Regent Place Birmingham B1 3NJ, United Kingdom

[5] North Bristol NHS Trust, Southmead Hospital, Bristol, BS10 5NB, United Kingdom

[6] Culture Collections, Public Health England, Porton Down, Salisbury SP4 0JG, United Kingdom

[7] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

[8] London School of Hygiene and Tropical Medicine, London, WC1E 7HT, United Kingdom

*Corresponding author: ker37@medschl.cam.ac.uk

Vancomycin-resistant *Enterococcus faecalis* (VREfs) is an important hospital-adapted pathogen. We undertook whole genome sequencing of *E. faecalis* associated with bloodstream infection in the United Kingdom and Ireland (UK&I) over more than a decade to determine the population structure and genetic associations with hospital adaptation. Three lineages predominated in the population, two of which (L1 &L2) were nationally distributed and L3 was geographically restricted. Genome comparison with a global collection identified that L1 and L3 were also present in the United States but were genetically distinct. Over 90% of VREfs belonged to L1-3, with resistance acquired and lost multiple times in L1 and L2, but only once followed by clonal expansion in L3. Putative virulence and antibiotic resistance genes were over-represented in L1, L2 and L3 isolates combined, versus the remainder. Each of the 3 main lineages contained a mixture of vancomycin-resistant and –susceptible *E. faecalis* (VSEfs), which has important implications for infection control and antibiotic stewardship.

Enterococci are the second and third most frequent cause of nosocomial infections in the United States (US) and Europe respectively,[1,2] with *Enterococcus faecalis* the most commonly isolated species.[2] Vancomycin is the first-line antimicrobial drug for enterococci with high-level resistance to ampicillin or for patients with penicillin allergy. Vancomycin resistance was first reported in 1988[3] and subsequently increased in prevalence. This rise was predominantly due to *E. faecium*, but *E. faecalis* accounted for 11% of VRE bacteremias in the UK&I between 2001-2013 (http://www.bsacsurv.org). Based on multi-locus sequence typing (MLST)[4], vancomycin resistance in *E. faecalis* has arisen in multiple genetic backgrounds,[5,6] and is associated with epidemic lineages[7] Microbial genome sequencing provides the opportunity to gain a detailed understanding of the molecular basis for hospital adaptation. The first whole-genome sequence of *E. faecalis* was published in 2003.[8] Subsequent genome studies have compared 18 *E. faecalis* strains, which demonstrated the contribution of mobile

47 genetic elements to the diversity of the species;[9] and 25 clinical and 7 non-clinical isolates,
48 which revealed that both were comparable in gene content.[10] The ability to sequence large
49 bacterial collections means that the molecular epidemiology and gene content of epidemic
50 and sporadic lineages can now be systematically defined.

51 We sequenced 168 *E. faecalis* isolates (58 VREfs, 110 VSEfs) from national (British
52 Society for Antimicrobial Chemotherapy (BSAC) n=94), local (Cambridge University Hospitals
53 NHS Foundation Trust (Addenbrooke's Hospital and The Rosie Hospital) (CUH, n=60), and
54 reference (National Collection of Type Cultures (NCTC) n=14) collections (see Supplementary
55 Table 1). BSAC isolates originated from 21 UK&I hospitals between 2001 and 2011 (see
56 Supplementary Fig. 1 for geographical and temporal distribution), and CUH isolates were
57 collected between 2006 and 2012. All BSAC and CUH isolates were associated with
58 bloodstream infection. NCTC isolates were from humans, livestock and food products and
59 were predominantly isolated prior to 1951 (10/14 isolates).

60 Comparison of these genomes against the core genome of *E. faecalis* V583 identified
61 124,194 single nucleotide polymorphisms (SNPs) over 2,886,189 nucleotides (Fig. 1). A
62 striking feature of the phylogenetic tree based on these SNPs was that 53% of isolates
63 clustered into three distinct lineages (termed L1, L2 & L3). L1 was represented in each year of
64 the collection, while L2 was most frequently represented between 2001 and 2006 after which
65 there were only two isolates identified that belonged to this lineage (Supplementary Fig. 1).
66 This suggests clonal replacement, a phenomenon observed for other hospital-related
67 pathogens such as methicillin-resistant *Staphylococcus aureus.*[11] Annotation of lineage-
68 specific trees with geographical location demonstrated that L1 and L2 were nationally
69 distributed (epidemic) clones, whilst L3 was only isolated in two locations (CUH and a hospital
70 in the East Midlands referral network[12]) (Fig. 1). L3 was the dominant lineage at CUH (17/60
71 study isolates) and accounted for 3/13 isolates from a hospital in the East Midlands, the
72 phylogenetic tree supporting a single introduction into this hospital followed by local

diversification. We also explored the phylogenetic origins of VREfs by including all available *E. faecalis* isolates from the NCTC collection. Eleven of 14 NCTC isolates clustered with recent clinical isolates (Fig. 1), including seven isolated prior to 1951, and two VREfs from 1986 (the first year that VREfs was recognised) that belonged to L1 or L2, which in the case of L2 may represent a founder of the circulating vancomycin-resistant *E. faecalis* lineage.

To place our collection into a global context we compared these to the *E. faecalis* genomes of isolates from around the world. This was achieved by retrieving all of the *E. faecalis* genomes (n=353) held by the European Nucleotide Archive (ENA) as of 10/09/2015, and combining our data with 347 of these (excluding 6 based on data quality). The phylogenetic tree based on 1,293 genes conserved in 99% of these 515 isolates revealed that isolates contained within L1 and L3 that originated from the UK and United States were genetically distinct (Fig. 2). This indicates independent clonal expansion of dominant lineages with limited international dissemination. One explanation for this is that these lineages are hospital-associated with limited carriage beyond hospitals. Studies investigating community carriage of VRE in the US and UK have failed to identify VREfs.[13,14] By contrast, global isolates from the non-dominant STs were closely related to UK isolates.

To compare our findings with those of published studies based on MLST, we assigned STs to all 168 study isolates (see Supplementary Table 1). Isolates in L1 were assigned to ST6, ST384 and ST642 (CC2), and L2 isolates were ST28 and ST640 (CC87). Both CCs have been described as high-risk lineages based on their association with hospital-derived isolates in Europe.[5–7] L3 isolates were ST103 (CC388), which has only been reported previously in relation to five clinical and two fecal isolates, all from the Americas.[10]

Comparison of the number of core genome SNPs for L1, L2 and L3 revealed lower genetic diversity for L3 (range 3-60 SNPs, median 33 SNPs), compared with L1 (range 2-375 SNPs median 30 SNPs), and L2 (range 0-237 SNPs, median 139 SNPs). This led us to use Bayesian Evolutionary Analysis Sampling Trees (BEAST) to date these lineages.[15] The last

common ancestor of L3 was estimated to be 1998 (95% highest posterior density (HPD) interval, 1980-2004) (Supplementary Fig. 2a), consistent with the earliest reported isolation of ST103 in the literature of 2002.[10] The last common ancestor of L1 was predicted to be 1918 (95% HPD interval, 1868-1960), with a clonal expansion in 1997 (95% HPD interval, 1992-2000) (Supplementary Fig. 2b). The early estimate for the last common ancestor of L1 relied on just three outlying isolates, and so a second algorithm was used to detect and remove recombination events and the BEAST analysis repeated to rule out the role of undetected recombination. This predicted a last common ancestor in 1852 (95% HPD interval, 1811-1956). It has been proposed previously that CC2 (L1) emerged recently, based on the lack of isolates identified prior to the 1980s.[16] Our analysis indicates that this lineage may have been in existence since the mid 1850s to early 1900s, with a clonal expansion in the 1990s. BEAST analysis of L2 failed, probably because of a limited number of isolates with high genetic diversity and wide temporal spread.

Establishing the rate of mutation in the core genome provides a molecular clock that contextualises analyses of bacterial genomes during putative outbreak investigations,[17–19] but has not been defined previously for *E. faecalis*. The rate of evolution was estimated to be 8.18×10^{-7} SNPs/site/year (approximately 2.5 SNPs/year) for L1 and 1.14×10^{-6} SNPs/site/year (approximately 3.4 SNPs/year) for L3. Based on these mutation rates and patient ward locations, we excluded direct patient-to-patient transmission of the CUH study isolates.

We explored the genetic basis for the success of the dominant *E. faecalis* lineages using a candidate gene approach by comparing the prevalence of putative virulence and antibiotic resistance genes in L1, L2 and L3 isolates combined, versus the remainder. *ace, gelE, asa1, agg, cyl, elrA*, and genes conferring resistance to tetracyclines, aminoglycosides, trimethoprim, chloramphenicol, macrolides/lincosamides/streptogramin B (MLSB), quaternary ammonium compounds (qacs) and vancomycin were over-represented in L1-3 compared to the rest (Fig. 3). There was a striking difference in the prevalence of genes

125 encoding aminoglycoside and vancomycin resistance, two commonly used antibiotics for
126 enterococcal infection, in dominant versus non-dominant lineages. Our findings extend
127 previous reports that epidemic lineages are enriched for multi-drug resistance and specific
128 virulence determinants.[6,7] We then compared the prevalence of the candidate virulence
129 genes in VREfs versus VSEfs contained in L1, L2 and L3 (Supplementary Fig. 3). This showed no
130 significant difference, indicating that over-representation of virulence genes is lineage- rather
131 than VRE-specific.

132 We then analysed the pangenome[20] of the 168 isolates to obtain a more detailed
133 understanding of their entire genomic repertoire. This indicated that *E. faecalis* has an open
134 genome with a gamma value of 0.21 (Supplementary Fig. 4), corroborating results derived
135 previously from the analysis of 5 genomes.[21] The pangenome contained 8,202 genes, of
136 which 1,967 genes were conserved across the collection. Of the 6,235 genes in the accessory
137 genome, 1,687 were present just once. The most common accessory genes encoded
138 hypothetical proteins (n=2,558), IS elements or transposons (n=177), phage or plasmid-
139 associated proteins (n=462 and n=113 respectively), transcriptional regulators (n=225), ABC
140 transporters or cassettes (n=124), and phosphotransferase systems (n=118). A total of 819
141 genes were only found in the three dominant lineages, of which 109 were present in more
142 than 10 isolates (Supplementary Table 2), including a WxL domain surface protein unique to
143 L1. Comparison of the amino acid sequence of the WxL protein from L1 to the proteome of
144 V583 revealed a 100% match to EF_3248, one of 27 WxL proteins identified by Brinster *et*
145 *al.*[22] No genes or homoplasic non-synonymous SNPs were ubiquitous in the dominant
146 lineages and absent from all sporadic lineages, suggesting that there is no single factor that
147 contributed to the emergence of these dominant clones, although antibiotic resistance and
148 virulence determinants are likely to represent multifactorial contributory factors. Analysis of
149 non-synonymous SNPs unique to L3 revealed 122 SNPs in 95 genes (Supplementary Table 3),

but no single genetic event was identified that might explain the geographically constrained success of this lineage.

Recombination is thought to be a major mechanism by which the *E. faecalis* genome evolves, which led us to estimate sites of recombination in the core genome[23] for L1, L2 and L3 (Supplementary Fig. 5). Recombination accounted for 12.3% of the core genome in L1 (6.5% related to a large recombination event in 2 isolates) and 3.9% in L2, with a single predicted 4 bp recombination event in one L3 isolate. This contrasts with reports that recombination across the species is high,[4] which led us to use an alternative algorithm (BratNextGen[24]) to detect recombination. This revealed similarly low levels of recombination in L2 and L3 (6.2% and 0.3% respectively) but higher rates in L1 (37%), although most of this (93%) was contained within two large recombination events (Supplementary Fig. 5). One possible explanation for the low levels of recombination is that this drove the initial diversification of the species but subsequently contributed little to short-term evolution.

Finally, we analysed the genetic basis of vancomycin resistance in the collection. Nearly all VREfs (57/58) carried *vanA*, with a single NCTC isolate carrying *vanB*. Annotation of the tree with resistance to vancomycin showed that all three dominant lineages contained a mixture of VREfs and VSEfs, with 89% of BSAC and 95% of CUH VREfs belonging to L1-3. Based on mapping to a reference *vanA* transposon (Tn*1546*) there was no SNP-based variation between transposons with the exception of one that had a C → T substitution at position 5745. However, there was substantial variation in gene content. The transposase, resolvase, *vanY* and *vanZ* genes were not detected in some isolates, but despite this the minimum inhibitory concentration (available for the 35 *vanA* positive BSAC isolates) was consistently very high (≥256 mg/L). There was considerable variation in genetic content within and between the L1 and L2 transposon, whilst L3 had limited variation with two variants relating to VREfs isolated in 2006-2009 and 2009-2012 respectively, and three partial deletions (Fig. 4). Analysis of the insertion sites for Tn*1546* revealed multiple insertion sites for L1 and L2,

but only one site was identified for L3 in the 11/14 genomes for which this analysis proved possible (Fig. 4 and Supplementary Table 4). Analysis using BLAST revealed that these insertion sites were best matched to plasmids, a finding corroborated using plasmid extraction and *vanA* hybridization for insertion site types 1A, 1B, 2B and 3 (data not shown). These data indicate multiple acquisition and loss of the *vanA* transposon in L1 and L2, suggesting a significant fitness cost. By contrast, the single acquisition followed by clonal expansion in L3 suggests that the transposon has negligible cost or confers a benefit in this lineage. Foucault *et al.*[25] demonstrated that its integration site in the chromosome predominantly determined the fitness cost of the *vanB* transposon. One possible reason for the retention of *vanA* in L3 is that the transposon has inserted into the plasmid at a location that lacks a fitness cost to the bacterium. However, *vanA* is inserted at the same site in 10 isolates from L1 and L2 and there is limited evidence for retention of *vanA* in these isolates.

In conclusion, whole genome sequencing of *E. faecalis* has highlighted the dominance of epidemic lineages in the UK&I, but also showed that a lineage with features of an epidemic lineage was confined to two hospitals. Additionally, we identified that the UK and US have genetically distinct populations belonging to two of these lineages, suggesting a lack of international transmission. The mutation rate defined here will have utility in clinical practice as sequencing technology is introduced into the investigation of putative outbreaks. Genome-level data provided comprehensive insights into the gene content of dominant versus sporadic lineages and allowed us to describe the evolution of vancomycin resistance in this collection, which included multiple loss and acquisition events. The observation that the major VREfs lineages were also the common lineages for VSEfs has important implications for infection control and antibiotic stewardship, since the control of VREfs is likely to depend on defining and addressing drivers for VSEfs and its transmission.

ONLINE METHODS

Ethical approval

The study was approved by the National Research Ethics Service (ref: 12/EE/0439) and the Cambridge University Hospitals NHS Foundation Trust (CUH) Research and Development (R&D) Department.

Isolate collection

The 168 *Enterococcus faecalis* isolates used in this study were selected from three collections: NCTC (n=14, deposited between 1927 and 2007), BSAC (n=94, isolated between 2001 and 2011) and CUH (n=60, isolated between Nov 2006 and Dec 2012). The collection was enriched for vancomycin-resistant isolates by selecting all of the available VREfs from NCTC (n=3) and BSAC (n=35) and the first stored isolate from all cases of VREfs bacteremia at CUH (n=20). To relate this to the underlying VSEfs population, 110 VSEfs were selected as follows: (i) all available VSEfs from the NCTC (n=11), (ii) 59 VSEfs from BSAC (35 matched to the BSAC VREfs cases by hospital and year of isolation where available, and an additional 24 VSEfs to gain greater representation of the VSEfs population); (iii) 40 VSEfs from CUH (the first stored bacteremia-associated isolate matched to CUH VREfs cases by isolation date (n=17), or that occurred 30 days or more after admission (n=19), and 4 additional VSEfs that were available to increase the representation of the local population). BSAC hospitals were assigned to referral networks described previously,[12] which are clusters of hospitals more likely to exchange patients within the cluster than outside of that cluster.

Microbiology and sequencing

Bacterial isolates were cultured on Columbia Blood Agar (Oxoid, Basingstoke, UK) and incubated at 37°C for 48 hours in air. Vancomycin susceptibility was determined using the agar dilution method[26] (BSAC isolates), or the Vitek2 instrument (Biomerieux, Marcy l'Etoile,

France) with the AST-P607 card (CUH and NCTC VREfs isolates). DNA was extracted using the QIAxtractor (QIAgen), according to the manufacturer's instructions. Library preparation was conducted according to the Illumina protocol, and sequencing was performed on an Illumina HiSeq2000 with 100-cycle paired-end runs. Sequence data for all isolates have been submitted to the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena) with the accession numbers shown in Supplementary Table 1.

Phylogenetic analyses

Sequence reads were mapped using SMALT (http://www.sanger.ac.uk/resources/software/smalt/) to the *E. faecalis* reference genome V583 (ENA accession number AE016830) for collection-wide analysis. This reference was selected because it is one of only two finished *E. faecalis* genomes from clinical isolates, and has been used in multiple studies (the second complete genome having only been published in 2014). For analysis of lineages L1, L2 and L3, the oldest isolate from each lineage was selected as a reference for mapping and an assembly created using Velvet. Mobile genetic elements were identified using gene annotation, PHAST[27] (phast.wishartlab.com), WebACT[28] (http://www.webact.org) and BLAST[29] (blast.ncbi.nlm.nih.gov) and were excluded in addition to contigs less than 500bp in length to create a 'core' genome. The core genome sizes were 2,886,189, 2,698,500bp, 2,372,434bp and 2,707,007bp for V583, L1, L2 and L3, respectively. Single nucleotide polymorphisms (SNPs) in the core genome were determined using an in-house script and used to estimate maximum likelihood trees using RAxML[30] with 100 bootstraps. Recombination was removed from the lineage-specific analyses using Gubbins.[23] To place the isolates into a global context, all of the available *E. faecalis* sequences listed in GenBank were downloaded from the ENA (n=353). Six isolates were excluded due to poor assemblies/annotation. The assemblies of the remaining 347 isolates were combined with the assemblies of the study isolates (created using Velvet), annotated with Prokka, and a pan-

genome estimated using Roary.[20] A 90% identity cut-off was used and core genes were defined as those in 99% of isolates. A maximum likelihood tree of the 25,294 SNPs in the 1,416 core genes was created using RAxML and 100 bootstraps. iTOL[31] and FigTree were used to visualise the trees. Assemblies were compared to the MLST database (pubmlst.org/efaecalis/) sited at the University of Oxford[32] using an in-house script.

Population history and mutation rate

Genetic diversity was calculated based on pairwise SNP differences. Bayesian Evolutionary Analysis Sampling Trees (BEAST)[15] was used to date the phylogeny and estimate a mutation rate for L1 and L3 using the core genome after removal of regions of recombination using Gubbins. BratNextGen[24] was used to verify the results of Gubbins for L1 using the following parameters: 10 iterations, 100 permutation runs and a significance threshold of 0.05. One NCTC isolate was excluded from the analysis for L1 because the isolation date was unknown. A Hasegawa, Kishino and Yano (HKY) model and gamma distribution was used, and the best molecular clock and tree selected based on Bayes factors calculated from path sampling and stepping stone sampling[33,34]: for L3 an exponential clock and constant tree were used, for L1 a lognormal clock and Bayesian skyline tree were used, and for the repeat analysis of L1 (with recombination events identified and removed based on BratNextGen) a lognormal clock and constant tree were used.

Detection of candidate genes

Virulence genes were chosen based on evidence from experimental mammalian models[35] and their presence determined by *in silico* PCR using previously published primers: *ace,*[36] *esp,*[37] *gelE,*[38] *asa1,*[39] *agg*[36] and *cyl,*[40] *elrA (OEF2* and *OEF8)*[41], *gls24,*[42] *tpx (ef1933for* and *tpxrev)*[43], *bgsA (bgsA for* and *bgsArev)*[44], *srtA (EF3056F* and *EF3056R)*[45], *sigV (SVRT1-2)*[46], *epaA (AB270_epa_F* and *AB271_epa_R)*[47], *epaB (AB272_epaB_F* and *AB273_epaB_R)*[47], *epaE*

(AB276_epaE_F and *AB277_epaE_R)*[47], *epaN (AB288_epaN_F* and *AN289_epaN_R)*[47] and *perA (perA-FF* and *perA-RR)*[48]. The presence of *msrA* and *msrB* were determined by coverage of EF1681 and EF3164 respectively, when mapped to the V583 reference genome. The presence of *vanA* and *vanB* were established by *in silico* PCR using published primers.[35,49] Genes encoding resistance to additional antimicrobial drugs were detected by comparing the whole genome of each isolate with the ResFinder database (compiled in 2012), [50] which has been manually curated since publication. Sequences were compared using an in-house script, and genes with 100% match to length and > 90% identity match were classified as present. *In silico* PCR using previously published primers was used for genes not in the ResFinder database: *dfrF*[51] and *qacZ*[52]. Statistical significance was determined using Fisher's exact test.

Pangenome and recombination

The pan genome was estimated using Roary.[20] Core genes were defined as those present in all 168 isolates with a 90% ID cut-off. The proteome of *E. faecalis* strain V583 was downloaded from the ENA and interrogated with the WxL protein described in this study using the protein version of BLAST. Recombination was identified within L1-L3 using Gubbins and verified using BratNextGen as described above.

Characterising the Tn*1546* transposon

Sequence reads for each isolate were mapped to Tn*1546* (accession number M97297) from the *E. faecium* strain BM4147 using SMALT. The depth of coverage was between ~30x and ~500x for all isolates, with 53/57 (93%) at a depth of > ~50x. To identify the insertion sites, the sequences adjacent to the start and end of the Tn*1546* transposons were extracted from the assemblies up to a maximum of 10,000bp or the end of the contig. The sequences adjacent to the start of Tn*1546* were too short to analyse but sequences adjacent to the end of Tn*1546* (termed "insertion site sequences") were compared between isolates. Insertion

306 site sequences that were identical for more than 200bp were grouped (groups 1-3 in
307 Supplementary Table 4), and then subgroups defined if there were any changes in the
308 downstream sequence, with no evidence that an insertion or deletion explain this change
309 (changes described in Supplementary Table 4). Where the insertion site sequence available
310 was too short to determine which subgroup it belonged to, this was categorised into
311 subgroup A (the most prevalent subgroup) for simplification of Fig. 4. Each subgroup was
312 identified as plasmid or chromosome-based using BLAST, with transposons considered
313 plasmid-borne if the highest match was to a plasmid and there was no match above 25%
314 coverage to an *E. faecalis* chromosome. Insertion site sequences less than 250bp were not
315 considered long enough for an accurate identification. To verify whether the *vanA*
316 transposons were located on plasmids, plasmid extraction followed by *vanA* hybridization
317 was performed. Representative isolates were selected for each of the transposon insertion
318 sites defined using the sequence data. Plasmids were extracted using the Kado and Liu[53]
319 method except that 100 mg/ml lysozyme was added with the E buffer followed by incubation
320 for 1 hour. Extracts were run on a 0.7% agarose gel and blotted using capillary transfer onto
321 Hybond N+ (Amersham, Buckinghamshire, UK). Hybridization was performed using the DIG-
322 high prime labeling and detection starter kit I (Roche Applied Science, Mannheim, Germany),
323 and luminescence was detected using CSPD (Roche Applied Science). *E. faecium* BM4147 and
324 NCTC 8132 were used as positive and negative controls, respectively, and two isolates with
325 known plasmid sizes were used as size markers (*Yersinia enterolitica* YE212/92 (BT 2, O:9) and
326 YE53/03 (BT 1A, O:5)). Probes were created using the following primers: vanA-1: 5'-
327 GGGAAAACGACAATTGC-3', vanA-2: 5'GTACAATGCGGCCGTTA-3'.[49]
328
329 **Accession codes**
330 The sequence data for the study isolates has been deposited in the ENA under the study
331 accessions PRJEB4344, PRJEB4345 and PRJEB4346, with the accession numbers for individual

isolates listed in Supplementary Table 1. Additional sequences used in this study were the *E. faecalis* reference genome V583 (ENA accession number AE016830) and Tn*1546* from the *E. faecium* strain BM4147 (ENA accession number M97297).

References

1. Sievert, D. M. *et al.* Antimicrobial-resistant pathogens associated with healthcare-associated infections: summary of data reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2009-2010. *Infect. Control Hosp. Epidemiol.* **34,** 1–14 (2013).

2. European Centre for Disease Prevention and Control. Point prevalence survey of healthcare-associated infections and antimicrobial use in the European acute care hospitals. *ECDC* doi 10.2900/86011 (2013).

3. Uttley, A. H. C., Collins, C.H., Naidoo, J. & George, R. C. Vancomycin-resistant enterococci. *Lancet* **2,** 57–58 (1988).

4. Ruiz-Garbajosa, P. *et al.* Multilocus sequence typing scheme for *Enterococcus faecalis* reveals hospital-adapted genetic complexes in a background of high rates of recombination. *J. Clin. Microbiol.* **44,** 2220–8 (2006).

5. Freitas, A. R., Novais, C., Ruiz-Garbajosa, P., Coque, T. M. & Peixe, L. Clonal expansion within clonal complex 2 and spread of vancomycin-resistant plasmids among different genetic lineages of *Enterococcus faecalis* from Portugal. *J. Antimicrob. Chemother.* **63,** 1104–11 (2009).

6. Kuch, A. *et al.* Insight into antimicrobial susceptibility and population structure of contemporary human *Enterococcus faecalis* isolates from Europe. *J. Antimicrob. Chemother.* **67,** 551–8 (2012).

7. Kawalec, M. *et al.* Clonal structure of *Enterococcus faecalis* isolated from Polish hospitals: characterization of epidemic clones. *J. Clin. Microbiol.* **45,** 147–53 (2007).

8. Paulsen, I. T. *et al.* Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science* **299,** 2071–2074 (2003).

9. Palmer, K. L., *et al.* Comparative Genomics of Enterococci: Variation in *Enterococcus faecalis*, Clade Structure in *E. faecium*, and Defining Characteristics of *E. gallinarum* and *E. casseliflavus*. *MBio* **3,** 1–11 (2012).

10. Kim, E. B. & Marco, M. L. Nonclinical and Clinical *Enterococcus faecium* Strains, but Not *Enterococcus faecalis* Strains, Have Distinct Structural and Functional Genomic Features. *Appl. Environ. Microbiol.* **80,** 154–165 (2014).

11. Hsu, L.-Y. *et al.* Evolutionary dynamics of methicillin-resistant *Staphylococcus aureus* within a healthcare system. *Genome Biol.* **16,** 81 (2015).

12. Donker, T., Wallinga, J., Slack, R. & Grundmann, H. Hospital networks and the dispersal of hospital-acquired pathogens by patient transfer. *PLoS One* **7,** e35002 (2012).

13. Coque, T. M., Tomayko, J. F., Ricke, S. C., Okhyusen, P. C. & Murray, B. E. Vancomycin-Resistant Enterococci from Nosocomial, Community, and Animal Sources in the United States. *Antimicrob. Agents Chemother*. **40,** 2605–2609 (1996).

14. Jordens, J. Z., Bates, J. & Griffiths, D. T. Faecal carriage and nosocomial spread of vancomycin-resistant *Enterococcus faecium*. *J. Antimicrob. Chemother.* **34,** 515–528 (1994).

15. Drummond, A. J., Suchard, M. A, Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29,** 1969–73 (2012).

16. Palmer, K. L. *et al.* Enterococcal Genomics. In Gilmore, M. S. *et al.*, editors. *Enterococci: From Commensals to Leading Causes of Drug Resistant Infection* [Internet]. Boston: Massachusetts Eye and Ear Infirmary (2014). Available from: http://www.ncbi.nlm.nih.gov/books/NBK190425/

17. Köser, C. U. *et al.* Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak. *N. Engl. J. Med.* **366,** 2267–2275 (2013).

18. Harris, S. R. *et al.* Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science* **327,** 469–474 (2010).

19. Walker, T. M. *et al.* Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13,** 137–146 (2013).

20. Page, A. J. *et al.* Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* pii: btv421 [Epub ahead of print] (2015).

21. The Human Microbiome Jumpstart Reference Strains Consortium. A Catalog of Reference Genomes from the Human Microbiome. *Science* **328,** 994–999 (2010).

22. Brinster, S., Furlan, S. & Serror, P. C-Terminal WxL Domain Mediates Cell Wall Binding in *Enterococcus faecalis* and Other Gram-Positive Bacteria. *J. Bacteriol*. **189,** 1244–1253 (2007).

23. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43,** e15 (2014).

24. Marttinen, P. *et al.* Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* **40,** e6 (2012).

25. Foucault, M., Depardieu, F., Courvalin, P. & Grillot-Courvalin, C. Inducible expression eliminates the fitness cost of vancomycin resistance in enterococci. *PNAS* **107,** 16964-16969 (2010).

Correspondence and requests for materials should be directed to Kathy Raven.

ACKNOWLEDGEMENTS

We thank the Wellcome Trust Sanger Institute library construction, sequence and core informatics teams, the staff at BSAC and the Cambridge Public Health England Microbiology and Public Health Laboratory, and Hayley Brodrick, Amy Cain, Derek Pickard, Kim Judge and Elizabeth Blane for their technical support. We thank the BSAC for allowing use of isolates from the BSAC Resistance Surveillance Project. This publication presents independent research supported by the Health Innovation Challenge Fund (HICF-T5-342 and WT098600), a parallel funding partnership between the UK Department of Health and Wellcome Trust. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health or the Wellcome Trust. This project was also funded by a grant awarded to the Wellcome Trust Sanger Institute (098051). MET is a Clinical Scientist Fellow supported by the Academy of Medical Sciences, the Health Foundation and the NIHR Cambridge Biomedical Research Centre.

AUTHOR CONTRIBUTIONS

SJP designed the study. KER performed bacterial identification, susceptibility testing and DNA extraction and analysed the data. SR assisted in bioinformatic analysis. TG, RR, JER, NMB and JP contributed materials and data, MET completed ethical approvals, and JP and SJP were responsible for supervision and management of the study.

COMPETING FINANCIAL INTERESTS

The authors declare that they have no competing financial interests.

427 **Figure legends**

428 **Figure 1. Phylogeny of *E. faecalis* isolates drawn from across the United Kingdom and**
429 **Ireland.** Left hand side: Midpoint rooted maximum likelihood tree of 168 *E. faecalis* isolates
430 based on SNPs in the core genome. Colored branches indicate the three dominant lineages
431 (L1, red; L2, purple; L3, turquoise). The vertical bars show the source of each isolate (dark
432 blue, BSAC; light blue, CUH; yellow, NCTC) and presence (red) or absence (blue) of
433 vancomycin resistance determinants. Bootstrap supports over 90% are labelled for the major
434 nodes. Scale bar indicates 10,000 SNPs. Right hand side: Maximum likelihood trees of the
435 three dominant lineages based on SNPs in the core genome after recombination was
436 removed, and rooted on an outlier. The trees are labelled by referral network, with '_1' and
437 '_2' indicating different hospitals within the referral network if more than one contributed to
438 the BSAC study collection, and year of isolation with CUH isolates highlighted blue. Bootstrap
439 supports over 90% are labelled. Scale bars indicate 25 SNPs.
440
441 **Figure 2: Global population structure of *E. faecalis***. Phylogeny of 168 study isolates
442 combined with 347 isolates from geographically diverse locations downloaded from the
443 European Nucleotide Archive (ENA). Maximum likelihood tree based on SNPs in the 1,293
444 genes conserved in 99% of isolates. Colored branches indicate the three dominant lineages L1
445 (red), L2 (purple) and L3 (turquoise). Inner colored ring indicates the country of isolation,
446 outer colored ring indicates the source of the isolate.
447
448 **Figure 3: Prevalence of virulence and antibiotic resistance genes in the dominant lineages**
449 **(L1-3, n=89) and remainder (n=79).** Graphs show the percentage of isolates for which
450 putative virulence genes (a) or antibiotic resistance genes (grouped by antibiotic class) (b)
451 were detected. Genes that were ubiquitous in the collection are not shown. p values are
452 shown when a significant difference was observed using Fisher's exact test. Virulence genes:

453 *ace* = collagen adhesion protein; *agg*= aggregation substance; *asa1*= aggregation substance;
454 *bgsA* = biofilm-associated glycolipid synthesis A; *cyl*= cytolysin; *elrA* = enterococcal leucine-
455 rich protein A; *esp*= enterococcal surface protein; *gelE* = gelatinase ; *perA* = pathogenicity
456 island-encoded regulator; *tpx* = thiol peroxidase. Antibiotic resistance genes: Am =
457 aminoglycosides (comprising one or more of *aac6'-2"*, *aph3"-III*, *aacA*, *ant-6-Ia*, *str*); Chlor =
458 chloramphenicol (*cat*); Linc = lincosamides (*lnuB*); MLSB = macrolide, lincosamide,
459 streptogramin B (comprising *ermB* or *ermT*); Tet= tetracycline (comprising one or more of
460 *tetL, tetM, tetO, tetS*); Trim = trimethoprim (comprising *dfrC*, *dfrD*, *dfrF* or *dfrG*); Qac =
461 quaternary ammonium compounds and other antiseptics (*qacZ*), Vanc = vancomycin.
462
463 **Figure 4. Mapping variation in the vancomycin resistance transposon**. Midpoint rooted
464 maximum likelihood tree of all 168 *E. faecalis* with the 3 dominant lineages highlighted (L1,
465 red; L2, purple; L3, turquoise) and the presence (red) or absence (blue) of a *van* transposon
466 indicated in the vertical bar. The right hand side shows the coverage plot (number of
467 sequence reads that map to that location) of the *vanA* transposon in each isolate, with black
468 indicating presence (30x coverage or above), graduating to white indicating absence (less
469 than 10x coverage). The genes are labelled in the top bar (*tn* = inverted repeat). Colors in the
470 vertical bar on the right indicate the different insertion sites in the three dominant lineages,
471 with a description of each color available in Supplementary Table 4. Scale bar indicates
472 10,000 SNPs.
473