Enterococcus faecium genome dynamics during long-term asymptomatic patient gut colonization

Bayjanov, Jumamurat R; Baan, Jery; Rogers, Malbert R C; Troelstra, Annet; Willems, Rob J L; van Schaik, Willem

DOI:
10.1099/mgen.0.000277

MICROBIAL GENOMICS

RESEARCH ARTICLE

Bayjanov *et al.*, *Microbial Genomics*

DOI 10.1099/mgen.0.000277

Enterococcus faecium genome dynamics during long-term asymptomatic patient gut colonization

Jumamurat R. Bayjanov[1,†], Jery Baan[1,†], Malbert R. C. Rogers[1], Annet Troelstra[1], Rob J. L. Willems[1] and Willem van Schaik[1,2,*]

Abstract

Enterococcus faecium is a gut commensal of humans and animals. In addition, it has recently emerged as an important nosocomial pathogen through the acquisition of genetic elements that confer resistance to antibiotics and virulence. We performed a whole-genome sequencing-based study on 96 multidrug-resistant *E. faecium* strains that asymptomatically colonized five patients with the aim of describing the genome dynamics of this species. The patients were hospitalized on multiple occasions and isolates were collected over periods ranging from 15 months to 6.5 years. Ninety-five of the sequenced isolates belonged to *E. faecium* clade A1, which was previously determined to be responsible for the vast majority of clinical infections. The clade A1 strains clustered into six clonal groups of highly similar isolates, three of which consisted entirely of isolates from a single patient. We also found evidence of concurrent colonization of patients by multiple distinct lineages and transfer of strains between patients during hospitalization. We estimated the evolutionary rate of two clonal groups that each colonized single patients at 12.6 and 25.2 single-nucleotide polymorphisms (SNPs)/genome/year. A detailed analysis of the accessory genome of one of the clonal groups revealed considerable variation due to gene gain and loss events, including the chromosomal acquisition of a 37 kbp prophage and the loss of an element containing carbohydrate metabolism-related genes. We determined the presence and location of 12 different insertion sequence (IS) elements, with IS*Efa5* showing a unique pattern of location in 24 of the 25 isolates, suggesting widespread IS*Efa5* excision and insertion into the genome during gut colonization. Our findings show that the *E. faecium* genome is highly dynamic during asymptomatic colonization of the human gut. We observed considerable genomic flexibility due to frequent horizontal gene transfer and recombination, which can contribute to the generation of genetic diversity within the species and, ultimately, can contribute to its success as a nosocomial pathogen.

DATA SUMMARY

Short-read data for the 96 genomes sequenced in this study are available at the European Nucleotide Archive (ENA), accession number PRJNA344739. The long-read sequence dataset used for the assembly of the genome of strain A_020709_82 is available at ENA, accession number CP018128.

INTRODUCTION

In recent decades, *Enterococcus faecium* has emerged as an important multidrug-resistant nosocomial pathogen. It is a major cause of hospital-acquired infections such as bacteraemia, urinary tract infection and endocarditis [1–4]. Furthermore, enterococcal infections contribute to patient mortality, increased length of hospital stay of patients and

Received 27 February 2019; Accepted 24 May 2019; Published 05 June 2019

Author affiliations: [1]Department of Medical Microbiology, University Medical Center Utrecht, 3584CX Utrecht, The Netherlands; [2]Institute of Microbiology and Infection, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK.

***Correspondence:** Willem van Schaik, w.vanschaik@bham.ac.uk

Keywords: *Enterococcus faecium*; gut colonization; microbial evolution; accessory genome; insertion sequences.

Abbreviations: μ, growth rate; A_{600}, absorbance at 600 nm (A_{600}); ARE, ampicillin-resistant *E. faecium*; BHI, Brain Heart Infusion; BS, Bayesian skyline; ENA, European Nucleotide Archive; HGT, horizontal gene transfer; HPD, highest posterior density; ICU, Intensive Care Unit; IS, insertion sequence; μ_{max}, maximum growth rate; MCC, maximum clade credibility; MCMC, Markov chain Monte Carlo (MCMC); OD_{600}, optical density at 600 nm (OD_{600}); OGs, orthologous genes; PS, path sampling; SNPs, single nucleotide polymorphisms; SS, stepping-stone sampling; VRE, vancomycin-resistant *E. faecium*; WGS, whole genome sequencing.

Short-read data for the 96 genomes sequenced in this study are available at the European Nucleotide Archive (ENA), accession number PRJNA344739. The long-read sequence dataset used for the assembly of the genome of strain A_020709_82 is available at ENA, accession number CP018128.

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. One supplementary table and two supplementary figures are available with the online version of this article.

higher healthcare costs [5]. Infections caused by *E. faecium* are difficult to treat due to the large repertoire of acquired antibiotic resistance determinants, of which vancomycin resistance is arguably the most problematic [6, 7].

The species *E. faecium* consists of distinct subpopulations or 'clades' [8–10]. A deep phylogenetic split distinguishes clades A and B from each other [11], with clade B containing most human commensal isolates. Clade A has been further subdivided in clades A1 and A2 [8]. Clade A1 contains the vast majority of strains isolated from clinical settings, and overlaps with the previously identified *E. faecium* sub-population clonal complex 17 [9, 12]. The polyphyletic clade A2 is enriched for strains that have been isolated from domestic animals and livestock [8, 10, 13]. While vancomycin resistance can be found among strains from both clade A1 and clade A2, clade A1 strains are almost always resistant to ampicillin, while strains from other clades are mostly ampicillin-susceptible [12].

E. faecium is a genetically dynamic organism with an open pan-genome [8, 14, 15]. Genomic changes in *E. faecium* are mostly driven by recombination and horizontal gene transfer (HGT), rather than by mutation [16]. Due to frequent HGT, *E. faecium* strains that have highly similar core genomes can have substantial differences in their accessory genomes [8, 15, 17]. Insertion sequence (IS) elements are abundant in the *E. faecium* genome [17]. IS elements are short transposable segments of DNA that can have an important role in shaping a bacterial genome. Insertion events can lead to the disruption of promoters, coding sequences or operon structures. In addition, they can catalyze genomic rearrangements, including deletions, inversions and duplications in bacterial genomes [18]. Complete genome sequences revealed that dozens of IS elements are scattered around the chromosome and plasmids of clinical *E. faecium* isolates [19, 20]. A number of IS elements, most notably IS*16*, are associated with clade A1 strains and have been hypothesized to contribute to the adaptation of *E. faecium* to the hospital environment [8, 17].

Patients that have been hospitalized for prolonged periods of time are potential reservoirs for drug-resistant *E. faecium* strains. Generally, infection by *E. faecium* is preceded by asymptomatic gut colonization by a resistant clade A1 strain [21, 22]. Patients that have been colonized by *E. faecium* can contaminate both their immediate surroundings and healthcare workers, leading to outbreaks [22, 23]. The ability of *E. faecium* to survive on inanimate objects creates an environmental reservoir of multidrug-resistant strains in hospital wards and makes outbreaks with *E. faecium* a challenge to control [24].

Recent studies have used whole-genome sequencing (WGS) to trace *E. faecium* transmission events between patients in hospital wards and between hospitals [10, 25, 26]. Recently, the relatedness of *E. faecium* strains from bloodstream infections, the gut and the immediate environment of four patients that were hospitalized for up to 2 months was studied using WGS [22]. Here, we present an analysis of the genome dynamics of vancomycin- and ampicillin-resistant *E. faecium* during the asymptomatic gut colonization of five patients for periods ranging from 15 months to 6.5 years. We describe the evolutionary trajectories, including the roles of gene gain and loss events and IS element excision and insertion, that shape the genome of *E. faecium*.

Impact Statement

Enterococcus faecium strains are common members of the gut microbiota of humans and animals. Over the last few decades, a multidrug-resistant clonal group of *E. faecium*, termed clade A1, has emerged to become a common cause of hospital-acquired infections. Gut colonization by clade A1 strains is common in patients. Previous work on *E. faecium* has suggested that recombination and horizontal gene transfer (HGT) are important drivers of diversity within the species. In this study we determined the diversity of *E. faecium* over periods of 15 to 6.5 years in patients that asymptomatically carried multidrug resistant *E. faecium* strains. While some patients harboured diverse *E. faecium* populations over this period of time, the gut of a single patient was colonized by a clonal *E. faecium* population for more than a year. This population was characterized by important gene loss and gene gain events and by the movement of insertion sequences, the simplest transposable elements in bacteria. Both HGT and insertion sequence insertion and excision contribute to the generation of considerable genetic diversity in *E. faecium* strains over relatively short periods of time, which can help this organism to efficiently adapt to novel or rapidly changing environments.

METHODS

Strain collection

Ninety-six *E. faecium* strains were isolated from five patients during routine diagnostic screenings at the University Medical Center Utrecht, a tertiary care facility in Utrecht, the Netherlands, as part of routine screening for colonization by multidrug-resistant *E. faecium* [27, 28]. Patients were screened for carriage of multidrug-resistant *E. faecium* by culturing rectal swabs in Enterococcosel Broth (Becton Dickinson) supplemented with aztreonam (75 mg l^{-1}) at 37 °C. If the cultures exhibited black colorization within 48 h, the broth was streaked on an Enterococcosel Agar Plate (Becton Dickinson) supplemented with aztreonam and vancomycin (25 mg l^{-1}) or with aztreonam and ampicillin (16 mg l^{-1}) and incubated at 37 °C for 48 h. Black colonies formed by Gram-positive cocci were subjected to multiplex PCR to detect vancomycin resistance genes and the *esp* gene, as well as additional antibiotic susceptibility testing. If a vancomycin- and/or an ampicillin-resistant isolate was found during a screening, a single colony was subsequently stored at −80 °C. When morphologically distinct colonies were

visible on a plate, multiple colonies (one per morphotype) were stored at −80 °C. The patients were selected because of their relatively high number of available screening isolates (between 17 and 25 per patient). One patient (patient C) was admitted to the hospital for recurring abscesses on the upper leg, while the other four patients were admitted for (haemo) dialysis procedures. None of the patients were diagnosed with enterococcal infections.

Growth curves and maximum growth rate

A BioScreen C instrument (Oy Growth Curves AB) was used to measure bacterial growth. One colony was picked per strain and grown overnight in brain heart infusion (BHI) broth at 37 °C with shaking at 200 r.p.m. and then diluted to an initial optical density at 600 nm (OD_{600}) of 0.1 in BHI. The cultures were incubated in triplicate in the Bioscreen C system at 37 °C with continuous shaking, and absorbance at 600 nm (A_{600}) was recorded every 15 min for 9 h. The growth rates (μ) were calculated using $\mu = \frac{\ln(A_2) - \ln(A_1)}{(t_2 - t_1)}$, where t_x signifies a time point and A_x signifies the associated A_{600} at this time point. The maximum growth rate (μ_{max}) was determined for each individual experiment by taking the highest μ over the course of the growth.

DNA isolation, genome sequencing and assembly

Genomic DNA of all strains was isolated from overnight cultures in BHI broth, incubated at 37 °C with shaking at 200 r.p.m., using the Wizard Genomic DNA purification kit (Promega). Library preparation for sequencing was performed using the Nextera XT kit and 150 nucleotide paired-end sequencing was performed by Edinburgh Genomics on an Illumina HiSeq 2500. An additional 70 publicly available *E. faecium* genomes, described by Lebreton *et al.* [8], were also included in our analyses and were used to represent the global diversity of the species *E. faecium*. The Nesoni (version 0.122) tool [29] was used to remove adapter sequences and homopolymers, and to trim low-quality bases in sequence reads that had a quality score below 10. If more than half of a read was composed of low-quality bases, the read was discarded. The SPAdes assembler (version 3.1.0) [30] with the *careful* option and k-mer sizes of 21, 31 and 41 was used for genome assembly. From the resulting contigs, those with less than 10-fold nucleotide coverage, as well as those smaller than 500 bases, were discarded.

Assembly quality was checked using QUAST [31] and contigs not originating from bacteria (presumably due to low-level contamination of datasets with eukaryotic reads) were identified by alignment to the National Center for Biotechnology Information (NCBI) GenBank database using BLAST+ (version 2.2.29) [32] and removed.

The genome of strain A_020709_82 (GenBank accession number CP018128) was sequenced to serve as a reference for the analysis of the accessory genomes and the distribution of IS elements in the genomes of the strains in group 1. DNA was prepared as described above and then prepared for sequencing according to the genomic DNA sequencing protocol for the MinION device (Oxford Nanopore Technologies, March 2016). Approximately 60 ng of the obtained pre-sequencing mix was loaded on a R7.3 flow cell and sequenced using an Oxford Nanopore MinION MkI instrument, which was run for a total of 48 h with a pre-sequencing mix top up (~60 ng) at the 24 h mark. A total of 18 629 high-quality two-directional (2D) reads were produced for a total of ~127 million bases. Poretools [33] was used to extract a FASTA-format file containing the reads. A hybrid assembly using these reads combined with 2×150 bp HiSeq 2500 Illumina reads was then generated using SPAdes 3.7.0 [30] using the *nanopore* option in SPAdes.

Genome annotation and clustering of orthologous proteins

We annotated the genome assemblies of all 166 isolates included in this study by using the Prokka [34] annotation tool (version 1.10) with its default parameters. To create clusters of orthologous proteins, the amino acid sequences of all genes in the 166 genomes were aligned against themselves using BLAST+ (version 2.2.29) [32]. Orthologous genes were identified with orthAgogue (version 1.0.3) [35] using the bit score information from the BLAST alignments, where the aligned sequence length between two genes should be at least half the size of the longer gene. Orthologous genes were grouped into orthologous groups (OGs) using the MCL algorithm (version 12-135) with an inflation parameter of 1.5 [36].

Phylogenetic analyses

We generated core genomes by concatenating the sequences of OGs that were present once in all genomes. To prevent bias in our data created by recombination, we filtered the core genomes to identify putative recombination regions using the Gubbins recombination filtering tool (version 1.3.4) [37]. We then used the SNPs in the core genome located outside of the identified recombination regions to create 2 phylogenetic trees: 1 for all 166 strains (96 patient isolates and 70 publicly available genomes) and 1 for the 114 clade A1 strains (95 patient isolates and 19 clade A1 isolates as defined in [8]), using FastTree2 (version 2.1.7; double precision mode enabled) [38]. We used a GTR substitution model for nucleotide sequences with a gamma site evolutionary rate correction and 1000 bootstrap samples to estimate the support for bifurcation points.

We observed 6 different groups of strains among the 96 newly sequenced strains based on their similarity in the tree of 114 clinical isolates. Each of these six groups was then analysed separately. For each group of strains, OG clustering as well as recombination filtering was applied, as described above. However, instead of using SNPs, a concatenated core genome containing all the core genes that were outside of recombination regions was used to obtain more accurate branch lengths and better estimates of time divergence in phylodynamic analysis [39].

Estimation of mutation rates

To further analyse the evolutionary dynamics of each group of strains, we first checked for the presence of sufficient temporal signal (R^2>0.30) using Path-O-Gen (version 1.4pre) [40]. We then used the BEAST molecular evolutionary analysis tool (version 1.8.2) [41] for those groups that had a sufficient level of temporal signal.

We used jModelTest2 [42] to identify the substitution model and site heterogeneity model, and to estimate the proportion of invariant sites, the transition/transversion ratio and the shape parameter of the Γ distribution. Five different clock models (strict, exponential, lognormal, fixed and random) and three different demographic models (constant, lognormal and Bayesian skyline plot) were used in the BEAST analysis. These different models were analysed with 100 million Markov chain Monte Carlo (MCMC) simulations with 10 million burn-ins, where sampling was performed after every 10 000 simulations. The best model among these 15 models (5 clock models×3 demographic models) was selected using path sampling (PS) and stepping-stone sampling (SS) model selection algorithms with 1 million simulations and 100 path steps, where logs after every 1000 simulations were screened as described previously [43]. A maximum clade credibility (MCC) tree was generated using TreeAnnotator using the median heights of the trees [41]. The estimated prior values for substitution and heterogeneity models as determined by jModelTest2 were HKY and I+G for both groups. The rest of the estimated coefficients were the same, with the exception of the transition/transversion ratio being 6.79 and 3.88 for groups 1 and 4, respectively. The best BEAST model for group 1 isolates was a lognormal relaxed clock (lognormal) with a constant coalescence model (lognormal-constant) based on SS model selection, and a lognormal relaxed clock with a Bayesian skyline (BS) coalescence model (lognormal-BS) based on the PS model selection. Although the SS model selection method is generally more accurate than the PS model selection method [44], we chose the lognormal-BS, as the BS coalescence model had higher effective sample size values than the constant model; and the difference between lognormal-constant and lognormal-BS models was negligible. For group 4 strains, the best BEAST model was the lognormal relaxed clock with an exponential coalescence model according to both the SS and PS model selection methods.

Analysis of accessory genome

In addition to core genome-based analysis, we studied the differential presence of accessory genes within the six groups. When an OG was either present in less than or absent in more than 90 % of the strains in the group, it was included in the accessory genome. In addition to the annotation information, we also considered which contigs differentially present genes were located on in order to identify potential genetic links. Thus, we aligned the corresponding contigs of each differentially present gene against the GenBank database using BLAST+ (version 2.2.29+) [32] to identify putative mobile elements on which the variably present genes were located. We aligned differentially present genes of group 1 to the A_020709_82 reference genome and visualized the location of these genes on the reference using Circos (version 0.69) [45]. The 2 largest clusters were aligned to all 166 genomes using BLAST+ (version 2.2.29+) [32] to determine the presence/absence of these regions in *E. faecium*. Abricate [46] was used to determine the presence of antimicrobial resistance determinants in the assembled genomes.

Gain and loss of insertion sequences

We used ISMapper [47] to find the gain and loss of IS elements among group 1 strains, which is entirely composed of isolates from the same patient (patient A). The sequences of the IS elements were found by uploading the complete genome sequence of the A_020709_82 reference strain at the ISfinder website [48]. The sequences of the identified IS elements were used in ISmapper, together with the sequence reads from the patient isolates. The genome of the patient A strain A_020709_82 was used as a reference to which reads were aligned, and the positions of the IS elements were ordered with respect to their positions in the reference genome.

RESULTS

Isolate collection and patient hospital stay

This study used vancomycin- and ampicillin-resistant *E. faecium* (VRE and ARE, respectively) isolates that were collected and stored in the period 2001–2008 as part of routine diagnostics and infection prevention interventions at the University Medical Center Utrecht, the Netherlands (Fig. 1). Analysis of the collected isolates with anonymized patient data showed that for five patients multidrug-resistant *E. faecium* isolates were collected over a period of >1 year. We sequenced the genomes of 96 isolates, all of which were previously determined to be ampicillin-resistant [27]. Using Abricate [46], we found that 38 and 21 isolates carried the *vanA* or *vanB* operon, respectively. Further information on the antibiotic resistance profiles of the strains sequenced in this study is provided in Table S1 (available in the online version of this article). The time span between the collection of the first and the last isolate from a single patient ranged from 15 months (patient B) to 6.5 years (patient C).

Genetic diversity of *E. faecium* patient isolates

To be able to place the collected isolates in the larger *E. faecium* population, we created an SNP-based, recombination-filtered phylogenetic tree using the 96 genomes sequenced in this study and 70 previously described *E. faecium* genome sequences that represent the global *E. faecium* population [8]. This phylogenetic tree was based on 1448 core genes and a total of 77 909 SNPs. Out of the 96 patient isolates, 95 clustered into clade A1, a clade of hospital-associated *E. faecium* strains (Fig. S1). The remaining isolate (C_050730_48) clustered with strains that were previously assigned to clade A2.

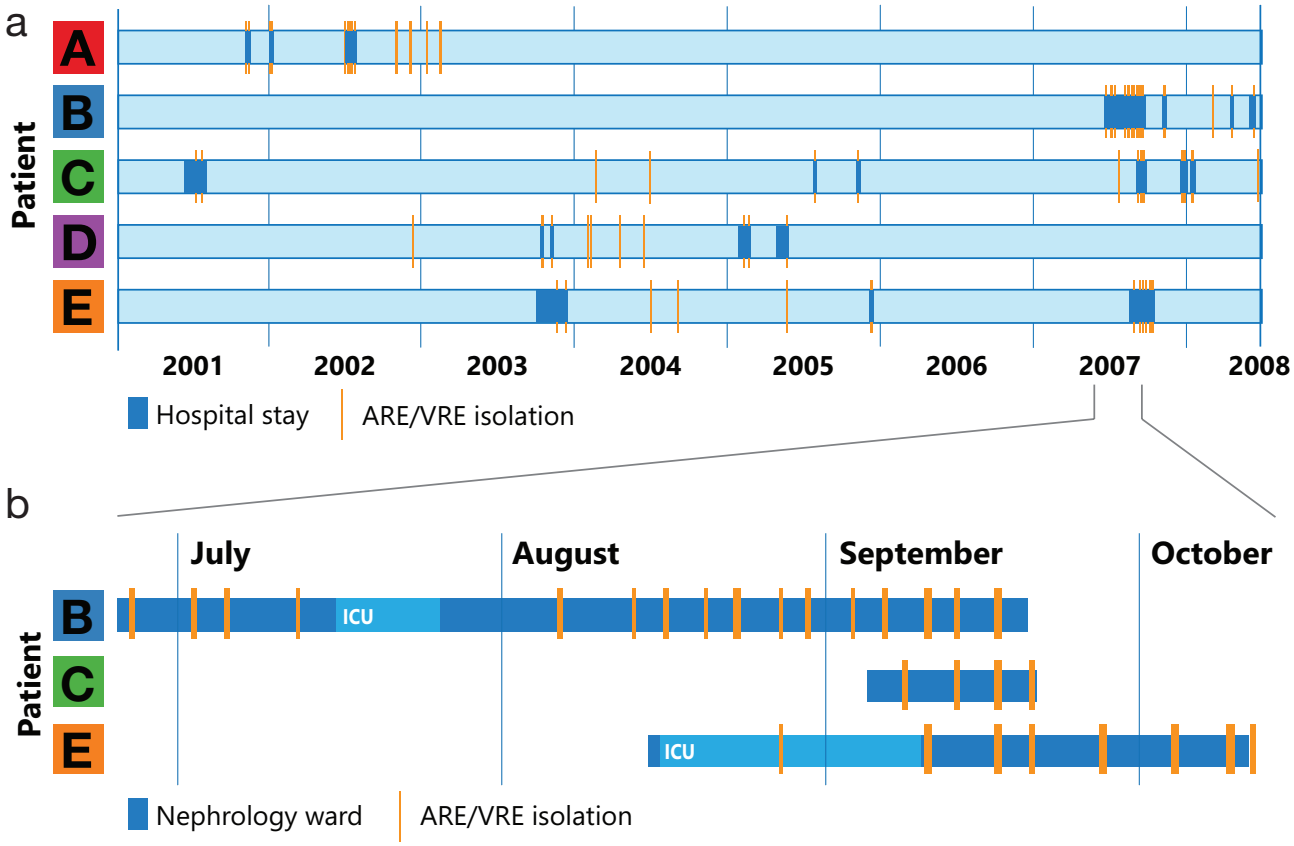

Fig. 1. (a) Timeline of hospital stay for five patients (A–E) and the time points at which multidrug-resistant *E. faecium* strains were isolated during routine screening between 2001 and 2008. (b) Detail for patients B, C and E, showing the overlap in their hospital stay on the nephrology ward in 2007 and the associated ARE/VRE isolations. Dark blue, patient hospital stay; orange, ARE/VRE-positive screening; ICU, patient in an intensive care unit. If an isolation time point does not overlap with hospital stay, the screening was performed at home as part of outbreak control studies.

The relatively large diversity of the 70 publicly available *E. faecium* genome sequences reduced the number of conserved genes in our dataset with 166 strains (Fig. S1), and thus limited the resolution of the phylogenetic tree within clade A1, where the vast majority of the patient isolates in this study were assigned. We therefore created a second tree in which we combined 19 genomes of previously sequenced clade A1 strains [8] with the genomes of 95 clade A1 patient isolates generated in this study. By only analysing the clade A1 *E. faecium* genomes, we were able to construct a tree based on 1805 core genes with 5092 SNPs, allowing us to accurately interpret the similarities between the hospital isolates (Fig. 2). The phylogenetic tree of the clade A1 strains revealed six groups of closely related isolates. Three of these groups (1, 2 and 3) only contained isolates from a single patient (A, B and D, respectively). While additional isolates from patients B and D were present in other groups, the patient A isolates clustered exclusively in group 1. In group 4, the isolates clustered closely together despite being from three different patients (B, C and E). By analysing the hospitalization dates and locations of these patients, we found that patients B, C and E were simultaneously present in a hospital ward (Fig. 1b), suggesting that we captured a small outbreak with this set of isolates. Groups 5 and 6 are two small, highly similar clusters of isolates. The isolates in group 5 originated from three patients (C, D, E), and were isolated between October 2003 and February 2004. In group 6, the isolates from patients C and E that were isolated between September 2007 and January 2008 cluster together. Both of these groups may thus also reflect the transmission of strains between different patients, but we were unable to retrospectively assign epidemiological links that could indicate the direct transmission of strains.

We also found that patients can be colonized by different populations of *E. faecium* at the same time, as shown by the genetic diversity found in strains that were isolated on the same date, e.g. those on 4 February 2004 (D_040204_23, D_040204_24 and D_040204_25).

We determined whether a temporal signal is present in the *E. faecium* genome sequences in each individual group. The temporal signal was defined using Path-O-Gen [40], which plots the time at which each isolate was identified versus the distance to the root of the tree. A temporal signal (R^2>0.3), was only found in groups 1 and 4. Analysis by BEAST resulted in

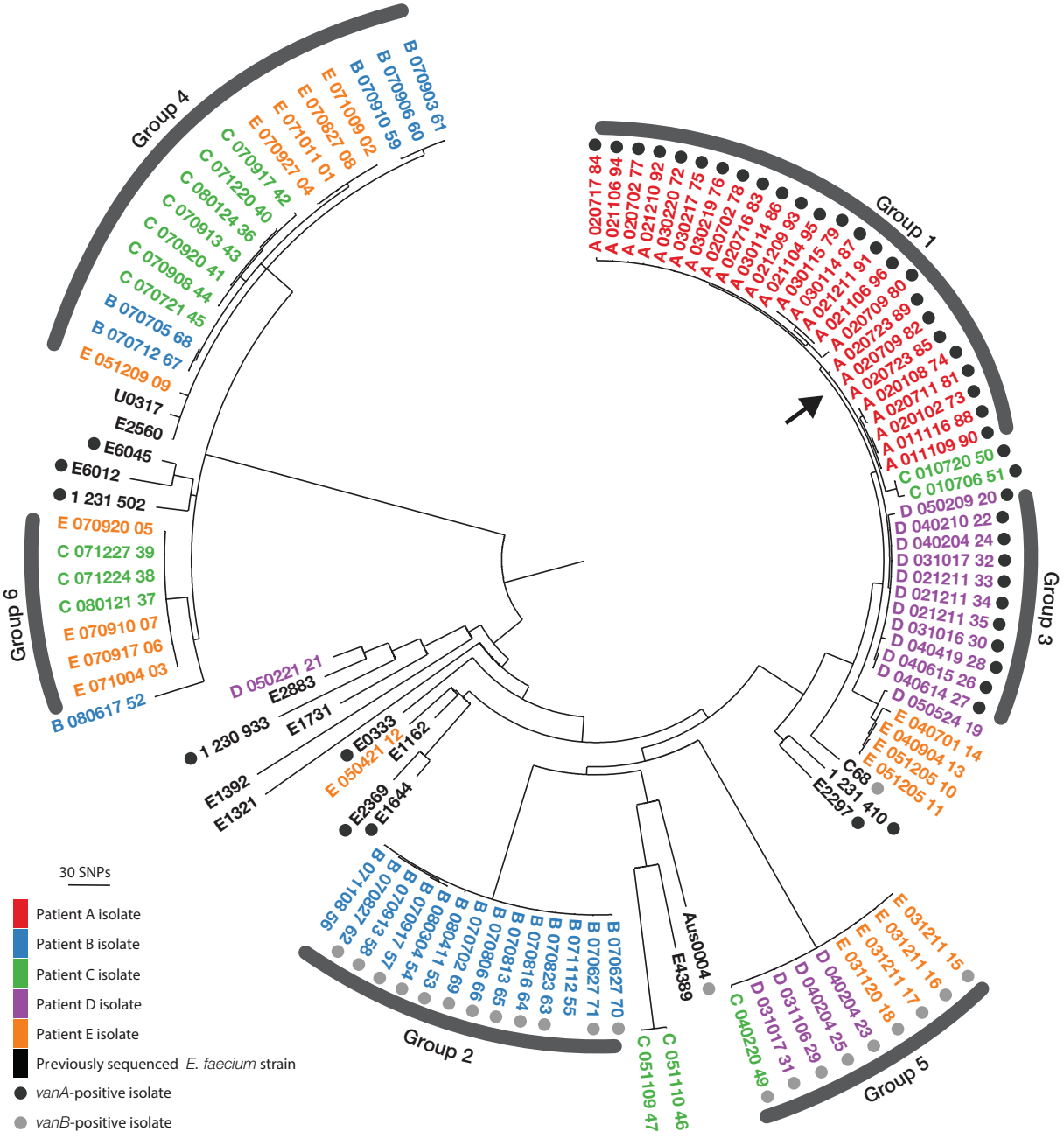

Fig. 2. Phylogenetic tree of clade A1 isolates. This maximum-likelihood tree includes 95 of the 96 genome sequences generated in this study and 19 publicly available *E. faecium* genome sequences. The core genome alignment consisted of 1 637 117 nucleotides. The position of strain A_020709_82 is marked with an arrow. The genome of this isolate was sequenced and assembled to completion using a combination of short- and long-read sequencing for use as a reference genome in further analyses. The genome sequences are coded as follows: the letter represents the patient, the six-number code represents the date of isolation in a year–month–day format and the final number is the unique identifier for each genome sequence. The colours indicate the patient the isolate was taken from. Black and grey marks indicate the presence of the *vanA* or *vanB* vancomycin resistance operon in the genome, respectively.

estimated mutation rates for groups 1 and 4 of 4.2×10^{-6} [with 95 % highest posterior density (HPD) of (2.2×10^{-6}, 6.3×10^{-6})] and 8.4×10^{-6} [with 95 % HPD of (4.7×10^{-6}, 1.2×10^{-5})] substitutions per nucleotide per year, with this being equivalent to 12.6 (6.6–19.8) SNPs/genome/year and 25.2 (14.1–36.0) SNPs/genome/year, respectively. The lack of temporal signal in groups 2, 3, 5 and 6 is likely caused by the relatively low number of strains in these groups. Because all group 1 isolates originate from the same patient, they provide a unique opportunity to study the genome dynamics of *E. faecium* during long-term asymptomatic patient gut colonization. Hence, we focused further analyses on the strains from this group.

The accessory genome of group 1 strains

A total of 74 OGs were found to be differentially present in the genomes of the group 1 isolates (Fig. 3a). Hierarchical clustering showed that most of the OGs were part of larger groups of OGs that showed the same presence/absence pattern across the genome sequences, suggesting that they are genetically linked. Further analysis revealed that the clustered OGs were co-located on contigs.

The two largest variably present clusters are phage-related OGs (cluster 1) and OGs related to carbohydrate metabolism (cluster 2). Cluster 1 contains 24 OGs, of which 10 are annotated as being hypothetical proteins. The annotations of the remaining genes suggest a phage origin of this element, as they include tail and terminase protein-encoding genes (Fig. 3a). Cluster 2 comprises several genes that are related to carbohydrate transport. Neither of the clusters contains genes related to antimicrobial resistance.

When aligning these gene clusters to the original collection of 166 genomes (96 genomes sequenced in this study and 70 genomes representing global *E. faecium* diversity) using BLAST, we find that they are mostly found in the newly sequenced isolates (Fig. S1), with cluster 1 being found in 42 genomes, of which 38 were sequenced as part of this study. Cluster 2 is present in 28 genomes, of which 26 were sequenced here.

To further investigate the genetic linkage of these variably present clusters in the accessory genome of group 1 strains, we fully sequenced the genome of isolate A_020709_82, combining Illumina reads with long reads generated via Oxford Nanopore's MinION platform to complete the genome assembly. The A_020709_82 strain has most of the genes of the accessory genome that are variably present among group 1 strains, including the two largest groups of OGs. The A_020709_82 strain has a chromosome of 2 740 566 nucleotides and 4 plasmids, ranging in size from 222 kbp to 4 kbp (Fig. 3b). By mapping all the differentially present OGs onto the A_020709_82 reference genome sequence, we found that the clustered OGs were located in close proximity to one another in the chromosome. A third, smaller, variably present cluster consisting of four OGs was found to be representing a 4.1 kbp plasmid that was lost in its entirety in four of the isolates, in a presence/absence pattern unrelated to that of the two larger clusters. To assess whether the differences in accessory genome sequences influence the fitness of the 25 group 1 isolates, we determined their *in vitro* maximum growth rates but found no statistically significant differences between the strains with different accessory genomes (Fig. S2).

Dynamics of IS elements in a clonal *E. faecium* population

We identified 12 different IS elements in the genome of A_020709_82. To identify the diversity and location of IS elements in the other strains from group 1, we used ISmapper [47] with the A_020709_82 genome as a reference and the sequencing reads of the other genomes in group 1 (Fig. 4). The positions of two IS elements (IS*16* and IS*6770*) are fixed in all 25 isolates. The IS element IS*Efa5* exhibited a particularly large diversity, with between 17 and 27 copies per genome. Twenty-four out of the 25 isolates in group 1 have a unique pattern, suggesting frequent excision and integration events of this IS element. The remaining nine IS elements showed an intermediate amount of diversity.

DISCUSSION

In this study, we use a collection of *E. faecium* carriage strains that were isolated from patients that had been admitted to hospital repeatedly over a time period ranging from 15 months to 6.5 years. Out of 96 isolates, 95 clustered to the hospital-associated A1 clade, which is expected, given their source, as ampicillin-resistant clade A1 strains cause the majority of hospital-acquired infections and are rarely carried by humans in community settings [8, 49]. The patients likely acquired these isolates during their hospital stay and were carriers for extended periods of time. Previous work has shown that ampicillin-resistant *E. faecium* clones can persist in the gut microbiota for several months after discharge from hospital [50], during which time further spread can occur.

The mutation rates we found in both the clonal group 1 and the non-clonal group 4 (12.6 and 25.2 substitutions/genome/year, respectively) are in line with previously described values for similar *E. faecium* populations [14, 51]. Others have described rates that were up to one order of magnitude higher [8, 22, 25]. This difference is postulated to be caused by increased genetic drift within patients, along with a limited time for purifying selection to act on a population, leading to the incomplete removal of strains with mildly deleterious mutations [22, 52]. Our estimate for the group 1 isolates in particular can be assumed to be a better approximation of the background mutation rate of *E. faecium*, given their clonality, the absence of enterococcal disease in the source patient and the longer time over which they were collected. However, it is also possible that the large differences in the mutation rates of different *E. faecium* clones reported in literature are a true biological signal. As in the Gram-negative gut commensal *Escherichia coli*, *E. faecium* clones with a higher mutation rate may be able to adapt more rapidly to novel environments, while this impacts negatively on their transmissibility and ability to recolonize similar hosts [53].

The pan-genome of *E. faecium* has previously been determined to be essentially open, meaning that it can easily acquire novel genes by HGT [8, 15]. This ability to acquire DNA was recently vividly illustrated by the description of a bovine *E. faecium* strain that had acquired a gene cluster encoding a botulinum-like neurotoxin [54]. In group 1 strains we observed a number of gene gain and loss events. The earliest isolates in group 1 carry a gene cluster that is predicted to be involved in carbohydrate metabolism, while later strains lose this element and acquire a phage. Five isolates carry both the phage element and the carbohydrate metabolism gene cluster, which shows that carriage of both elements is not mutually exclusive. We did not observe differences in the *in vitro* growth rates in rich medium of strains with different combinations of the carbohydrate metabolism element and

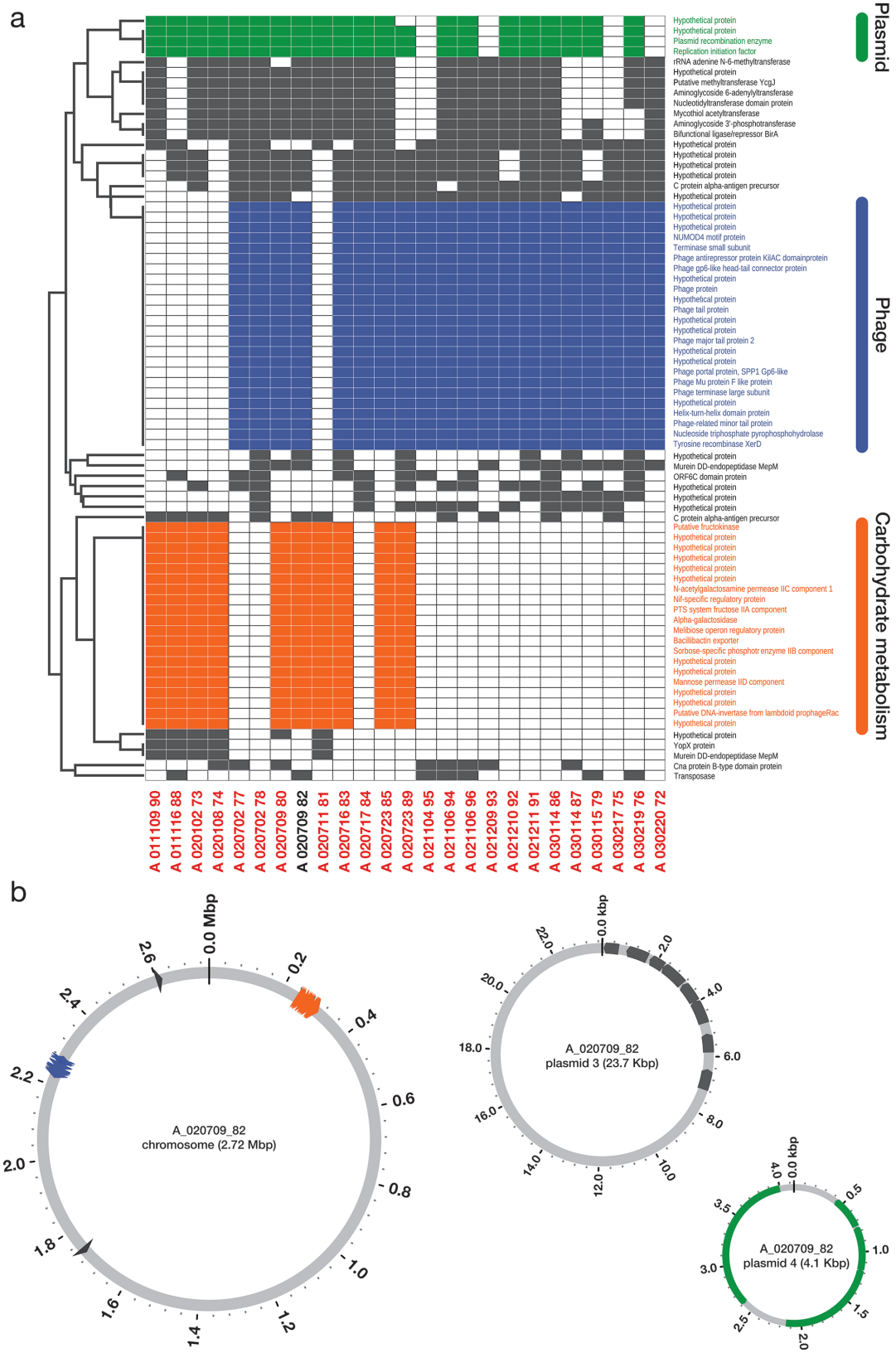

Fig. 3. The accessory genome of group 1 isolates. (a) Plot showing genes that were differentially present in the different isolates, ordered chronologically. The colours indicate gene clusters that were variably present or absent and are annotated on the basis of their predicted function or origin. (b) The differentially present genes mapped onto the A_020709_82 genome, with colours corresponding to gene clusters in (a). Chromosome and plasmid sizes are not shown to scale.

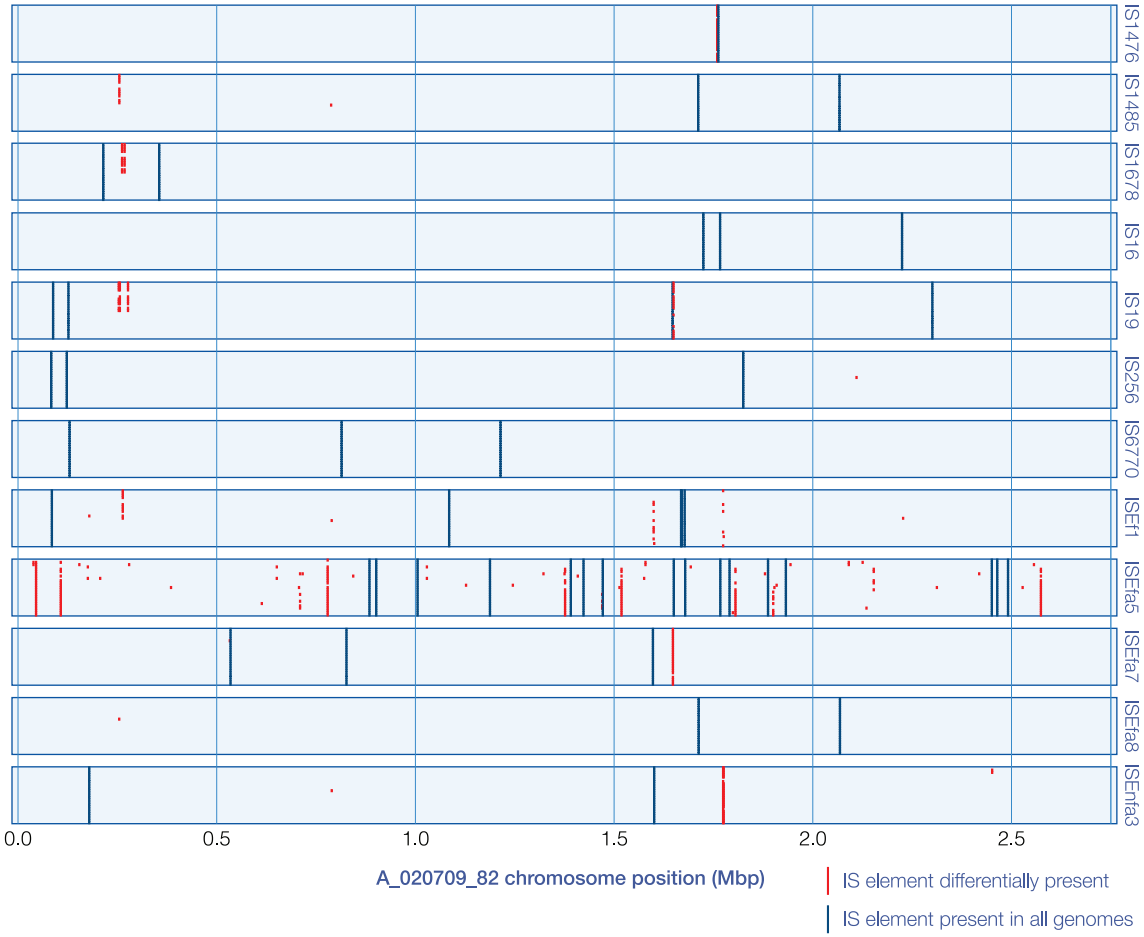

Fig. 4. Variable presence of IS elements in a clonal population of *E. faecium* during asymptomatic gut colonization. Overview of the different IS elements found in the genomes of the 25 clonal patient A isolates, plotted on the chromosomal sequence of isolate A_020709_82. A total of 12 different IS elements were found in this group. Blue marks indicate the presence of that IS element in all isolates. Red marks indicate that the IS element was present in the indicated isolate, but not in all isolates. Each row of an individual IS element represents a single isolate, with the oldest isolate on top and the newest at the bottom.

the phage element. However, these experiments are unlikely to reflect *in vivo* conditions, and the presence or absence of these elements might affect the strains' fitness in colonizing the gut of this patient. It is possible that the changes in the accessory genome allow the clone to adapt optimally for colonization in the context of the patient's gut microbiota.

As described in previous studies, there is an abundance of IS elements in the *E. faecium* genome [17]. We find that some IS elements, such as IS*256*, IS*6770* and the clade A1-associated IS*16* [8], show little to no variation in insertion location and number in the genomes of patient A isolates. IS*16* was previously proposed to confer a degree of genomic flexibility to the hospital-adapted sub-population of *E. faecium* that could contribute to its success as a nosocomial pathogen [17]. However, the fixed position of IS*16* in the group A isolates appears to contradict a major role for this IS element in shaping the *E. faecium* genome. Conversely, we find a large number of IS*Efa5* copies in the genome of group 1 strains and evidence for frequent excision and insertion events. IS*Efa5* was first described as part of Tn*1546*-like elements, which are responsible for VanA-type vancomycin resistance, in South American *E. faecium* isolates [55, 56], but it was later also found in European [57] and Australian strains [58]. In the whole-genome sequence of strain Aus0085, 25 copies of IS*Efa5* were found [58]. Its high copy number in *E. faecium* strains and the evidence for frequent integration and excision events provided in this study, suggest that IS*Efa5* may be contributing significantly to the genomic flexibility of the species. Further work will need to be performed to elucidate the functional impact of IS element excision and insertion in clonal populations of *E. faecium*.

Our observation that patients can be colonized by multiple strains simultaneously is in line with the findings of previous studies [22, 26, 52, 59]. Concurrent colonization by multiple clones can have an important impact on infection prevention efforts if only single colonies are selected for further typing. Potentially pathogenic or multidrug-resistant strains can then be inadvertently missed, leading to the erroneous

reconstruction of transmission networks. When isolates are missed, transmission networks may also be reconstructed erroneously [22], making outbreak control more challenging. This is illustrated by the small outbreak we detected in our dataset, where patients B, C and E appear to have been colonized by isolates with high inter-patient similarity, as well as more different ones. Sampling and typing of multiple colonies when performing screening for colonization by multidrug-resistant *E. faecium* is thus required to capture the full within-patient diversity of this organism. The use of metagenomic shotgun sequencing, combined with tools to reconstruct microbial genomes and resolve strains [60], may become a useful alternative to culture-based approaches to determine the presence of different *E. faecium* clones in the gut microbiome [61].

Our findings show that the *E. faecium* genome is highly dynamic during asymptomatic colonization of the patient gut. We demonstrate *E. faecium*'s remarkable genomic flexibility, which is characterized by frequent gene gain and gene loss due to HGT and recombination and the movement of IS elements. The ability of *E. faecium* to diversify rapidly may contribute to its success as a nosocomial pathogen, as it allows clones that circulate in a hospital to rapidly optimize their ability to effectively colonize individual patients that may differ in their underlying illnesses, antibiotic therapy and gut microbiota composition. Improving our understanding of the mechanisms that underpin this trait is crucial for combating issues related to the emergence of multidrug-resistant *E. faecium* as an important nosocomial pathogen.

Funding information
This work was supported by a grant from the Netherlands Organization for Scientific Research (VIDI: 917.13.357) and a Royal Society Wolfson Research Merit Award to W. v. S. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
The authors wish to thank Mick Watson (Edinburgh Genomics, University of Edinburgh) for generating the Illumina sequence data used in this study.

Author contributions
W. v. S. designed the study. A. T. provided the strains. J. R. B., J. B. and M. R. C. R. performed data analyses. J. R. B., J. B., R. J. L. W. and W. v. S. wrote the manuscript. All authors read and approved the final manuscript.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
Strains were isolated as part of routine diagnostic procedures during a VRE outbreak. This aspect of the study did not require consent or ethical approval by an institutional review board.

References

1. Gao W, Howden BP, Stinear TP. Evolution of virulence in *Enterococcus faecium*, a hospital-adapted opportunistic pathogen. *Curr Opin Microbiol* 2018;41:76–82.
2. Arias CA, Murray BE. The rise of the *Enterococcus*: beyond vancomycin resistance. *Nat Rev Microbiol* 2012;10:266–278.
3. Mendes RE, Castanheira M, Farrell DJ, Flamm RK, Sader HS *et al*. Longitudinal (2001–14) analysis of enterococci and VRE causing invasive infections in European and US hospitals, including a contemporary (2010–13) analysis of oritavancin in vitro potency. *J Antimicrob Chemother* 2001–14;2016:3453–3458.
4. de Kraker MEA, Jarlier V, Monen JCM, Heuer OE, van de Sande N *et al*. The changing epidemiology of bacteraemias in Europe: trends from the European antimicrobial resistance surveillance system. *Clin Microbiol Infect* 2013;19:860–868.
5. Cheah ALY, Spelman T, Liew D, Peel T, Howden BP *et al*. Enterococcal bacteraemia: factors influencing mortality, length of stay and costs of hospitalization. *Clin Microbiol Infect* 2013;19:E181–E189.
6. Miller WR, Murray BE, Rice LB, Arias CA. Vancomycin-resistant enterococci: therapeutic challenges in the 21st century. *Infect Dis Clin* 2016;30:415–439.
7. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M *et al*. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis* 2018;18:318–327.
8. Lebreton F, van Schaik W, Manson McGuire A, Godfrey P, Griggs A *et al*. Emergence of epidemic multidrug-resistant *Enterococcus faecium* from animal and commensal strains. *MBio* 2013;4:00534–13.
9. Guzman Prieto AM, van Schaik W, Rogers MRC, Coque TM, Baquero F *et al*. Global emergence and dissemination of enterococci as nosocomial pathogens: Attack of the clones? *Front Microbiol* 2016;7:788.
10. Raven KE, Reuter S, Reynolds R, Brodrick HJ, Russell JE *et al*. A decade of genomic history for healthcare-associated *Enterococcus faecium* in the United Kingdom and Ireland. *Genome Res* 2016;26:1388–1396.
11. Palmer KL, Godfrey P, Griggs A, Kos VN, Zucker J *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* 2012;3:00318–00311.
12. Willems RJL, Top J, van Santen M, Robinson DA, Coque TM *et al*. Global spread of vancomycin-resistant *Enterococcus faecium* from distinct nosocomial genetic complex. *Emerg Infect Dis* 2005;11:821–828.
13. Gouliouris T, Raven KE, Ludden C, Blane B, Corander J *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;9:01780–18.
14. Howden BP, Holt KE, Lam MMC, Seemann T, Ballard S *et al*. Genomic insights to control the emergence of vancomycin-resistant enterococci. *MBio* 2013;4:00412–00413.
15. van Schaik W, Top J, Riley DR, Boekhorst J, Vrijenhoek JEP *et al*. Pyrosequencing-based comparative genome analysis of the nosocomial pathogen *Enterococcus faecium* and identification of a large transferable pathogenicity island. *BMC Genomics* 2010;11:239.
16. 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;5:1524–1535.
17. Leavis HL, Willems RJL, van Wamel WJB, Schuren FH, Caspers MPM *et al*. Insertion sequence-driven diversification creates a globally dispersed emerging multiresistant subspecies of *E. faecium*. *PLoS Pathog* 2007;3:e7–96.
18. Ooka T, Ogura Y, Asadulghani M, Ohnishi M, Nakayama K *et al*. Inference of the impact of insertion sequence (IS) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in *Escherichia coli* O157 genomes. *Genome Res* 2009;19:1809–1816.
19. Qin X, Galloway-Peña JR, Sillanpaa J, Roh JH, Nallapareddy SR *et al*. Complete genome sequence of *Enterococcus faecium* strain TX16 and comparative genomic analysis of *Enterococcus faecium* genomes. *BMC Microbiol* 2012;12:135.
20. Lam MMC, Seemann T, Bulach DM, Gladman SL, Chen H *et al*. Comparative analysis of the first complete *Enterococcus faecium* genome. *J Bacteriol* 2012;194:2334–2341.

21. Ubeda C, Taur Y, Jenq RR, Equinda MJ, Son T *et al*. Vancomycin-resistant *Enterococcus* domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans. *J Clin Invest* 2010;120:4332–4341.
22. Moradigaravand D, Gouliouris T, Blane B, Naydenova P, Ludden C *et al*. Within-host evolution of *Enterococcus faecium* during longitudinal carriage and transition to bloodstream infection in immunocompromised patients. *Genome Med* 2017;9:119.
23. de Regt MJA, van der Wagen LE, Top J, Blok HEM, Hopmans TEM *et al*. High acquisition and environmental contamination rates of CC17 ampicillin-resistant *Enterococcus faecium* in a Dutch hospital. *J Antimicrob Chemother* 2008;62:1401–1406.
24. Wendt C, Wiesenthal B, Dietz E, Rüden H. Survival of vancomycin-resistant and vancomycin-susceptible enterococci on dry surfaces. *J Clin Microbiol* 1998;36:3734–3736.
25. Raven KE, Gouliouris T, Brodrick H, Coll F, Brown NM *et al*. Complex routes of nosocomial vancomycin-resistant *Enterococcus faecium* transmission revealed by genome sequencing. *Clin Infect Dis* 2017;64:886–893.
26. Raven KE, Gouliouris T, Parkhill J, Peacock SJ. Genome-based analysis *of Enterococcus faecium* bacteremia associated with recurrent and mixed-strain infection. *J Clin Microbiol* 2018;56:01520–17.
27. Mascini EM, Troelstra A, Beitsma M, Blok HEM, Jalink KP *et al*. Genotyping and preemptive isolation to control an outbreak of vancomycin-resistant *Enterococcus faecium*. *Clin Infect Dis* 2006;42:739–746.
28. Mascini EM, Jalink KP, Kamp-Hopmans TEM, Blok HEM, Verhoef J *et al*. Acquisition and duration of vancomycin-resistant enterococcal carriage in relation to strain type. *J Clin Microbiol* 2003;41:5377–5383.
29. https://github.com/Victorian-Bioinformatics-Consortium/nesoni.
30. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
31. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
32. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J *et al*. BLAST+: Architecture and applications. *BMC Bioinformatics* 2009;10:421.
33. Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* 2014;30:3399–3401.
34. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
35. Ekseth OK, Kuiper M, Mironov V. orthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics* 2014;30:734–736.
36. van Dongen S, Abreu-Goodger C. Using MCL to extract clusters from networks. *Methods Mol Biol* 2012;804:281–295.
37. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15.
38. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.
39. Bertels F, Silander OK, Pachkov M, Rainey PB, van Nimwegen E. Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Mol Biol Evol* 2014;31:1077–1088.
40. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* 2016;2:vew007.
41. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the beast 1.7. *Mol Biol Evol* 2012;29:1969–1973.
42. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 2012;9:772.
43. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA *et al*. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol* 2012;29:2157–2167.
44. Baele G, Lemey P, Vansteelandt S. Make the most of your samples: bayes factor estimators for high-dimensional models of sequence evolution. *BMC Bioinformatics* 2013;14:85.
45. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res* 2009;19:1639–1645.
46. https://github.com/tseemann/abricate.
47. Hawkey J, Hamidian M, Wick RR, Edwards DJ, Billman-Jacobe H *et al*. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics* 2015;16:667.
48. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* 2006;34:D32–D36.
49. van den Bunt G, Top J, Hordijk J, de Greeff SC, Mughini-Gras L *et al*. Intestinal carriage of ampicillin- and vancomycin-resistant *Enterococcus faecium* in humans, dogs and cats in the Netherlands. *J Antimicrob Chemother* 2018;73:607–614.
50. Weisser M, Oostdijk EA, Willems RJL, Bonten MJM, Frei R *et al*. Dynamics of ampicillin-resistant *Enterococcus faecium* clones colonizing hospitalized patients: data from a prospective observational study. *BMC Infect Dis* 2012;12:68.
51. Duchêne S, Holt KE, Weill FX, Le Hello S, Hawkey J *et al*. Genome-scale rates of evolutionary change in bacteria. *Microb Genom* 2016;2:e000094.
52. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. *Nat Rev Microbiol* 2016;14:150–162.
53. Giraud A, Matic I, Tenaillon O, Clara A, Radman M *et al*. Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science* 2001;291:2606–2608.
54. Zhang S, Lebreton F, Mansfield MJ, Miyashita SI, Zhang J *et al*. Identification of a botulinum Neurotoxin-like toxin in a commensal strain of *Enterococcus faecium*. *Cell Host Microbe* 2018;23:169–176.
55. Camargo ILBC, Zanella RC, Brandileone MCC, Pignatari ACC, Goldman GH *et al*. Occurrence of insertion sequences within the genomes and tn1546-like elements of glycopeptide-resistant enterococci isolated in Brazil, and identification of a novel element, ISEfa5. *Int J Med Microbiol* 2005;294:513–519.
56. Khan MA, Northwood JB, Loor RGJ, Tholen ATR, Riera E *et al*. High prevalence of ST-78 infection-associated vancomycin-resistant *Enterococcus faecium* from hospitals in Asunción, Paraguay. *Clin Microbiol Infect* 2010;16:624–627.
57. Wardal E, Kuch A, Gawryszewska I, Żabicka D, Hryniewicz W *et al*. Diversity of plasmids and Tn1546-type transposons among vanA *Enterococcus faecium* in Poland. *Eur J Clin Microbiol Infect Dis* 2017;36:313–328.
58. Lam MMC, Seemann T, Tobias NJ, Chen H, Haring V *et al*. Comparative analysis of the complete genome of an epidemic hospital sequence type 203 clone of vancomycin-resistant *Enterococcus faecium*. *BMC Genomics* 2013;14:595.
59. Dubin KA, Mathur D, McKenney PT, Taylor BP, Littmann ER *et al*. Diversification and evolution of vancomycin-resistant *Enterococcus faecium* during intestinal domination. *Infect Immun* 2019.
60. Segata N. On the road to strain-resolved comparative metagenomics. *mSystems* 2018;3:e00190–17.
61. Ravi A, Halstead FD, Bamford A, Casey A, Thomson NM *et al*. Loss of microbial diversity and pathogen domination of the gut microbiota in critically ill patients. *bioRxiv* 2019. https://doi.org/10.1101/582494.