Article

Transposable elements are driving rapid adaptation of *Enterococcus faecium*

https://doi.org/10.1038/s41586-026-10373-2

Received: 15 March 2025

Accepted: 6 March 2026

Published online: 22 April 2026

Open access

Check for updates

Matthew P. Grieshop[1,2,3,13], Aaron A. Behr[4,13], Sierra Bowden[1], Jordan D. Lin[5], Marco Molari[6,7], Gabriella ZM Reynolds[1,5], Erin F. Brooks[5,12], Boryana Doyle[3], Ashley A. Moore[5], Guillermo Rodriguez-Nava[8], Jorge L. Salinas[8], Niaz Banaei[8,9,10] & Ami S. Bhatt[1,5,11]✉

Bacterial pathogens adapt rapidly to clinical and within-host selective pressures[1]. Insertion sequences (IS) are transposable elements that can contribute to pathogenic adaptation[2], but their activity and consequences in contemporary clinical populations are not well characterized. Here, combining large-scale genomic surveys with long-read sequencing of clinical isolates and longitudinal gut metagenomes, we quantify pathogen IS dynamics from global patterns to within-host evolution. Across 19,485 publicly available high-contiguity ESKAPEE pathogen genomes, *Enterococcus faecium* genomes are the most IS dense, dominated by replicative ISL3 family elements, which have proliferated in clinical lineages over the past 30 years. We find extensive chromosomal structural variation, largely involving ISL3, within a new single-hospital collection of bloodstream isolates. Long-read metagenomic sequencing of 28 longitudinal stool samples from 12 haematopoietic cell transplantation (HCT) recipients demonstrates within-host IS dynamics and their regulatory consequences. In one patient, an ISL3 insertion upstream of a folate transporter formed a strong promoter, increasing transcription and improving relative fitness under folate limitation. Enhanced folate scavenging may enable *E. faecium* to thrive in the setting of microbiome collapse, which is common in HCT and other critically ill patients[3]. Together, these results show that a recent ISL3 expansion is driving rapid evolution in healthcare-associated *E. faecium*, with consequences for its metabolic fitness that may help explain its increasing clinical burden. Several other pathogens also show elevated IS loads in our survey, which suggests that IS expansion-mediated evolution might be more broadly relevant.

Bacteria experience extreme selective pressures in the modern healthcare environment[1]. The ability to adapt rapidly is a defining feature of successful healthcare-associated pathogens[1]. Single-nucleotide variants (SNVs) and small insertions and deletions (indels) are well known to drive clinical adaptation[4–7]. By contrast, the role of structural variation mediated by transposable elements, particularly insertion sequences (IS), is understudied. Although IS expansions have been well described in the emergence of pathogens such as *Shigella* species[8] and *Bordetella pertussis*[9], the understanding of IS-mediated contributions to short-term clinical adaptation is relatively limited.

IS elements can impact sensitivity to phage predation[10] and antibiotics[11,12], and can also improve metabolic efficiency[13] in laboratory conditions. IS elements, especially replicative ones[14], can disrupt and upregulate genes through mobile regulatory elements[2,15] and mediate structural variations. We postulate that these IS-mediated activities contribute to the evolution of bacterial pathogens on clinical timescales. Understanding how these mechanisms operate could provide crucial insights into the ongoing evolution of healthcare-associated pathogens and inform strategies to combat their persistence and spread.

The ESKAPEE pathogens (*Enterobacter* spp., *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Escherichia coli* and *Enterococcus faecium*) are disproportionately responsible for healthcare-associated infections[1]. Some reports have associated their pathogenicity with an enrichment of IS elements[16–24]. Previous studies of *E. faecium*[23,24] and *A. baumannii*[21]—two bacterial species that can be pathogenic when found in the bloodstream, wounds or abscesses—identified signatures of IS expansion in the emergence of healthcare-associated lineages. For example, the presence of the replicative IS16 element of the IS256 family distinguishes healthcare-associated lineages of *E. faecium* from commensal lineages, implicating IS16 elements in the pathogenicity of *E. faecium*[23]. However, a comprehensive understanding of how replicative IS elements

[1]Department of Genetics, Stanford University, Stanford, CA, USA. [2]Stanford Medical Scientist Training Program, Stanford, CA, USA. [3]Stanford University School of Medicine, Stanford, CA, USA. [4]Department of Biology, Stanford University, Stanford, CA, USA. [5]Department of Medicine, Division of Hematology, Stanford University, Stanford, CA, USA. [6]Swiss Institute of Bioinformatics, Basel, Switzerland. [7]Biozentrum, University of Basel, Basel, Switzerland. [8]Department of Medicine, Division of Infectious Diseases and Geographic Medicine, Stanford University, Stanford, CA, USA. [9]Clinical Microbiology Laboratory, Stanford Health Care, Stanford, CA, USA. [10]Department of Pathology, Stanford University, Stanford, CA, USA. [11]Department of Medicine, Division of Blood and Marrow Transplantation and Cellular Therapy, Stanford University, Stanford, CA, USA. [12]Present address: School of Medicine, Case Western Reserve University, Cleveland, OH, USA. [13]These authors contributed equally: Matthew P. Grieshop, Aaron A. Behr. ✉e-mail: asbhatt@stanford.edu

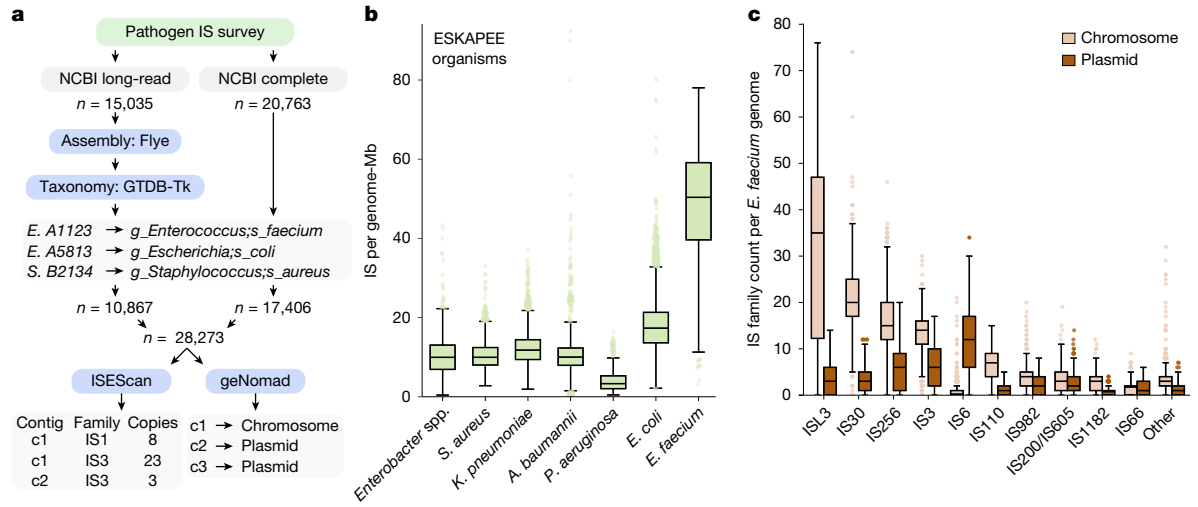


Fig. 1 | Survey of IS in ESKAPEE pathogen genomes reveals ISL3 family elements in *E. faecium* as the most abundant. a, Workflow for counting IS from publicly available pathogen data, either NCBI long-read assemblies (n = 10,867 passing quality control out of 15,035 long-read datasets) or NCBI complete genomes (n = 17,406 passing quality control out of 20,763 initial assemblies). **b**, From 28,273 genomes meeting quality control criteria, 19,485 are ESKAPEE pathogen genomes (*Enterobacter* spp. (536), *S. aureus* (2,672), *K. pneumoniae* (4,627), *A. baumannii* (1,082), *P. aeruginosa* (1,399), *E. coli* (8,530) and *E. faecium* (639)). Total insertion sequence copy per genome normalized by genome assembly length are visualized as boxplots. Centre line, median; box, 25th to 75th percentiles; whiskers, 1.5× IQR; individual values outside 1.5× IQR are shown. **c**, Among 639 *E. faecium* genomes, mean (± s.d.) copy per IS family per genome with counts assigned to chromosome or plasmid contigs as classified by geNomad.

contribute to the ongoing evolution of pathogens, at either the population level or in individual patients, is lacking.

In this study we investigate the role of IS elements in the rapid evolution of healthcare-associated pathogens. We report a large-scale survey of IS elements in highly contiguous pathogen genomes, revealing a very high density of IS, particularly ISL3, in *E. faecium*. A temporal analysis demonstrates that ISL3 family elements have undergone marked expansion since the 1990s, coinciding with the increasing clinical importance of *E. faecium*[25,26]. Using long-read DNA sequencing of clinical *E. faecium* isolates from a modern hospital-endemic lineage, we find that ISL3 contributes disproportionately to extensive recent genome rearrangements. Subsequently, longitudinal metagenomic analyses of gut microbiome samples from people undergoing allogeneic haematopoietic cell transplantation (HCT) demonstrates IS-mediated structural variants in *E. faecium* occurring within individual patients over the course of their treatment. These variants showed different metabolic behaviour, driven by an ISL3 insertion. Specifically, an ISL3 insertion upstream of *folT* generated a strong hybrid promoter, substantially enhancing folate scavenging—a trait that seems to confer an important advantage in the gut environment of severely ill patients. Taken together, careful quantification and placement of highly repetitive genetic elements in bacterial pathogen genomes reveals a dynamic, ISL3-dependent mechanism for *E. faecium* evolution and proposes a path for its gut domination.

IS in ESKAPEE pathogens

Early surveys in prokaryotes found substantial IS content heterogeneity, even among closely related species[17,18], but were constrained by the small number of sequenced genomes at the time. High-throughput sequencing increased greatly the quantity of available genomic data. However, as repetitive elements such as high-copy IS are often fragmented or collapsed in short-read draft assemblies, accurate quantification remained difficult. The growing adoption of long-read sequencing now enables comprehensive comparison of IS abundance and genomic context across thousands of pathogen genomes.

To characterize the distribution and diversity of IS elements systematically across clinically important bacteria, we quantified IS content with a standardized workflow (Fig. 1a and Supplementary Tables 1–3) in 19,485 highly contiguous genomes from the ESKAPEE pathogens: *Enterobacter* spp. (536), *S. aureus* (2,672), *K. pneumoniae* (4,627), *A. baumannii* (1,082), *P. aeruginosa* (1,399), *E. coli* (8,530) and *E. faecium* (639) (Fig. 1b), as well as 8,788 genomes from these and other genera of pathogenic species, some with notable IS expansions[8,9] (*Shigella*, *Bordetella*, *Streptococcus* and *Salmonella;* Extended Data Fig. 1).

Among the ESKAPEE pathogens, *E. faecium* had the highest IS density (mean, μ = 47.4 copies per Mb; s.d., σ = 14.9; Fig. 1b). Although each ESKAPEE taxon in this dataset comprised numerous lineages that vary in IS content, the same pattern was observed at the subspecies level, with several high-identity clusters of *E. faecium* genomes (Methods) displaying the highest IS density (Extended Data Fig. 2). In *E. faecium*, ISL3 was the most abundant family (μ = 35.2 copies per genome; σ = 21.0), followed by IS30 (μ = 24.6; σ = 8.5), IS256 (μ = 21.7; σ = 9.7), IS3 (μ = 19.4; σ = 6.6), IS6 (μ = 13.2; σ = 7.2), and IS110 (μ = 7.1; σ = 4.3) (Fig. 1c). As IS elements are often carried on plasmids[2], we distinguished chromosomal and plasmid contigs (Fig. 1c and Extended Data Fig. 3). ISL3 elements were enriched particularly on chromosomal contigs. By contrast, IS6 was predominantly plasmid-borne, consistent with the presence of IS6 in the Tn1546 composite transposon conferring clinical vancomycin resistance[27]. ISL3 copy number correlated with antimicrobial resistance gene content at low but not high ISL3 copy number (Extended Data Fig. 4a). Together, these data indicate that ISL3 copy number increases independently of its association with antimicrobial resistance genes.

Having identified *E. faecium* as the ESKAPEE pathogen with the highest IS density, we investigated sequence features and signatures of recent activity of the most common IS in *E. faecium*. ISL3 elements encoded the largest transposases of all high-copy IS elements in *E. faecium*, and they also had longer 5′ and 3′ flanking regions than the other IS elements with a DDE catalytic triad (Extended Data Fig. 4b,c). We reasoned that, for recently active, replicative IS elements, more transposase copies would exist from fewer sequence variants, reflecting limited time for divergence[17]; indeed, we found that ISL3 transposase nucleotide sequences were more conserved than those of other abundant IS transposases (Extended Data Fig. 4d). The four most common variants alone, corresponding to ISEfa11, ISEfa5 and two unnamed ISL3 transposases, accounted for 70% of the total (14,176 of 20,289)

ISL3 transposase copies in the dataset. Notably, IS1251—a previously described ISL3 element known for its integration in the vancomycin resistance-conferring *vanA* operon[28]—was less abundant (909 of 20,289) (Extended Data Fig. 4c).

Together, these results provide a population-scale perspective on IS distributions in prominent healthcare-associated human pathogens. *E. faecium* is distinguished among the ESKAPEE taxa by the greatest IS burden and an enrichment of chromosomal ISL3 elements. The high proportion of identical ISL3 transposase copies suggests recent replicative activity and raises the possibility that ISL3 transposition may contribute dynamically to the ongoing adaptation of *E. faecium* in clinical settings.

ISL3 elements in Gram-positive pathogens

Previous studies proposed that mobile genetic elements, including replicative IS, drove the emergence of a subclade of *E. faecium* (clade A) enriched with pathogenic isolates from its commensal ancestor, *Enterococcus lactis* (formerly clade B *E. faecium*, until 2022)[23,24,29]. Clade A is further separated into healthcare- (A1) and animal-associated (A2) subclades and, although one study[24] noted enriched IS3, IS110 and IS256 in A1, it did not mention ISL3. From a collection of high-contiguity *Enterococcus* genomes (defined in Methods), we examined the phylogenetic distribution of ISL3 across 785 *E. faecium* genomes. Each was assigned to clade A1, A2 or *E. lactis*[29] (Fig. 2a,b). As suspected, ISL3 was enriched in the healthcare-associated A1 clade (median: 44 copies per genome) compared with A2 (2 copies) and *E. lactis* (1 copy) (Fig. 2a,b). ISL3 copy number was high across diverse clinical *E. faecium* lineages, defined both by sequence type (ST) and genomic distance (Fig. 2a and Extended Data Fig. 2), with substantial variability even among closely related genomes (maximum 86 copies; interquartile range (IQR) 19–59), potentially consistent with ongoing ISL3 activity (Fig. 2a).

Non-*faecium*, non-*faecalis* (NFF) enterococci capable of human infection are of growing clinical concern[26,30]. To determine whether ISL3 abundance extends beyond *E. faecium*, we analysed 2,112 high-contiguity *Enterococcus* genomes spanning 32 *Enterococcus* and outgroup taxa as defined by Lebreton and colleagues[29], quantifying copy numbers of the six most abundant IS families found in *E. faecium* (Fig. 1c). ISL3 is also abundant in several NFF enterococci associated with human infection, including *Enterococcus avium*, *Enterococcus raffinosus* and *Enterococcus durans*[30] (Fig. 2b). Extending this analysis to other Gram-positive pathogens revealed abundant ISL3 in several *Streptococcus* and *Staphylococcus* species, including a subset of *S. aureus* (Fig. 2c). Additional species-level analyses and taxonomic context are provided in Supplementary Note 1.

Together, these results support a connection between ISL3 abundance and pathogenic potential across several Gram-positive cocci species. As ISL3 abundance is not restricted taxonomically, this raises the hypothesis that independent acquisition and enrichment of ISL3 elements in a clinical context has occurred among Gram-positive pathogens.

ISL3 expanded recently in *E. faecium*

Earlier reports of IS expansions in healthcare-associated *E. faecium* did not identify ISL3, perhaps because it is not particularly abundant in the healthcare-associated clade reference genome isolated in 1998 (refs. 23,24,31). Therefore, we postulated that the high copy number (Fig. 1c) and sequence conservation of ISL3 (Extended Data Fig. 4d) might reflect a relatively recent expansion of these elements.

Highly contiguous genomes enable direct quantification of IS copy number, but are scarce for isolates collected before the 2010s. Thus, we developed an approach for estimating genomic IS copy number from short-read isolate sequencing data (Fig. 3a). We benchmarked the performance of our heuristic by generating both short- and long-read sequencing data from the same set of isolates (Fig. 3b and Extended Data Fig. 5a), and observed comparable performance in both *E. faecium* and *E. faecalis* (Extended Data Fig. 5b). We then estimated IS family copy per genome from 5,422 National Center for Biotechnology Information (NCBI) short-read *E. faecium* assemblies sampled from before 1990 through 2025 (Fig. 3c and Extended Data Fig. 5c). Although IS256, IS30 and IS3 were abundant in the earliest genomes, they remained relatively stable over time (Fig. 3c and Extended Data Fig. 5c). By contrast, ISL3 elements increased from around 17 mean copies per genome in 1995 to around 35 per genome in 2025, with a peak of around 40 per genome in 2020. Furthermore, IS200/IS605 increased over the sampling window from approximately five to approximately ten copies per genome.

Increased sampling of *E. faecium* lineages enriched for ISL3 could explain this observation. To address this possibility, we considered the prevalence of different lineages over time in the collection (Fig. 3d). For isolates before 2005, the most common STs were ST16, ST17 and ST18. After 2015, other lineages, including ST80 and ST117, became prominent. In ST17 and ST18, which are relatively common in both early (pre-2005) and recent (post-2015) time periods, we found ISL3 counts were substantially higher in recent (post-2015) samples compared with early (pre-2005) samples in both ST17 (median $\Delta = +26.3$; two-sided Mann–Whitney $P = 2.1 \times 10^{-26}$) and ST18 ($+31.7$, $P = 1.5 \times 10^{-10}$). Furthermore, high ISL3 abundance was observed across the prevalent post-2015 lineages (Fig. 3e). These findings demonstrate ISL3 has proliferated in healthcare-associated *E. faecium* populations over the last several decades.

To assess whether similar IS expansions have occurred in other organisms, we applied our approach to *Staphylococcus epidermidis*, another species for which IS element proliferation has played a role in adaptation of hospital clades (ST2, ST5 and ST23) from low-virulence commensals to pathogens[32,33]. Among 1,604 non-skin-associated *S. epidermidis* genomes, we find trends of increasing IS1182, decreasing IS256 and clade-specific IS copy number changes over time (Extended Data Fig. 6).

Extensive ISL3 structural variation

Longitudinal studies of *E. faecium* isolates collected from individual patients have measured IS transposition activity[34,35]. However, the broader landscape of IS-driven genome plasticity in a present-day endemic hospital lineage has not been explored. To address this gap and investigate how ISL3 might underpin the success of *E. faecium* in contemporary clinical settings, we sought to determine (1) the extent of short-term structural variation in a single-hospital lineage of *E. faecium* and (2) whether ISL3 activity contributes to any observed variation.

Highly contiguous and minimally divergent genomes facilitate the accurate detection of structural variation. Therefore, we collected every bloodstream isolate of *E. faecium* ($n = 109$) from Stanford Health Care (SHC) in 2021 and 2023, performed long-read sequencing, assembled complete or near-complete genomes ($n = 100$), and identified clusters of highly related genomes (Fig. 4a). Consistent with epidemiological studies from other institutions[36–38], a dominant lineage, ST117 was present (66 of 100; Fig. 4b).

We also employed the same approach for *E. faecalis* ($n = 166$)—a species with reported IS expansion[19,20], similar genome size and infection burden at the same hospital (Fig. 4a). We identified an ST179 cluster of 44 *E. faecalis* (Fig. 4e). The recombination-filtered core genome divergence of the ST179 *E. faecalis* and ST117 *E. faecium* clusters are comparable (about 0.0009 and about 0.0005 substitutions per core genome-aligned position, respectively), indicating a slightly higher total evolutionary distance for the *E. faecalis* cluster (Extended Data Fig. 7). Therefore, the relative importance of recent structural variation between these pathogens can be compared meaningfully.

In the *E. faecium* ST117 and *E. faecalis* ST179 clusters, we measured recent IS activity by employing a reference-free, gene-agnostic

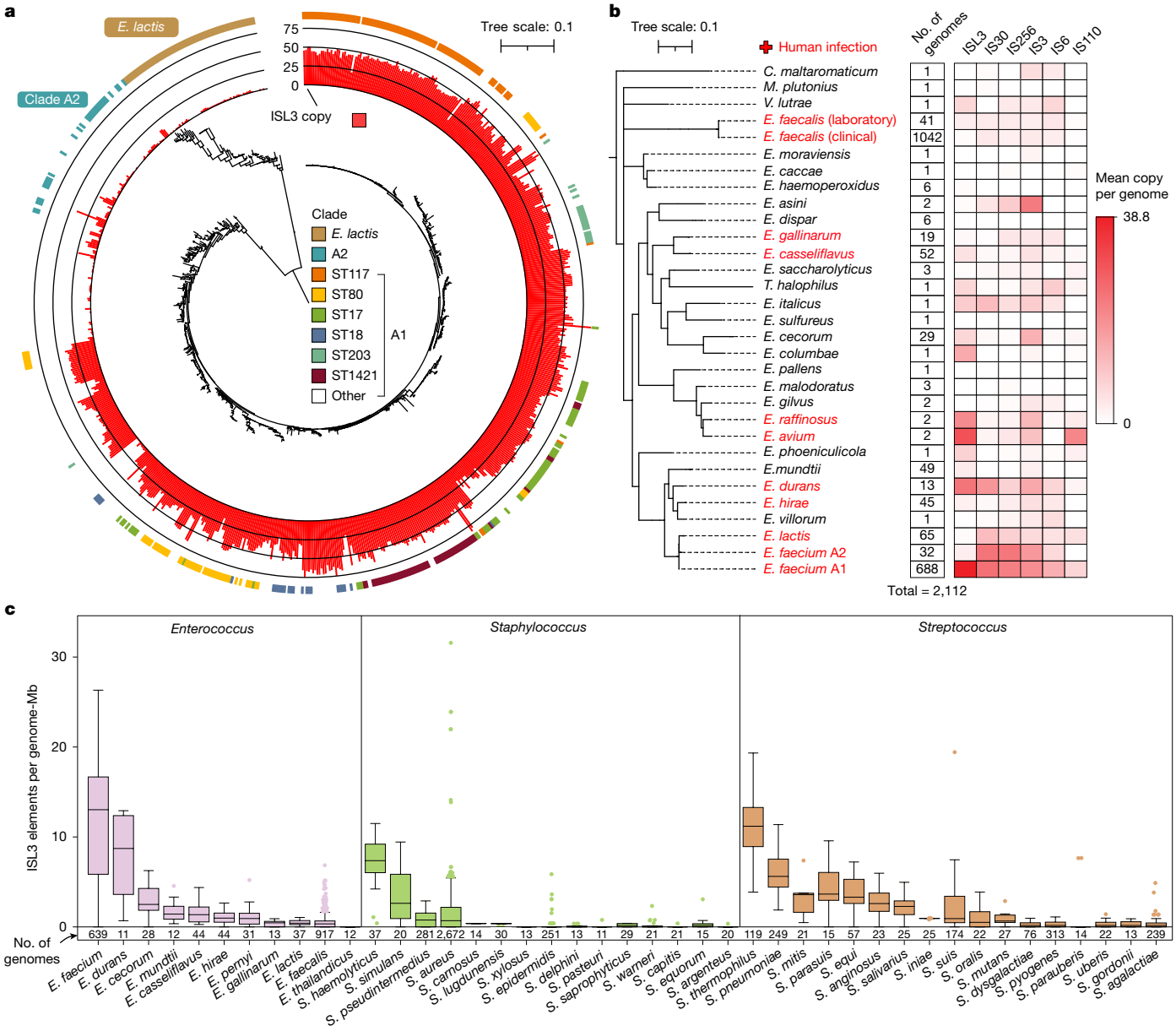


Fig. 2 | ISL3 elements are enriched in clinical *E. faecium* and abundant in other Gram-positive pathogens. a, ISL3 elements (red bar) per genome among 785 *E. faecium* and *E. lactis* (formerly clade B *E. faecium*) genomes from the high-contiguity *Enterococcus* collection displayed across a core pangenome phylogeny (midpoint rooted between *E. faecium* and *E. lactis*). IS counts are provided in Supplementary Tables 2 and 3. Taxonomic clades (outer ring, coloured squares), including *E. lactis*, clade A2, and multi-locus sequence typing for six common STs, are labelled. Genomes defined by our approach as clade A1 with an undetermined or less common ST lack a label. Note that the close relationship and probable genetic exchanges between clades A1 and A2 complicate their accurate delineation[42,43]. **b**, GTDB-Tk based phylogenetic tree of *Enterococcus* genomes from 2,112 high-contiguity *Enterococcus* collapsed by nearest distance to the list of representative *Enterococcus* genomes and outgroups from Lebreton and colleagues[29]. *C*, *Carnobacterium*; *E*, *Enterococcus*; *M*, *Melissococcus*; *T*, *Tetragenococcus*; *V*, *Vagococcus*. Heatmap values are the mean IS copy per genome among all genomes assigned to a representative. *Enterococcus* species are classified as 'human-infectious' and coloured red if they were identified to cause at least one bloodstream infection per year in the 10-year Australian *Enterococcus* Surveillance Outcome Program[30]. **c**, Distribution of ISL3 elements per genome-Mb across surveyed genomes from NCBI complete and NCBI long-read datasets for each species, separated by genus, are shown as boxplots. Centre line, median; box, 25th to 75th percentiles; whiskers, 1.5× IQR; individual values outside 1.5× IQR are shown. Note, this does not include the SHC long-read *Enterococcus* isolates or representative *Enterococcus* included in **a** and **b**.

approach based on that described by Molari and colleagues[39] (Fig. 4a). In the ST117 cluster, we identified 245 structurally variable loci (Fig. 4c,d and Supplementary Table 4). IS insertions accounted for 67% of these 'junctions', with ISL3 predominating (Fig. 4c,d). By contrast, the ST179 *E. faecalis* cluster had fewer structurally variable loci ($n = 77$) and fewer IS-mediated structural variations (Fig. 4f,g and Supplementary Table 4). This is consistent with a relatively prominent role for recent IS activity in the ongoing evolution of *E. faecium*. This analysis also identifies low-activity IS elements in *E. faecium*. Although the combined abundance of IS30, IS256, IS3 and IS110 elements exceeds that of ISL3 (Fig. 1c), ISL3 elements account for twice as many recent variants as the other four combined (100 and 50, respectively; Fig. 4h).

As genome reductive effects of IS expansion have been observed repeatedly in other species[8,9,40], we next investigated whether active ISL3 variants might cause insertional disruptions. Among 100 distinct ISL3 insertions observed within ST117, only 11 disrupted open

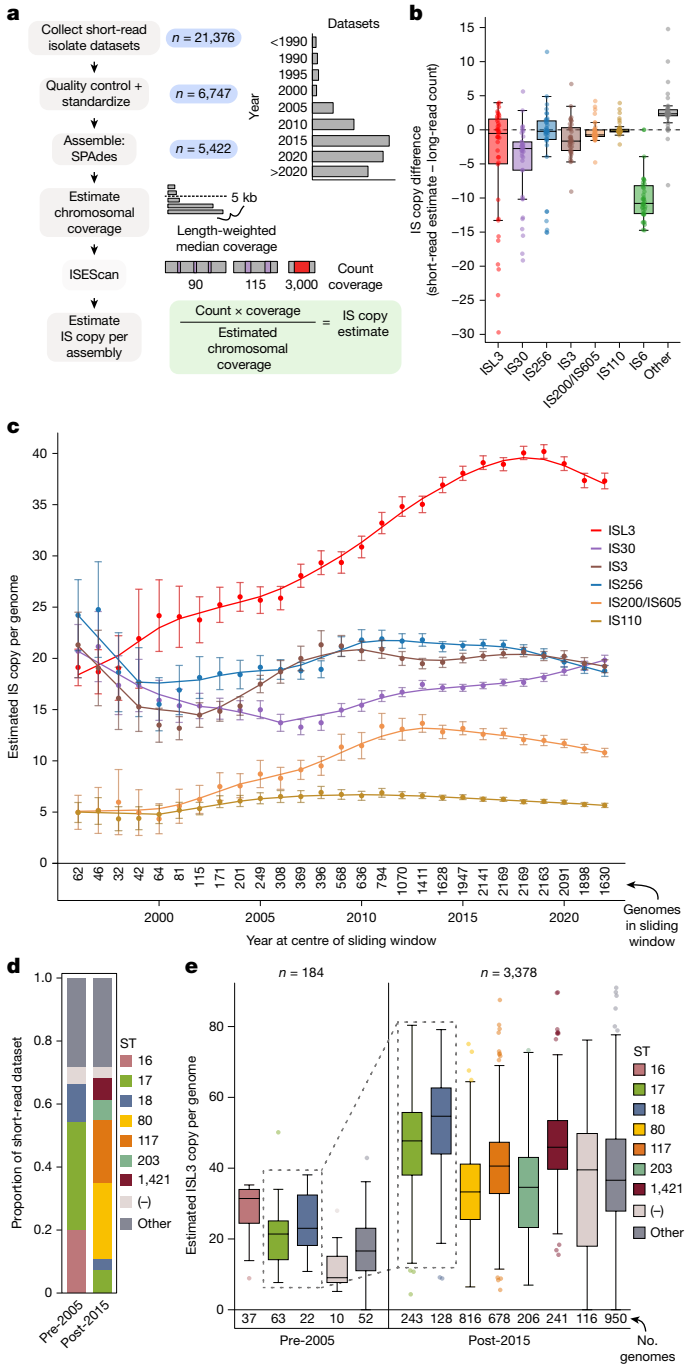


Fig. 3 | Temporal analysis of IS abundance in healthcare-associated *E. faecium* demonstrates recent ISL3 element proliferation. a, Workflow for estimating IS abundance in short-read *E. faecium* data using assembly coverage. A total of 21,376 publicly available datasets were retrieved from SRA, filtered for coverage and collection date ($n = 6,747$), processed and assembled, and 5,422 high-quality assemblies were annotated with ISEScan. IS family copy number per genome was estimated by multiplying IS count per contig by normalizing contig coverage to chromosomal coverage. **b**, *E. faecium* isolates ($n = 37$ biologically independent samples) from the SHC long-read *Enterococcus* isolates were short-read sequenced. Per IS family counts were obtained from long-read data as before or estimated from short-read data as in **a**. Comparison of long-read IS counts and short-read estimates are shown as boxplots with individual values overlaid. Centre line, median; box, 25th to 75th percentiles; whiskers, 1.5× IQR. **c**, Mean IS copy number per family per genome across 5-year sliding windows is shown as LOWESS-smoothed trend lines with 95% confidence intervals. The number of genomes considered per window is printed below each point. **d**, Proportion of genomes belonging to each multi-locus sequence typing ST before 2005 ($n = 184$) and after 2015 ($n = 3,378$). **e**, Estimated ISL3 copy number per genome among common STs before 2005 and after 2015, shown as boxplots as in **b**. For STs with at least 20 samples per time bin, ISL3 abundance increased significantly over time, including ST17 (median 21.4 to 47.7 copies; $\Delta = +26.3$; two-sided Mann–Whitney $P = 2.1 \times 10^{-26}$) and ST18 (23.0 to 54.7; $\Delta = +31.7$; $P = 1.5 \times 10^{-10}$).

reading frames (ORFs)—a markedly lower proportion than for other common IS families (Fig. 4h). Instead, ISL3 elements inserted frequently into intergenic regions, typically on the opposite strand of and near the downstream gene (Fig. 4i). These findings highlight the striking genome plasticity mediated by active ISL3 transposition taking place over a few years in a single hospital-endemic *E. faecium* lineage.

ISL3 insertion drives folate scavenging

E. faecium can colonize and dominate the gastrointestinal tract when the microbiome 'collapses' in people undergoing HCT[41], which often precedes bloodstream infections[42]. Previous studies have reported SNVs in *E. faecium* during intestinal domination and the transition to bloodstream infection[5,43], whereas the impact of IS-mediated structural variation has not been evaluated clearly in vivo. Replicative ISL3 activity might induce structural variation in *E. faecium*. To investigate this hypothesis, we performed longitudinal long-read metagenomics on a set of stool samples from patients (34 samples; 14 patients) who underwent HCT at Stanford Hospital[44,45], with evidence of *E. faecium* domination at several time points (at least two samples, greater than 10% relative abundance; Extended Data Fig. 8). Direct metagenomic sequencing from stool avoided culture bottlenecks and recovered an *E. faecium* chromosomal contig in 28 of 31 samples that passed quality filtering (Extended Data Figs. 8 and 9). By comparing the reads from each sample with their own de novo assembly, we detected evidence of within-sample *E. faecium* structural heterogeneity in 27 of 28 stool samples with a chromosomal contig (Extended Data Fig. 9f). Additional metagenomic sequencing statistics and epidemiological analyses are described in Supplementary Note 2 and Supplementary Table 5. Next, we searched for IS variants that increased in frequency over time within individual *E. faecium* populations. In those patients with confirmed longitudinal strain identity (11 of 12 patients; Extended Data Fig. 9g), structural variants were called using reads from each time point aligned to the earliest metagenomic assembly. IS variants were dynamic in 7 of 11 cases (Extended Data Fig. 9h and Supplementary Table 6). In p11568—an individual patient who had undergone allogeneic HCT for treatment of acute myeloid leukaemia—two ISL3 insertions reached high frequency by the final time point.

We further examined the ISL3 structural variants of *E. faecium* in four longitudinal, *E. faecium*-dominated stool samples (t1–t4, corresponding to day −2, 12, 19 and 26 relative to transplant, day 0). Two variants emerged in this time frame: a 1,493-bp ISL3 variant inserted 42 bp upstream of and co-oriented with a folate transporter gene (*folT*) (Fig. 5a–c), and another disrupted *agrC* (Extended Data Fig. 10b–d). Neither variant was detectable in t1 but both neared fixation by t4 (Fig. 5a and Extended Data Fig. 10b–d).

To investigate whether the *folT*-adjacent ISL3 regulates the expression of *folT*, we isolated *E. faecium* from the t1 and t4 stool samples. The t1 and t4 isolates were found to be nearly identical (no SNVs or indels), differing only by the position of the two additional ISL3 elements as described above and a putative plasmid present only in t4. RNA sequencing (RNA-seq) demonstrated that *folT* was the most upregulated gene in t4 relative to t1 (Fig. 5b, Supplementary Table 7 and Extended Data Fig. 10a). Furthermore, the disrupted *agr* operon was silenced in t4 (Extended Data Fig. 10). The ISL3 insertion upstream of the *folT* locus did not disrupt the native −10 promoter sequence (Pribnow box; Fig. 5c), but did donate a 'perfect' −35 sequence (5′-TTGACA-3′),

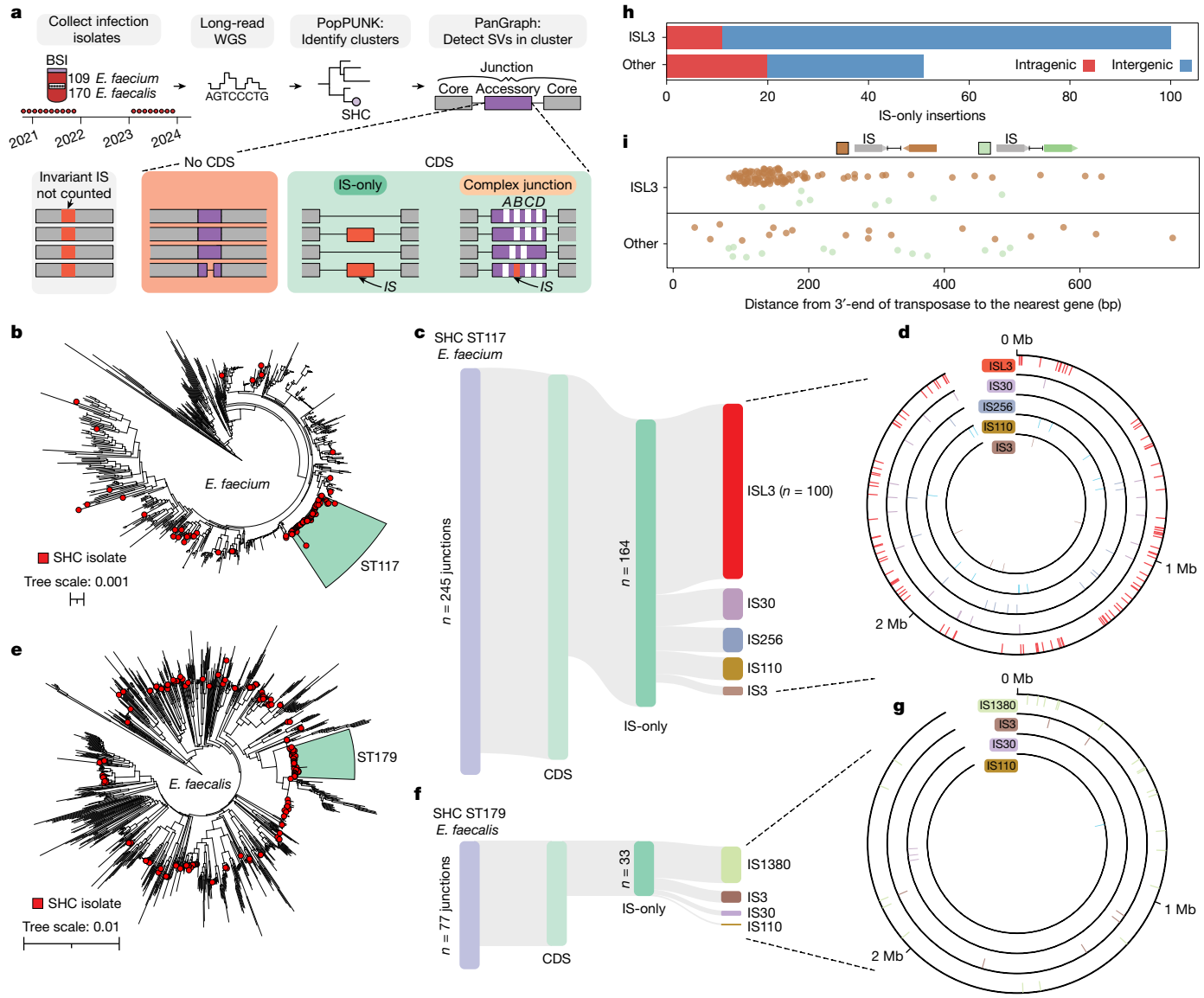


Fig. 4 | ISL3 mediates extensive recent structural variation in a hospital-endemic *E. faecium* lineage. a, Analysis of recent IS activity among a closely related hospital lineage using PanGraph. CDS, coding sequence. **b**, PopPUNK clustering and phylogeny of *E. faecium* isolates from SHC (this study; red tips) among publicly available genomes. An *E. faecium* cluster of ST117 genomes is highlighted ($n = 66$). **c**, IS variants (164 of 245 total) detected in ST117 *E. faecium*. **d**, Circos plot showing the genomic location of each IS variant from **c**. Rings are ordered and elements coloured the same as in **c** with ISL3 variants in the outermost ring and IS3 in the innermost. **e**, PopPUNK clustering and phylogeny of *E. faecalis* isolates from the SHC (this study; red tips) among publicly available genomes. An *E. faecalis* cluster of ST179 genomes is highlighted ($n = 44$). **f**, IS variants (33 of 77 total) detected in ST179 *E. faecalis*. **g**, Circos plot showing the genomic location of each IS variant from **f**. Rings are ordered and elements coloured the same as in **f** (IS1380 in the outer ring, IS110 in the inner ring). **h**, Among simple insertion variants in SHC ST117 *E. faecium*, the number of inter- versus intra-genic insertions for ISL3 and the other IS elements named in **c** is displayed. **i**, For each category, the distance from the 3′ end of the transposase coding sequence to the next gene is displayed. Each dot is coloured by whether the IS is oriented in the same (green) or opposite (brown) direction as the downstream gene.

replacing the native −35 sequence (5′-ATGACA-3′) and the surrounding sequence context (Fig. 5c). These findings demonstrate this ISL3 element carries a 3′-outwardly directed −35 promoter sequence that can drive expression when positioned correctly upstream of a Pribnow box and transcriptional start site.

Healthcare-associated *E. faecium* is clinically resistant to antifolate antibiotics such as trimethoprim through the uptake of exogenous folates[46]. We wondered whether the increased expression of *folT*, which encodes the substrate-specific component of an energy-coupling factor folate transport complex[47], might have been selected for by treatment with antifolate medications. However, patient 11568 was not exposed to methotrexate or trimethoprim-sulfamethoxazole (TMP-SMX) during the time frame of our sampling (Fig. 5d). Therefore, we reasoned that increased *folT* expression had some other adaptive benefit, such as enhancing growth of *E. faecium*—a known folate auxotroph[48]—in the setting of limited folate in the gut. To test whether higher *folT* expression in the t4 isolate confers a fitness advantage over t1 under folate-limiting conditions, we measured in vitro growth during TMP-SMX treatment and evaluated whether exogenous folate in the medium differentially rescues antifolate toxicity (Fig. 5e,f and Extended Data Fig. 10e,f). Indeed, the t4 isolate had a marked growth advantage over t1 at all drug concentrations tested (Fig. 5e). This difference was abrogated when the medium was supplemented with folinic acid (a synthetic folate), which bypasses antifolate toxicity (Extended Data Fig. 10e,f). Conversely, we observed that folic acid supplementation increased drug susceptibility in both isolates and particularly in t4, probably

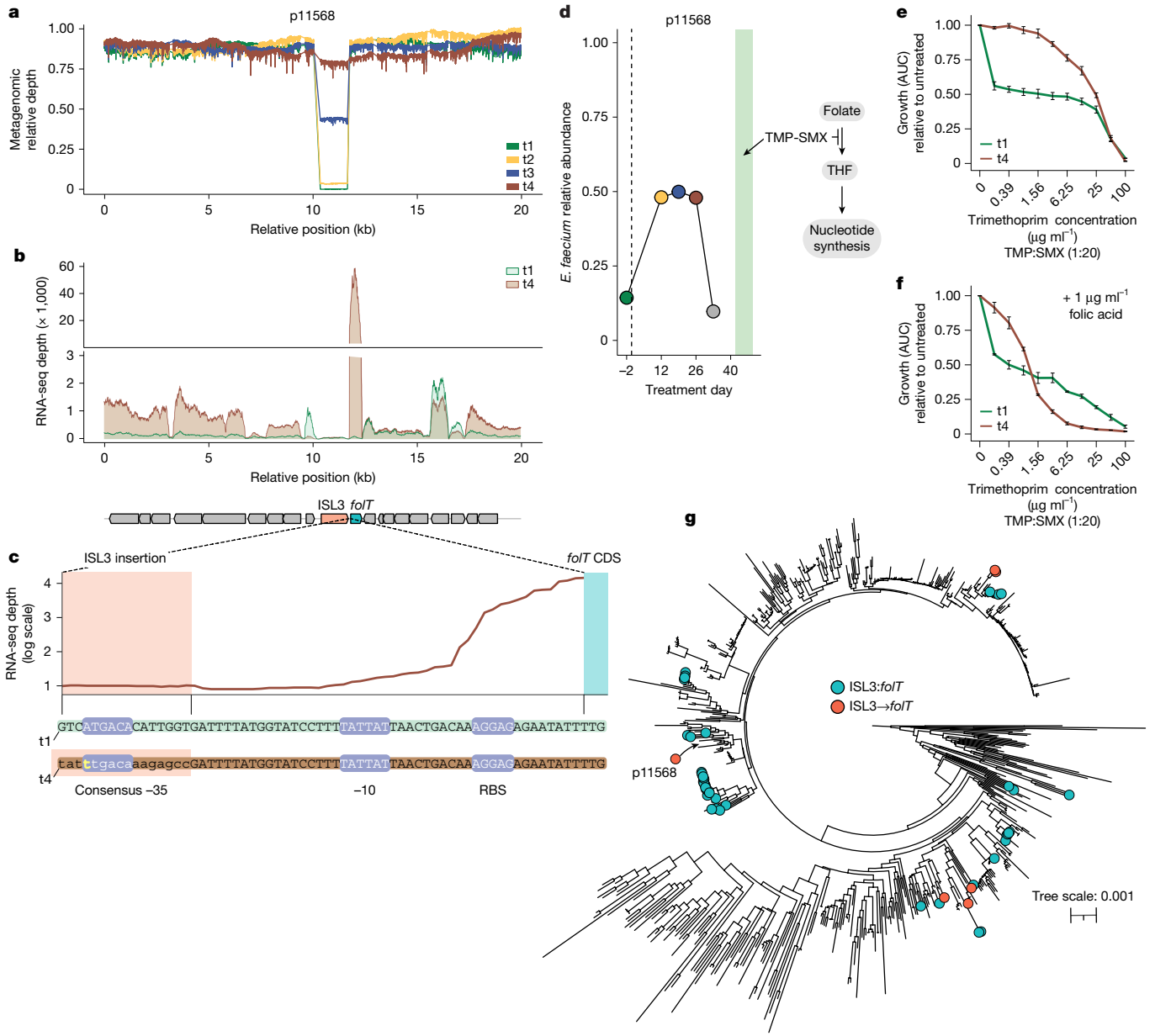


Fig. 5 | ISL3 insertion drives folate scavenging. a, Long-read coverage across a 20-kb region around the ISL3 insertion upstream of *folT* in the *E. faecium* p11568 t4 metagenome-assembled chromosome, normalized to mean chromosomal coverage. **b**, RNA-seq coverage across the same locus in *E. faecium* isolates from p11568 t1 (technical triplicate) and t4 (technical quadruplicate). **c**, Sequence context of the ISL3-derived hybrid promoter between the 3′ end of ISL3 (lowercase, red) and the *folT* CDS (cyan). Mean t4 RNA-seq depth (log scale) is shown above the aligned t1 and t4 isolate sequences, with predicted −35, −10 and ribosome-binding sites marked. **d**, Relative abundance of *E. faecium* in p11568 stool samples over time (t1–t4 correspond to days −2, 12, 19 and 26 relative to transplant, day 0, indicated with a dashed line). TMP–SMX treatment is indicated (green box), alongside a schematic of its targets in folate metabolism. **e**, Dose–response of t1 and t4 isolates to TMP–SMX (TMP concentrations shown; 1:20 TMP:SMX). Growth is quantified by normalized AUC (mean ± s.e., $n = 3$ independent experiments in technical duplicate). **f**, Same as **e**, with folic acid supplementation (1 μg ml^{-1}). **g**, Frequency of ISL3 insertions upstream of *folT* in the same orientation (orange) or adjacent but oppositely oriented (blue) across a PopPUNK phylogeny of 687 *E. faecium* genomes; p11568 was placed on the phylogeny by minimum Mash distance. BSI, bloodstream infection; RBS, ribosome-binding site; WGS, whole-genome sequencing.

due to more rapid accumulation of non-productive folate intermediates[49] (Fig. 5f). Together, these findings indicate that ISL3-mediated increased expression of *folT* enhances the ability of p11568 t4 *E. faecium* to scavenge folate metabolites from the environment, conferring a fitness advantage over p11568 t1 under folate-limiting conditions.

Although a variant such as the *folT*-ISL3 variant observed here may reach high frequency in an individual strain due to genetic drift, observing it repeatedly in distinct strains and/or as several distinct genotypes provides compelling evidence for selection. To determine whether several such independent ISL3 insertions have occurred, we investigated the neighbourhood of the *folT* locus. Returning to a global collection, ISL3 was present directly adjacent to *folT* in 50 of the 687 high-contiguity *E. faecium* genomes with a *folT* annotation, suggesting that the *folT* locus might be an insertional hotspot (Fig. 5g and Supplementary Table 8). In five of these, as in the p11568 t4 *E. faecium* genome, an ISL3 element was positioned upstream of, and co-oriented with, *folT*. These ISL3→*folT* variants were found in

bloodstream infection and HCT patient gut samples from the Netherlands, Germany and South Korea. As these insertions occurred in distinct phylogenetic lineages and involved different ISL3 sequence variants, they probably represent independent evolutionary events. Given their global spread, these variants implicate folate scavenging as an important function under selection in some patients. Several other genes not explored in detail here—including *rhgT*—a gene predicted to be involved in pectin metabolism—have frequent proximal ISL3 insertions and warrant further exploration (Supplementary Table 8).

Discussion

Bacterial pathogens are under incredible pressure to evolve and adapt, and IS movement and expansion are prominent mechanisms by which rapid evolution can occur. Here we report the recent replicative ISL3 expansion in healthcare-associated *E. faecium*. Long-read metagenomic studies showed several (up to six) 'IS variant strains' in a single gut microbiome. This enables direct evaluation of the natural population of *E. faecium* in the clinical environment, without the bottlenecking effect of isolation and culture. Finally, we demonstrate ongoing, extensive structural variation mediated by ISL3 elements, and find that these variants can alter gene transcription, driving the expression of genes such as *folT*, which contributes to improved competitive fitness. Open questions remain, including how generalizable the ISL3 expansion phenomenon is, the mechanism of ISL3 expansion, the driving forces that led to the dominance of ISL3-replete organisms and the potential downstream consequences of expanded copies of repetitive elements on the genome.

Although *E. faecium* is particularly enriched for ISL3, these elements are not strictly taxonomy bound; they are present in various *Bacilli*, such as other enterococci, streptococci and staphylococci. However, in those organisms, ISL3 copy number is generally much lower than in *E. faecium*, except in very specific strains of other *Bacilli*. ISL3 is also present in Gram-negative organisms[50], and is abundant in *Bordetella holmesii*. Previously described mechanisms that control IS copy number include both sequence-level variation in transposases[2], host-encoded defence systems that constrain mobile elements[51] and linkage with other elements under strong selection[52]. The documented loss of CRISPR–Cas and restriction–modification pathways during the emergence of healthcare-associated *E. faecium* may have created a permissive backdrop for ISL3 amplification[51].

The exact mechanism that has driven the recent expansion of ISL3 in the past 30 years is as yet unclear. Other abundant IS elements in healthcare-associated *E. faecium*, such as IS*256*[23], have been at high but stable copy number for the past several decades. ISL3 copy number might have increased due to the emergence or introduction of a hyperactive variant of ISL3. Furthermore, selective fitness benefits may arise when the outwardly directed promoter region of ISL3 upregulates genes that confer adaptive traits, such as enhanced folate scavenging or pectin metabolism. Consistent with this, previous work has demonstrated examples of outwardly directed promoter elements on IS that can drive transcription of neighbouring genes and adaptive fitness benefits such as antibiotic resistance[2,15]. Finally, insertion into intergenic regions might reflect tropism to AT-rich regions or may be due to the fact that such insertions are less likely to impact fitness negatively[2,17].

Another consequence of having many identical elements within a genome is recombination between the elements, which can result in deletion of regions of the genome. This has been described previously with expansion of an IS element followed by genome reduction, for example, in *Bordetella pertussis*[9], *Yersinia pestis*[40] and in the evolution of *E. coli* into *Shigella* lineages[8]. In these cases, genome reduction contributed to a narrower host range, and a more pathogenic lineage with a tremendous impact on human society. For example, *Shigella*, unlike *E. coli*, has no known reservoirs outside of humans. The notion that *E. faecium* may be following a similar path toward hospital specialization through ISL3 expansion and subsequent genome reduction[53] is supported by the observation that healthcare-associated strains are restricted mainly to critically ill patients and have limited prevalence in healthy gut microbiomes[41]. Future, more focused, investigations into whether this type of genome reduction is occurring in healthcare-associated *E. faecium* lineages will inform our understanding of additional consequences of ISL3 expansion.

Our work has several limitations. First, the isolate datasets we compiled probably underestimate the population dynamics of transposable elements due to genetic bottlenecks in the isolation process, collapsing of population heterogeneity during assembly, the non-randomness of sampling and the relative paucity of older samples. Second, although we find extensive evidence of IS-mediated structural variation in the SHC hospital-endemic ST117 lineage, our analysis was limited to bloodstream infection isolates. We did not sample hospital surfaces or patient guts—the main reservoirs of *E. faecium* in the hospital. Third, we analysed chromosomal structural variants; future studies will be required to assess plasmids, which are often enriched in transposable elements[52,54] and are important mediators of *E. faecium* pathogenicity[55]. Finally, our study design does not allow us to differentiate definitively active transposition on clinical time scales from selection on standing variation in the community.

In conclusion, ISL3 is a contemporary evolutionary force in hospital-associated *E. faecium*, and IS elements probably play a larger role in the evolution of pathogens than has been appreciated previously. Using both isolate and metagenomic datasets, we demonstrate how bacterial pathogens leverage mobile genetic elements to swiftly respond to the intense selective pressures of clinical settings. These findings advance our knowledge of bacterial evolution and may contribute to new, effective strategies for combating the rising threat of healthcare-associated infections through regulation of IS elements. Future studies that measure the types and consequences of integrated mobile genetic elements in large-scale datasets, as well as efforts to understand what regulates IS element activity and the long-term consequences of IS expansion, will improve our understanding of how these elements shape the evolution of clinically relevant phenotypes in pathogens.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-10373-2.

1. Miller, W. R. & Arias, C. A. ESKAPE pathogens: antimicrobial resistance, epidemiology, clinical impact and therapeutics. *Nat. Rev. Microbiol.* **22**, 598–616 (2024).
2. Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.* **38**, 865–891 (2014).
3. Wirbel, J. et al. Differential effects of two common GVHD prophylaxis regimens on the gut microbiome: Results from the BMT CTN 1801 study. Preprint at *bioRxiv* https://doi.org/10.64898/2026.02.19.706769 (2026).
4. Gouliouris, T. et al. Quantifying acquisition and transmission of *Enterococcus faecium* using genomic surveillance. *Nat. Microbiol.* **6**, 103–111 (2021).
5. Chilambi, G. S. et al. Evolution of vancomycin-resistant *Enterococcus faecium* during colonization and infection in immunocompromised pediatric patients. *Proc. Natl Acad. Sci. USA* **117**, 11703–11714 (2020).
6. Van Tyne, D. et al. Impact of antibiotic treatment and host innate immune pressure on enterococcal adaptation in the human bloodstream. *Sci. Transl. Med.* **11**, eaat8418 (2019).
7. Moradigaravand, D., Martin, V., Peacock, S. J. & Parkhill, J. Evolution and epidemiology of multidrug-resistant *Klebsiella pneumoniae* in the United Kingdom and Ireland. *mBio* **8**, e01976–16 (2017).
8. Hawkey, J., Monk, J. M., Billman-Jacobe, H., Palsson, B. & Holt, K. E. Impact of insertion sequences on convergent evolution of *Shigella* species. *PLoS Genet.* **16**, e1008931 (2020).
9. Parkhill, J. et al. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat. Genet.* **35**, 32–40 (2003).

10. Reif, H. J. & Saedler, H. IS1 is involved in deletion formation in the gal region of *E. coli* K12. *Mol. Gen. Genet.* **137**, 17–28 (1975).
11. Baym, M. et al. Spatiotemporal microbial evolution on antibiotic landscapes. *Science* **353**, 1147–1151 (2016).
12. Durrant, M. G., Li, M. M., Siranosian, B. A., Montgomery, S. B. & Bhatt, A. S. A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. *Cell Host Microbe* **27**, 140–153 (2020).
13. Cooper, V. S., Schneider, D., Blot, M. & Lenski, R. E. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J. Bacteriol.* **183**, 2834–2841 (2001).
14. Sousa, A., Bourgard, C., Wahl, L. M. & Gordo, I. Rates of transposition in *Escherichia coli*. *Biol. Lett.* **9**, 20130838 (2013).
15. Zlitni, S. et al. Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale. *Genome Med.* **12**, 50 (2020).
16. Moran, N. A. & Plague, G. R. Genomic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.* **14**, 627–633 (2004).
17. Wagner, A. Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Mol. Biol. Evol.* **23**, 723–733 (2006).
18. Touchon, M. & Rocha, E. P. C. Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.* **24**, 969–981 (2007).
19. Paulsen, I. T. et al. Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science* **299**, 2071–2074 (2003).
20. Shankar, N., Baghdayan, A. S. & Gilmore, M. S. Modulation of virulence within a pathogenicity island in vancomycin-resistant *Enterococcus faecalis*. *Nature* **417**, 746–750 (2002).
21. Xie, F. et al. Large-scale genomic analysis reveals significant role of insertion sequences in antimicrobial resistance of *Acinetobacter baumannii*. *mBio* **16**, e02852-24 (2025).
22. Armbruster, C. R. et al. Adaptation and genomic erosion in fragmented *Pseudomonas aeruginosa* populations in the sinuses of people with cystic fibrosis. *Cell Rep.* **37**, 109829 (2021).
23. Leavis, H. L. et al. Insertion sequence-driven diversification creates a globally dispersed emerging multiresistant subspecies of *E. faecium*. *PLoS Pathog.* **3**, e7 (2007).
24. Lebreton, F. et al. Emergence of epidemic multidrug-resistant *Enterococcus faecium* from animal and commensal strains. *mBio* **4**, e00534-13 (2013).
25. Murray, B. E. The life and times of the Enterococcus. *Clin. Microbiol. Rev.* **3**, 46–65 (1990).
26. Arias, C. A. & Murray, B. E. The rise of the Enterococcus: beyond vancomycin resistance. *Nat. Rev. Microbiol.* **10**, 266–278 (2012).
27. Vehreschild, M. J. G. T., Haverkamp, M., Biehl, L. M., Lemmen, S. & Fätkenheuer, G. Vancomycin-resistant enterococci (VRE): a reason to isolate? *Infection* **47**, 7–11 (2019).
28. Handwerger, S. & Skoble, J. Identification of chromosomal mobile element conferring high-level vancomycin resistance in *Enterococcus faecium*. *Antimicrob. Agents Chemother.* **39**, 2446–2453 (1995).
29. Lebreton, F. et al. Tracing the enterococci from Paleozoic origins to the hospital. *Cell* **169**, 849–861 (2017).
30. Mullally, C. A., Fahriani, M., Mowlaboccus, S. & Coombs, G. W. Non-faecium non-faecalis enterococci: a review of clinical manifestations, virulence factors, and antimicrobial resistance. *Clin. Microbiol. Rev.* **37**, e00121–e00123 (2024).
31. Lam, M. M. C. et al. Comparative analysis of the first complete *Enterococcus faecium* genome. *J. Bacteriol.* **194**, 2334–2341 (2012).
32. Kozitskaya, S. et al. The bacterial insertion sequence element IS256 occurs preferentially in nosocomial *Staphylococcus epidermidis* isolates: association with biofilm formation and resistance to aminoglycosides. *Infect. Immun.* **72**, 1210–1215 (2004).
33. Schoenfelder, S. M. K. et al. Success through diversity - how *Staphylococcus epidermidis* establishes as a nosocomial pathogen. *Int. J. Med. Microbiol.* **300**, 380–386 (2010).
34. Zhang, X. et al. RNA-seq and Tn-seq reveal fitness determinants of vancomycin-resistant *Enterococcus faecium* during growth in human serum. *BMC Genomics* **18**, 893 (2017).
35. Kirsch, J. M. et al. Targeted IS-element sequencing uncovers transposition dynamics during selective pressure in enterococci. *PLoS Pathog.* **19**, e1011424 (2023).
36. Willems, R. J. L. et al. Global spread of vancomycin-resistant *Enterococcus faecium* from distinct nosocomial genetic complex. *Emerg. Infect. Dis.* **11**, 821–828 (2005).
37. Huycke, M. M., Spiegel, C. A. & Gilmore, M. S. Bacteremia caused by hemolytic, high-level gentamicin-resistant *Enterococcus faecalis*. *Antimicrob. Agents Chemother.* **35**, 1626–1634 (1991).
38. Murray, B. E. et al. Evidence for clonal spread of a single strain of β-lactamase-producing *Enterococcus* (*Streptococcus*) *faecalis* to six hospitals in five States. *J. Infect. Dis.* **163**, 780–785 (1991).
39. Molari, M., Shaw, L. P. & Neher, R. A. Quantifying the evolutionary dynamics of structure and content in closely related *E. coli* genomes. *Mol. Biol. Evol.* **42**, msae272 (2025).
40. Parkhill, J. et al. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**, 523–527 (2001).
41. Taur, Y. et al. Intestinal domination and the risk of bacteremia in patients undergoing allogeneic hematopoietic stem cell transplantation. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **55**, 905–914 (2012).
42. Dubin, K. A. et al. Diversification and evolution of vancomycin-resistant *Enterococcus faecium* during intestinal domination. *Infect. Immun.* **87**, e00102-19 (2019).
43. Moradigaravand, D. et al. Within-host evolution of *Enterococcus faecium* during longitudinal carriage and transition to bloodstream infection in immunocompromised patients. *Genome Med.* **9**, 119 (2017).
44. Tamburini, F. B. et al. Precision identification of diverse bloodstream pathogens in the gut microbiome. *Nat. Med.* **24**, 1809–1814 (2018).
45. Siranosian, B. A. et al. Rare transmission of commensal and pathogenic bacteria in the gut microbiome of hospitalized adults. *Nat. Commun.* **13**, 586 (2022).
46. Zervos, M. J. & Schaberg, D. R. Reversal of the in vitro susceptibility of enterococci to trimethoprim-sulfamethoxazole by folinic acid. *Antimicrob. Agents Chemother.* **28**, 446–448 (1985).
47. Rodionov, D. A. et al. A novel class of modular transporters for vitamins in prokaryotes. *J. Bacteriol.* **191**, 42–51 (2009).
48. Ramsey, M., Hartke, A. & Huycke, M. in *Enterococci: From Commensals to Leading Causes of Drug Resistant Infection* (eds Gilmore, M. S. et al.) 18–19 (Massachusetts Eye and Ear Infirmary, Boston, 2014).
49. Dale, B. A. & Greenberg, G. R. Effect of the folic acid analogue, trimethoprim, on growth, macromolecular synthesis, and incorporation of exogenous thymine in *Escherichia coli*. *J. Bacteriol.* **110**, 905–916 (1972).
50. Holmes, D. S. et al. ISAfe1, an ISL3 family insertion sequence from *Acidithiobacillus ferrooxidans* ATCC 19859. *J. Bacteriol.* **183**, 4323–4329 (2001).
51. Palmer, K. L. & Gilmore, M. S. Multidrug-resistant enterococci lack CRISPR–cas. *mBio* **1**, e00227-10 (2010).
52. Sastre-Dominguez, J. et al. Plasmid-encoded insertion sequences promote rapid adaptation in clinical enterobacteria. *Nat. Ecol. Evol.* **8**, 2097–2112 (2024).
53. Gilmore, M. S., Lebreton, F. & van Schaik, W. Genomic transition of enterococci from gut commensals to leading causes of multidrug-resistant hospital infection in the antibiotic era. *Curr. Opin. Microbiol.* **16**, 10–16 (2013).
54. DelaFuente, J. et al. Within-patient evolution of plasmid-mediated antimicrobial resistance. *Nat. Ecol. Evol.* **6**, 1980–1991 (2022).
55. Arredondo-Alonso, S. et al. Plasmids shaped the recent emergence of the major nosocomial pathogen *Enterococcus faecium*. *mBio* **11**, e03284-19 (2020).

Methods

Bacterial genomic datasets compiled for this study

Several bacterial genomic datasets were compiled for this study, and different selections of these datasets were used, as appropriate, for the analyses described herein. We refer to these datasets as: 'NCBI complete genomes', 'NCBI long-read assemblies', 'NCBI short-read *E. faecium* assemblies', 'SHC long-read *Enterococcus* isolates' and 'SHC long-read metagenomes'. We also define the 'high-contiguity *Enterococcus* genomes', which consist of *Enterococcus* genomes from (1) NCBI complete genomes, (2) NCBI long-read assemblies, and (3) SHC long-read *Enterococcus* isolates. Datasets were combined without deduplication.

To aid readers in interpreting our manuscript and for transparency, Supplementary Table 1 contains (1) a detailed description of bacterial genomic datasets compiled for this study, (2) descriptions of any additional filtering for specific analyses, and (3) a description of how these datasets are used in each main figure and Extended Data figure.

NCBI long-read assemblies: long-read dataset curation, preprocessing, assembly and quality control

Candidate isolate long-read sequencing datasets were identified on NCBI with the following search criteria on 29 February 2024: ((((Enterococcus[Organism]) OR (Staphylococcus[Organism] OR (Klebsiella [Organism]) OR (Acinetobacter [Organism]) OR (Pseudomonas [Organism]) OR (Enterobacter [Organism]) OR (Escherichia [Organism]) OR (Salmonella[Organism]) OR (Streptococcus[Organism])) AND 'oxford nanopore'[Platform])). Datasets for *Shigella* and *Bordetella* were queried separately on 28 October 2024 in the same way. Metadata for these accessions was downloaded with the Sequence Read Archive (SRA) Run Selector interface. Datasets with less than 50 Mb of total sequencing or less than 2 kb average read length were not included for further analysis. Raw reads were downloaded using SRA tools prefetch and fasterq-dump, and assembled with flye[56] (v.2.9) using the --nano-raw method. Genome taxonomy was verified with GTDB-Tk[57] (v.2.4) with database release v.220.

Assemblies were then filtered for appropriate contiguity (≤20 contigs), genome size (±2 s.d. per genus), coverage (≥40×) and with identified species-level taxonomy. We excluded the GTDB-Tk taxonomy result for the *Shigella* datasets because the GTDB-Tk database does not include *Shigella* genomes and thus cannot classify them beyond *E. coli*. Finally, we filtered the set of genomes to include only species with more than ten samples in this and the NCBI complete genomes (see 'NCBI complete genomes' below) combined. The final set of genomes with assembly and taxonomy information is available in Supplementary Table 2. We refer to this dataset in the manuscript as NCBI long-read assemblies. Relevant source code and data can be found in our repository ('Code Availability').

NCBI complete genomes

The NCBI command line tool ncbi-datasets-cli (v.16.22.1)[58] was used to download all available NCBI genomes with the 'complete' assembly level from the following taxa on 3 July 2024: *Enterobacter*, *Staphylococcus*, *Klebsiella*, *Acinetobacter*, *Pseudomonas*, *Enterococcus*, *Escherichia*, *Salmonella*, *Streptococcus*, *Shigella* and *Bordetella*. For example, the following command was used to download *Enterococcus* assemblies: datasets download genome taxon enterococcus --assembly-level complete --include genome,gff3 --dehydrated --filename enterococcus.zip. Note, this command collects a reference to the relevant genomic data. Data must be downloaded using the 'rehydrate' command. After rehydrating, duplicate assemblies were removed, preferring the RefSeq over GenBank as the source database. Assemblies were then filtered according to the same contiguity, size, coverage and sample count thresholds as above, and the final set is likewise available in Supplementary Table 2. We refer to this dataset in the manuscript as NCBI complete genomes. Relevant source code and data can be found in our repository ('Code Availability').

SHC long-read *Enterococcus* isolates: collection and sequencing

A total of 279 bloodstream infection isolates identified by biotyping with a MALDI-TOF based Bruker Biotyper as *E. faecalis* (n = 170) or *E. faecium* (n = 109) by the Stanford Health Care Clinical Microbiology Laboratory between the periods 1 October 2020 to 31 December 2021 and 1 January 2023 to 31 December 2023 were collected under a protocol approved by the Stanford University Institutional Review Board (Protocol no. 64591; Principal Investigator: A. Bhatt, Committee IRB-61). Informed consent was obtained from all human research participants. An additional eight isolates collected in 2024 identified as atypical *Enterococcus* as above were stored as glycerol stocks. Metadata, including date, site of collection and relevant patient information, were extracted. Each isolate was cultured aerobically for approximately 16 h with shaking at 37 °C in 3 ml Brain Heart Infusion (BHI) broth (Sigma-Aldrich). Then, a 1:100 subculture into 10 ml BHI broth was prepared. Subcultures were allowed to reach an OD of approximately 1.0. The subculture was centrifuged at 3,000*g* for 10 min at room temperature, resuspended in PBS, then centrifuged again. The PBS was poured off and the subsequent pellet was resuspended in 1 ml 1× Zymo DNA/RNA Shield (Zymo, catalogue no. R1100-250) and shipped to Plasmidsaurus for Oxford Nanopore Technologies (ONT) 'Standard Bacterial Genome with DNA extraction' sequencing. After inspection of the assemblies, one *E. faecium* isolate (EF_B_40) was mischaracterized as *E. faecalis* and one *E. faecalis* isolate (FM_B_40) was mischaracterized as *E. faecium*, on the basis of biotyping. Of the 170 *E. faecalis* isolates, 167 were confirmed as *E. faecalis*. Of the 109 *E. faecium* isolates, 107 were confirmed as *E. faecium* with standard quality control performed by Plasmidsaurus. Finally, assemblies were filtered according to the same contiguity and coverage thresholds as above, resulting in 166 *E. faecalis* and 100 *E. faecium* genomes, respectively. We refer to this dataset in the manuscript as SHC long-read *Enterococcus* isolates.

IS counting

ISEScan[59] (v1.7.2.3) was run with default parameters. Each resulting {sample}.csv file was used to count IS elements per sample, per contig and per IS family. Summary results from NCBI long-read assemblies and NCBI complete genomes are included in Supplementary Table 2, and summary results from SHC long-read *Enterococcus* isolates (*E. faecium* and atypical *Enterococcus*) are included in Supplementary Table 3.

Assembly annotation

Genomes were annotated with Bakta[60] (v.1.9.1–v.1.10.3) using database v.5.1 with the following flags --skip-tmrna --skip-ncrna --skip-ncrna-region --skip-crispr --skip-sorf --skip-gap --keep-contig-headers. Output files, including .ffn, .faa and .gff3, were used for subsequent analyses.

Determination of IS flanking regions and boundaries

To investigate the sequence features of high-copy IS elements in *E. faecium* we used custom scripts to identify the IS ends, yielding transposase flanking and boundary regions. From the analysis described in the preceding sections, we identified 17 IS transposases with at least one copy per genome on average. Of these, we exclude the three IS3 family sequences as they are annotated as two coding sequences (CDSs) (orfA/B). The resulting 14 sequences comprise 30,847 copies across the 593 *E. faecium* genomes surveyed. In brief, we extracted 500 bp of flanking sequence on either side of the transposase CDS. For a given transposase variant, as their CDS is identical, we expect the IS flanks to be highly conserved. However, as we are considering variants from many similar genomes, we must account for identity by descent where sequence context beyond the true IS boundary is repeated. Therefore,

we first deduplicate the flanking sequence. Then, per transposase variant, a multiple sequence alignment was generated for the set of deduplicated flanks with MAFFT[61] (v.7.526). After collapsing positions whose consensus is 'GAP' (representing low-frequency insertions), per base entropy was calculated as bits (log_2-scale) with scores approaching 0 representing highly conserved sites. The flank is determined as the (longest) region of continuous low entropy (<0.5), exempting short (≤3 bp) high-entropy segments representing SNVs or short indels. Sequence information for these IS, including their consensus flank lengths, is displayed in Extended Data Fig. 4c. Relevant source code can be found in our repository ('Code Availability').

Multiple sequence alignment of IS transposase ORFs

Transposase protein sequences from the most common IS elements in *E. faecium* (Fig. 1c) were aligned using MAFFT (v.7.526) with default parameters. This excludes IS3 elements as they are not encoded as single ORFs. A maximum-likelihood phylogeny was inferred with IQ-TREE2 (ref. 62) (v.2.2.0) under the VT+F+G4 substitution model, selected according to the Bayesian information criterion, with 1,000 ultrafast bootstrap replicates. Branch lengths represent the expected number of amino acid substitutions per site. The resulting tree was visualized with the R package ggtree (v.3.10.1).

Frequency and distribution of IS transposase variants in *E. faecium*

For the identification of unique IS transposase ORF variants and quantification of their frequencies in *E. faecium*, we analysed the 593 *E. faecium* genomes from NCBI complete genomes and SHC *Enterococcus* isolates. We excluded the NCBI long-read assemblies from this analysis because only 5 out of 174 datasets (Extended Data Fig. 1) were sequenced with the more recent R10.4.1 nanopore chemistry. Earlier-generation ONT data are known to exhibit higher rates of homopolymer errors[63], which do not notably hamper hidden-Markov-model-based approaches such as ISEScan but would complicate accurate counting of exact copies of transposase nucleotide sequences.

Genome annotations were performed using Bakta, and the resulting GFF files were used to extract all CDS features longer than 500 nucleotides and labelled as 'transposase'. These were then consolidated into a non-redundant transposase database, enabling us to quantify the frequency of each unique transposase nucleotide sequence across the dataset. We identified the six most abundant IS transposase families—ISL3, IS30, IS256, IS6, IS3 and IS110—and visualized their distribution (Extended Data Fig. 4d). Relevant source code can be found in our repository ('Code Availability').

geNomad prediction of plasmid contigs

To predict whether a contig is probably from a chromosome or plasmid, we used geNomad[64] (v.1.8.0, v.1.6 DB) with the end-to-end workflow. Contigs predicted as chromosomal with a score of >0.6 were included in analyses as chromosomes. Contigs predicted as a plasmid with a score of >0.7 were included in analyses as plasmid. Contigs not reaching those thresholds were considered undetermined and excluded from analyses where plasmid predictions were necessary (Fig. 1c and Extended Data Fig. 3). Information about the number of contigs reaching these thresholds for each genome from the NCBI complete genomes and NCBI long-read assemblies are included in Supplementary Table 2 and are visualized in Extended Data Fig. 3i.

IS counts among ESKAPEE lineages

To control for the possibility that certain lineages drive species-level results in Fig. 1b, IS counts were measured among ESKAPEE taxa lineages. All pairwise genomic distances were calculated within each ESKAPEE taxon using Mash[65] (v.2.3) with 100,000 (-s), length 21 *k*-mer (-k) sketches. Genomes were then grouped into clusters defined by <0.01 pairwise distance across all members (Extended Data Fig. 2).

IS expansion in enterococci

Genomes from the high-contiguity *Enterococcus* collection and the 32 reference genomes described by Lebreton and colleagues[29] (2,134 in total) were classified phylogenetically with de_novo_wf from GTDB-Tk v.2.4.0, Release 220. Following Lebreton and colleagues[29], the term g__Vagococcus was used to define the outgroup. A table summarizing the IS counts as defined by ISEScan for each genome, as well as the Newick and tree.dist files output by de_novo_wf were processed with custom scripts. In addition to four genomes being excluded by GTDB-Tk, genomes were assigned to a representative by the minimum phylogenetic distance. A total of 18 genomes were excluded for having a minimum distance >0.01, yielding 2,112 assigned genomes. IS counts per family were summarized for each set of genomes assigned to a given representative. In total, 785 genomes were assigned to the representatives *EnGen0015*, *Aus0004* and *EnGen0043*, defined as *E. lactis*, clade A1 (healthcare-associated) and clade A2, respectively. To visualize counts per IS family across enterococci (Fig. 2b), the mean count of each IS family for the 32 sets of genomes was calculated and a file containing these values was uploaded to iToL[66]. For better visualization of *Enterococcus*, *Lactococcus garvieae* was excluded from the tree. Relevant source code and data can be found in our repository ('Code Availability').

ISL3 counts across *E. faecium* lineages

To investigate whether certain lineages of *E. faecium* were enriched for ISL3 copy, ISL3 copy was associated with a phylogenetic tree of all *E. faecium* genomes from the high-contiguity *Enterococcus* collection. In brief, a pangenome was generated with panaroo[67] (v.1.5.2, --core_threshold 0.99). The resulting core genome alignment was input into iqtree2 (v.2.2.0) to build a phylogenetic tree using a GTR+G substitution model and 1,000 bootstraps. The maximum-likelihood tree was visualized in iToL. The tree was midpoint rooted between *E. lactis* and clades A1 and A2. For each of the 785 genomes, ISL3 copy number as measured previously by ISEScan was uploaded to iToL. Furthermore, we assigned STs with multi-locus sequence typing (MLST)[68] (v.2.23.0; available at https://github.com/tseemann/mlst) to each genome using the scheme efaecium. Clade information was uploaded to iToL. In addition to *E. lactis* and clade A2 *E. faecium*, common STs (17, 18, 80, 117, 203 and 1,421) were also visualized.

Correlation of antibiotic resistance and IS copy in *E. faecium*

To predict antibiotic resistance genes (ARGs), AMRFinder[69] (v.4.0.23, DB 2025-07-16.1) was run on 785 *E. faecium* genomes from the high-contiguity *Enterococcus* collection with additional parameters -plus and -organism '*Enterococcus faecium*'. Hits were deduplicated by selecting the highest identity per ORF coordinate. Then, unique ARGs per genome were counted. ARGs were then related to IS counts per genome; 65 genomes predicted to be *E. lactis* as defined above in 'IS expansion in enterococci' were considered a single 'Non-clade A1/A2' category. The remaining 720 predicted clade A1/A2 genomes were converted into a categorical variable on the basis of IS copy, for each of the six most common IS families as in Extended Data Fig. 4d. The genomes were separated into quintiles, with the lowest quintile as Q1 (fewest IS) and the highest quintile as Q5 (most IS).

Defining human-infectious NFF *Enterococcus*

NFF *Enterococcus* causing human infection were based on the Australian *Enterococcus* Surveillance Outcome Program[30]. During the 10 years from 2013 to 2022, 11,681 *Enterococcal* bacteremia isolates were collected and identified. Any *Enterococcus* species responsible for at least ten bacteremia isolates (average of one per year) in that collection is considered here as human-infectious. These species are *E. faecalis* (6,319), *E. faecium* (4,712), *E. casseliflavus* (198), *E. gallinarum*

(166), *E. avium* (107), *E. raffinosus* (60), *E. hirae* (45), *E. lactis* (29) and *E. durans* (27).

NCBI short-read *E. faecium* assemblies: curation of isolate datasets for ISL3 expansion timeline

A list of accessions was obtained by querying the NCBI e-utilities application programming interface, resulting in a set of 21,376 Illumina-sequenced *E. faecium* genomic whole-genome sequencing data with available collection year. This set of accessions was filtered to paired-end data with at least around 20× estimated coverage and then downsampled to ≤500 samples per year, resulting in a final set of 6,747 samples. Next, we ran custom Nextflow[70] pipelines to process the samples. In brief, raw reads were downloaded, deduplicated, trimmed, coverage-filtered/downsampled (20–150×) and assembled, and then FastANI[71] (versus clade A1 [Aus0004], clade A2 [EnGen0015] and *E. lactis* [EnGen0043] references) and CheckM2 (ref. 72) were run on each assembly. Finally, assemblies were filtered: (1) FastANI closest reference = A1, min_dist ≤ 0.015 and ≥97% average nucleotide identity (ANI); (2) genome size within $\mu \pm 2$ s.d.; (3) CheckM2 ≥ 99% completeness and ≤5% contamination; (4) contig N50 ≥ 25 kb. This resulted in 5,422 final short-read assemblies; we refer to this dataset in the manuscript as NCBI short-read *E. faecium* assemblies. Relevant source code and data can be found in our repository ('Code Availability').

NCBI short-read *S. epidermidis* assemblies: curation of isolate datasets for IS expansion timeline

As with *E. faecium*, a list of accessions was obtained querying the NCBI e-utilities application programming interface, resulting in the set of 5,938 Illumina-sequenced *S. epidermidis* genomic whole-genome sequencing data with available collection year. To enrich for hospital-adapted isolates and avoid the potential confounding influence of skin commensals, we excluded isolates with sampling metadata indicating superficial body sites (for example 'skin', 'scalp', 'hand', 'ear'). We also required ≥99% ANI to the *S. epidermidis* RefSeq reference (GCF_006094375.1) and applied additional quality control thresholds analogous to those in our *E. faecium* analysis. This resulted in 1,604 final short-read assemblies after applying all filtering steps. Relevant source code and data can be found in our repository ('Code Availability').

Coverage-based IS count estimates in short-read data

To quantify IS abundance in these short-read assemblies, we first annotated IS present on individual contigs with ISEScan. Contigs shorter than (600 × number of predicted IS on the contig) were not counted to ensure contigs that are putatively only part of an IS are not counted. This limits the inclusion of noisy coverage data from very short contigs, as well as the possibility of double counting fragmented IS elements. This probably also leads to conservative estimates of IS count, which we prefer to the alternative. For each assembly, chromosomal coverage was approximated as the length-weighted median coverage of contigs exceeding 5 kb in length. We then calculated a relative coverage for each contig by dividing the contig coverage by the chromosomal coverage. The number of IS copies on each contig was multiplied by its estimated coverage, and these values were summed subsequently across all included contigs to obtain the final per family IS count per assembly.

To ensure our IS counting heuristic from fragmented assemblies was appropriate, especially to avoid differential bias towards certain IS families, we benchmarked our approach against highly contiguous genome IS counts. We generated paired short- and long-read data for 117 samples (37 *E. faecium*; 80 *E. faecalis*) from our SHC collection and compared IS counts per family and per species (Fig. 3b and Extended Data Fig. 5). These results show that our IS counting heuristic is reasonably accurate across IS families and that this is not specific to *E. faecium*. The results also highlight the following caveats of the heuristic: (1) it tends towards being slightly conservative (due partially to ignoring short noisy contigs); (2) it is less precise, although not less accurate, for higher-copy IS families and (3) it does not perform as well for IS6, which tends to be plasmid-borne (Fig. 3b). This result is not surprising because of the greater difficulty in resolving repetitive/plasmid sequences in short-read data. As a result, we do not include analyses of IS6 elements from short-read-based assemblies.

To control for the possibility changing IS abundance over time could be explained by recent isolates being disproportionately clinical in origin, whereas older isolates might derive from nonclinical sources, we evaluated several IS families. Healthcare-associated *E. faecium* strains were reported previously to be enriched for IS256, IS30 and IS3 family elements. Consequently, if this kind of sampling bias were driving our findings, we would expect all IS families associated with clinical strains to follow the same temporal pattern (that is, low in older and high in more recent samples).

Relevant source code and data can be found in our repository ('Code Availability').

Trendline of IS copy number over time

To discern whether there is a general trend in changing IS copy number over time (raw data shown in Extended Data Fig. 5), while limiting short-term fluctuations, we computed the average genomic IS copy number per year over a 5-year sliding window (95% confidence interval, mean ± t × s.e.), and plotted a trendline using locally weighted scatterplot smoothing (LOWESS) (smoothing fraction 0.25). Relevant source code can be found in our repository ('Code Availability').

PopPUNK clustering of *E. faecalis* and *E. faecium* genomes for PanGraph input

To characterize structural variation among a set of nearly clonal genomes, we returned to the SHC long-read *Enterococcus* isolates as described above. With PopPUNK[73] (v.2.6.7), we clustered 722 *E. faecium* genomes from the high-contiguity *Enterococcus* collection with the v.2 full *E. faecium* PopPUNK database downloaded from https://ftp.ebi.ac.uk/pub/databases/pp_dbs/Enterococcus_faecium_v2_full.tar.bz2. These genomes are the 720 clade A genomes as in Fig. 2a, with FM003 and FM057 included. They were excluded previously for coverage lower than our 40× threshold (39× and 31×). We chose to include them here because they are both single-contig chromosomal assemblies. Clustering was done with the poppunk_assign command with the additional flag of run-qc. PopPUNK's poppunk_visualize command was used to generate a Newick file, viz_core_NJ.nwk, describing the resulting phylogeny. The output tree was visualized with iToL. We assigned sequence types with MLST (v.2.23.0) to each genome using the scheme efaecium. The SHC hospital-endemic lineage corresponds to ST117 and formed a single cluster. The 66 genomes in the cluster, all from the SHC collection, were selected for further analysis. Similarly, the 1,083 *E. faecalis* genomes from NCBI complete and SHC long-read *Enterococcus* isolates were assigned to the v.2 full *E. faecalis* PopPUNK database downloaded from https://ftp.ebi.ac.uk/pub/databases/pp_dbs/Enterococcus_faecalis_v2_full.tar.bz2. None were removed by quality control. We attempted to assign STs as above using the scheme efaecalis. Here the largest cluster corresponding to ST179 contained 57 genomes, 44 of which were SHC isolates. The 44 genomes from the SHC collection were selected for further analysis.

Filtering genomes and stitching contigs for input into PanGraph

For each of the SHC genomes in those clusters, we extracted the contigs predicted to be from the chromosome by geNomad as described before. As PanGraph[74] requires single-contig inputs, we filtered for genomes with five or fewer contigs, yielding 66 *E. faecium* and 44 *E. faecalis* genomes. Multi-contig genomes were stitched together using 2 kb 'N' spacers in the order output by Flye.

PanGraph analysis of structural variation in the SHC *E. faecium* ST117 and *E. faecalis* ST179 clusters

For each of the two clusters defined above, we adapted the 'Structural genome evolution' workflow as defined by Molari and colleagues[39] and found at https://github.com/mmolari/structural-evo.

In brief, the genomes were processed with PanGraph (v.0.7.1) to build a pangenome graph, which defines all 'core blocks' (regions of shared homology across all genomes). Regions of structural variation are captured as edges between consecutive core blocks, and the set of sequences that can occur between two core blocks is defined as a 'junction'. PanGraph was rerun on each junction (with its flanking core blocks), and the resulting 'subgraph' defines each unique arrangement of homologous blocks through the junction as a unique 'junction path'. Thus, for a given junction, the set of unique junction paths neatly describes the extent of structural sequence variation occurring in the cluster for a given structurally variant locus.

Next, the set of junction subgraphs was analysed with a custom Python workflow: first, we confirmed that the spacers (2 kb 'N' as defined above) always occurred fully in the accessory region of junctions. This was expected, as disruptions in assembly contiguity are probably due to structural variation within the sample population itself. We removed any spacer-containing path from downstream analysis as their true sequence length and genotype is unclear. Next, for each junction, the workflow enumerates the distinct paths within the subgraph and characterizes the accessory genome of each non-empty path, using the Bakta-derived annotations (described above) for each source genome (Supplementary Table 4). In ST117, the most common type of junction is a 'binary' IS insertion. In that case, the junction comprises two paths, in which one path is empty (that is, there is no accessory region between the core blocks), and the other path contains only an IS element transposase annotation. IS-only junctions containing more than two paths and/or several IS families also occurred. Fewer large and more complex junctions occurred as well, including large indels and phages, although we did not characterize these in detail. Relevant source code and data can be found in our repository ('Code Availability').

Mobile genetic elements have been reported previously as important to *E. faecalis* virulence. The analysis of *E. faecalis* ST179 served as a control for rates of structural variation over the same period in a related organism with similar hospital epidemiology and taxonomic relatedness. The *E. faecalis* samples were collected from the same environment, location and time frame, and were processed and analysed in an identical manner (Supplementary Table 4).

For compatibility with our high-performance computing cluster, the PanGraph workflow was adapted to work for Snakemake (v.8.18.2). The workflow was run per cluster with existing installations of ISEScan (v.1.7.2.3), DefenseFinder[75] (v.1.3.0), geNomad (v.1.8.0) and IntegronFinder[76] (v.2.0.5).

Generation of circos plots of IS variants in PanGraph clusters

The isolate with earliest collection date, EF006 for the ST179 *E. faecalis* cluster and FM001 for the ST117 *E. faecium* cluster, was chosen as the reference coordinate system. The set of IS-only junctions was filtered to 'backbone-only' junctions. In other words, IS insertions occurring in junctions not shared by all genomes (for example, within other junctions or between core blocks that are inverted for some genomes) were not shown, as their position on the core backbone is ambiguous. Each junction was given a fixed width of 3 kb, as smaller sizes were not visible on the plot. Relevant source code and data can be found in our repository ('Code Availability').

ISL3 gene disruption and intergenic distance analysis in ST117

The *E. faecium* ST117 PanGraph analysis defined 100 ISL3 junctions (that is, IS-only junctions containing only ISL3 family transposase annotation(s)) (Fig. 4c). Inspection of these junctions revealed that the vast majority are indeed parsimoniously explained by insertions of one or more ISL3 elements in the region. To identify candidates for intra-genic disruptions, we considered junctions where an annotated gene in the 'empty' path (lacking the ISL3 insertion) overlapped fully or partially with the accessory region of a 'filled' path. A junction was counted as gene-disruptive only if it contained such an overlapping gene and if that gene's annotation was also missing or truncated in the filled path (Fig. 4h). The remaining ISL3 junctions were presumed to be intergenic. Of these, we considered every filled (ISL3-containing) path for every junction and extracted the distance from the 3′ end of the transposase to the next nearest gene, as well as its relative orientation. Paths that contained several accessory ISL3 transposases were counted only if they were all oriented the same way, otherwise the orientation was considered ambiguous and the path was excluded from analysis (Fig. 4i). As a control, the same analyses were done on the 51 IS30, IS256, IS110, IS3 and IS6 junctions as a combined set ('Other' in Fig. 4h,i).

E. faecium domination longitudinal sample curation

To identify a set of metagenomic samples that were (1) clinically relevant, (2) longitudinal and (3) technically suitable for assembling complete, closed *E. faecium* genomes directly, we selected longitudinal cases from our laboratory's biobank of stool samples from people undergoing HCT with high relative abundance of *E. faecium*. As previously described[44,45], these stool samples were collected under protocols approved by the Stanford University Institutional Review Board (Protocol nos. 8903; principal investigator: D. Miklos, co-investigator: A. Bhatt and T. Andermann, Committee IRB-3 and 42053; principal investigator: A. Bhatt, Committee IRB-8). Informed consent was obtained from all human research participants. On the basis of previous metagenomic sequencing results, samples with at least 10% relative abundance of *E. faecium* were considered. Any patients with at least two such samples were then identified and those samples 'dominated' (defined as ≥10% relative abundance) by *E. faecium* were selected for long-read metagenomic sequencing, yielding 34 samples from 14 patients (range, two to five samples per patient).

Metagenomic high molecular weight DNA extraction

To extract high molecular weight DNA from stool samples in preparation for metagenomic long-read sequencing, we used the QIAamp PowerFecal Pro DNA Kit (Qiagen, catalogue no. 51804). In brief, working in a biosafety hood, from each stool sample kept on dry ice, roughly 200 mg stool (two to three biopsy punches) were added to the bead beating tube. Bead beating tubes were added in sets of 12 to a Vortex-Genie 2 with the horizontal Vortex Adapter described in the extraction kit manufacturer's instructions and vortexed at maximum speed for 12 min at room temperature. All subsequent steps were consistent with manufacturer instructions using centrifugation. The final DNA was eluted in 100 µl C6 buffer and incubated at room temperature for 15 min before centrifugation. DNA concentration was measured by NanoDrop. DNA was then cleaned up with the Genomic DNA Clean & Concentrator-10 kit (Zymo Research, catalogue no. D4011) according to manufacturer instructions for large genomic DNA. In brief, 200 µl of DNA binding buffer was added to 100 µl of DNA eluate from before. DNA was eluted in 25 µl nuclease-free water. After clean-up, by Nanodrop, DNA concentrations ranged from 7.5 ng μl^{-1} to 768.7 ng μl^{-1} per sample. Two samples had both DNA concentrations below 25 ng μl^{-1} and purity values (OD_{260}/OD_{230}) less than 1.5 and were not sequenced. DNA quality was confirmed with Genomic DNA ScreenTape Analysis (Agilent, catalogue nos. 5067-5365 and 5067-5366) on the Agilent 4150 TapeStation system. Samples had peak concentrations at DNA lengths greater than 20 kb. The loss of one sample, p11580_03 led to a singleton, but we continued with sequencing of p11580_06. Therefore 32 samples from 14 patients (range 1 to 4) were sequenced as described below.

SHC long-read metagenomes: ONT sequencing, basecalling, quality control and assembly

Two pooled libraries containing 16 samples each were prepared with the Oxford Nanopore Native Barcoding 96-kit (catalogue no. SQK-NBD114.96) according to manufacturer's instructions. In brief, 400 ng of input DNA per sample was prepared, diluting the sample in nuclease-free water if necessary. DNA concentration was confirmed for each sample with the Qubit DNA High Sensitivity kit at each checkpoint. On the basis of previous use of this approach for metagenomic sequencing of stool samples in our laboratory, and on TapeStation results, we expected our fragment length to be between 7 kb and 10 kb and therefore loaded 50 fmol of each library. Libraries were loaded onto separate R10.4.1 PromethION flow cells (FLO-PRO114M) according to the manufacturer's instructions and sequenced simultaneously on a PromethION 2 solo instrument. We anticipated assembling complete *E. faecium* genomes with at least 40× coverage. Therefore, we estimated that to reach 40× coverage of a 3-Mb *E. faecium* genome making up 10% relative abundance of the sample, we would need 1.2 Gb of sequencing per sample. To increase the likelihood of reaching this threshold for each sample and accounting for library imbalance, we let sequencing proceed until both libraries reached at least 2.5 Gb per sample of total sequencing. We obtained 60 Gb and 40 Gb of sequencing for each library, respectively.

Raw.pod5 files were transferred to our computing cluster and basecalling was performed with Oxford Nanopore's open-source Dorado basecaller (v.0.7.3) available at https://github.com/nanoporetech/dorado. The following command was used: dorado basecaller −kit-name SQK-NBD114-96 -b 1024 −trim 'all' sup@v5.0.0 {pod5} > {basecalled}.bam. Output bam files were demultiplexed and emitted as.fastq files with: dorado basecaller −kit-name SQK-NBD114-96 -b 1024 −trim 'demux'. Of 32 sample barcodes, 31 had at least 1 Gb of sequencing after demultiplexing. One barcode corresponding to p11569_01 had less than 10 Mb of sequencing and was not considered further. Unfortunately, this made p11569_02 a singleton. Ultimately, 31 samples from 14 patients (range 1 to 4) were included for metagenomic assembly (Supplementary Table 5).

NanoPlot[77] (v.1.41.6) was used to evaluate sequencing quality per sample with default parameters. All 31 samples had median read quality of at least Q22.1 (22.1–24.2). Two samples, p11463_06 and p11580_06 had 9,150 and 11,220 reads, respectively. Other samples had at least 56,643 reads (56,643–659,377). The minimum read N50 for these 29 samples was 7.77 kb. Metagenomic assembly was performed with Flye[78] (v.2.9.2) on all 31 sequenced samples using the following command: flye --nano-hq {reads} −meta −genome-size 200 m −keep-haplotypes −no-alt-contigs (Supplementary Table 5). We refer to this dataset in the manuscript as the SHC long-read metagenomes.

Taxonomic profiling of assembled contigs with sourmash

To identify *E. faecium* contigs from the metagenomic assemblies, all contigs 5 kb or longer were profiled taxonomically with sourmash[79] (v.4.8.11) using the successive commands: sourmash sketch dna -p k = 51,abund −singleton {sample1}.fa and sourmash scripts fastmultigather -k 51 -m DNA −threshold-bp 5000 {samples}.zip gtdb-rs214-reps.k51.zip. All contigs predicted as *g_Enterococcus;s_faecium* longer than 2.5 Mb, and with at least 80× coverage were considered 'complete' chromosomal contigs. One such contig was identified for 28 of 31 (p11463_06, p11569_02, p11580_06 did not) successfully sequenced samples as described above, which resulted in 12 longitudinal cases (Extended Data Fig. 8).

Split *k*-mer analysis of *E. faecium* assemblies

To determine cases of 'isogenic' *E. faecium* chromosomes such that within strain heterogeneity and structural variation could be assessed clearly, the relatedness of the 28 longitudinal *E. faecium* chromosomal contigs was considered with SKA[80] (v.1.0). All pairwise distances and clusters of highly related contigs were identified using the following commands: first, ska alleles -f inputs.txt and then, ska distance -f skfs.txt -o results -S -s 40, where inputs.txt contained paths to the 28 *E. faecium* chromosomal contig fasta files. Patients with longitudinal samples with fewer than ten SNPs per megabase or higher than 99.999% identity over time were considered for longitudinal within-patient analyses. This threshold excluded p11342.

Within-sample and within-patient structural variant calling

To assess within-sample *E. faecium* population heterogeneity, reads from each sample were aligned to their own metagenome assembly using ngmlr[81] (v.0.2.7) and default ONT sequencing presets. Structural variants were called with Sniffles2 (ref. 82) (v.2.6.0), which was run in both normal and mosaic modes to capture both dominant and subclonal variants. The outputs from these two modes were combined and any duplicate variants were removed, as well as any variants mapping to contigs other than the *E. faecium* chromosome and variants smaller than 100 nucleotides. The resulting per sample structural variant counts are shown in Extended Data Fig. 9f (Supplementary Table 6).

To determine whether ISL3 variants persisted or became fixed across sequential isolates from the same patient, we performed longitudinal variant calling: for each patient, the *E. faecium* chromosome from the sample collected at the earliest time point was chosen as the reference. Reads from each time point were mapped back to the reference metagenome assembly and passed as input to Sniffles2 as above (Supplementary Table 6). To look for structural variants caused by IS element transpositions, all identified structural variants mapping to the *E. faecium* chromosome were screened to include only insertions or deletions in the 750–3,000 bp size range, as this corresponds approximately to the length of IS elements. For each candidate variant, the predicted sequence was extracted and analysed with ISEScan to confirm the presence of IS elements. The resulting IS variant calls were compiled, and patterns of persistence or fixation were examined by comparing the frequencies and presence or absence of these IS-containing variants across each patient's time points.

p11568 ISL3 structural variant neighbourhood coverage

To measure the changing population frequency of the two identified ISL3 structural variants in the four sequenced p11568 samples over time, ONT reads from each of p11568 t1–t4 were aligned against the t4 *E. faecium* contig with minimap2 (ref. 83) with the flags -x asm20, -c and --eqx. Per base read depth was then counted for each position with samtools[84] (v.1.21) using the command: samtools depth -a -Q 10 -b t4.bed {sample}_sorted.bam > {sample}.coverage.txt. Poor quality alignments are removed by the -Q 10 flag. First, the mean alignment depth across the entire t4 *E. faecium* contig for each sample was calculated. Then, relative depth per base was calculated by dividing the depth at each position by the mean depth per sample. The t4 *E. faecium* contig was annotated with Bakta as described above and the neighbourhood of each ISL3 structural variant was visualized.

E. faecium isolation from stool

A single punch biopsy of stool samples from the first (t1) and last (t4) time points from p11568 were suspended in 300 µl PBS. Each sample was serially diluted tenfold into 900 µl PBS to 1:10,000. Then, 30 µl of the 1:10,000 dilutions were plated onto BHI agar and incubated overnight at 37 °C. The next day, six colonies from each plate were selected and identified by biotyping with a MALDI-TOF based Bruker Biotyper per manufacturer instructions. Each identified *E. faecium* colony was streaked to isolation on a BHI agar plate. To confirm the insertion genotypes, overnights of isolates were prepared in 3 ml BHI broth and incubated with shaking at 37 °C. Overnights were diluted 1:1,000 into

10-ml subcultures and grown to an optical density of approximately 1.0. Cell pellets were then prepared according to Plasmidsaurus 'Standard Bacterial Genome with DNA extraction' specifications. The resulting genomes were annotated with Bakta as described above and the lack of both ISL3 insertions in t1, and the presence of both ISL3 insertions in t4 were confirmed in the assemblies.

RNA-seq of *E. faecium* isolates

From glycerol stocks of both p11568 t1 and t4 confirmed to have the insertional genotypes (Fig. 5), four overnight cultures were prepared in BHI as before. Each overnight culture was diluted 1:100 into 5 ml of BHI and subcultured for around 3 h until OD_{600} reached 0.4; 5-ml microtubes (AXYGEN, catalogue no. MCT-500-C) were prepared with 0.5 ml of quenching solution. The quenching solution contained 90% (v/v) 100% ethanol and 10% (v/v) saturated phenol for RNA extraction (pH 4–5). Microtubes were stored at −20 °C before quenching occurred. After OD_{600} reached 0.4, 4 ml of each culture was promptly quenched. The microtube was vortexed until the solution was well mixed. All microtubes were centrifuged at 10,000*g* for 5 min at 4 °C. The supernatant was poured off. The pellet was then lysed in 260 µl of a prepared mastermix containing 25:1 PBS:10 mg ml^{-1} lysozyme (Sigma-Aldrich, catalogue no. L3790-10X1ML). Each pellet was resuspended in lysis buffer and then incubated at 37 °C for 30 min while rocking on a BenchRocker. Then, 30 µl of 20% SDS was added followed by another round of incubation at 37 °C. Then, 1.5 ml of Trizol was added and mixed well by pipette. All subsequent steps were performed at room temperature. The solution was incubated for 10 min. Then, 0.5 ml of chloroform was added and ten vigorous complete inversions of the tube were made before 3 min of incubation. After incubation, each tube was centrifuged at 12,000*g* for 10 min. Carefully avoiding any non-aqueous layer, approximately 800 µl of the aqueous phase was transferred to a clean 2-ml Eppendorf tube. An equal volume of 100% ethanol was added and mixed by pipette until emulsions were no longer visible. RNA was then purified using the Zymo RNA Clean & Concentrator-25 kit (catalogue no. R1018) according to manufacturer instructions. RNA was eluted in 25 µl nuclease-free water. RNA concentration was measured by NanoDrop and quality was confirmed by Bioanalyzer with the RNA 6000 Pico Kit (Agilent; catalogue no. 5067-1513). RNA was then stored at −80 °C and sequenced with Novogene's Prokaryotic Total RNA sequencing service (NovaSeq PE150, 3G raw data per sample).

Analyses of RNA-seq data

Raw RNA-seq reads were trimmed with Trim Galore, available at https://github.com/FelixKrueger/TrimGalore/, using the command trim_galore --quality 30 --length 60 --paired {sample}1.fq.gz {sample}_2.fq.gz --retain_unpaired. Bowtie2 (ref. 85) (v.2.5.4) was used to align processed reads against the p11568 t4 isolate reference genome with default parameters. On the basis of a poor alignment fraction (20.99%) compared with the other replicates (minimum: 97.63%), the fourth replicate from the t1 isolate (t1_4) was excluded. To get the per base depth of reads, samtools depth was used as before except no quality threshold was implemented. This meant reads mapping equally well to several locations were distributed randomly. Furthermore, a feature count matrix filtering alignments against the t4 isolate reference genome with mapping quality < 10 was generated with bedtools (v.2.27.1) multicov using the command bedtools multicov -p -q 10 -bams {sample1}_sorted.bam {sample2}_sorted.bam. The resulting feature count matrix, along with the Bakta annotation of the t4 isolate genome, were used to perform differential expression analysis with the DESeq2 (ref. 86) package (v.1.42.1) in R. Only features with at least one count in both t1 ($n = 3$) and t4 ($n = 4$) were considered for differential expression. The adjusted P value and $\log_2$(fold change) for each of the 2,668 resulting features were visualized as a volcano plot. Thresholds of 1×10^{-25} and $|\log_2(\text{fold change})| \geq 2$ were used to define differentially expressed genes.

E. faecium growth and antibiotic dose–response assays

E. faecium isolates were grown as described above. Antibiotic susceptibility testing was performed as described previously[46], with some modifications. In brief, each strain was grown overnight (approximately 18 h) to stationary phase in BHI medium at 37 °C. Overnight cultures were then diluted to $OD_{600} = 0.001$ in Mueller-Hinton broth (Sigma-Aldrich) and incubated with twofold serial dilutions of TMP (Sigma-Aldrich) and SMX (Fisher) at the indicated concentrations. A ratio of 1:20 (TMP:SMX) was maintained for all treatments. Folic acid (Fisher) and folinic acid (Sigma-Aldrich) were supplemented at 1 µg ml^{-1} where indicated. Cultures were incubated at 37 °C with continuous shaking in technical duplicate (150 µl volume per well) in a flat bottom 96-well microtiter plate (Corning) and sealed with a Breathe-Easy membrane (Diversified Biotech). Absorbance (600 nm) was monitored every 10 min for 24 h using an Agilent BioTek Synergy H1 microplate reader. Growth was calculated as the percent area under the curve (AUC) relative to a control with no antibiotics added using the R package gcplyr[87].

Data visualization

Initial plots were generated by custom Python (Plotly) and R (ggplot) code. Plots were organized and refined with AffinityDesigner v.2.5.7.

Scripting

Custom bash, Python, R and Nextflow scripts referenced in the methods section are available in the GitHub repository ('Code Availability'). All scripts were run in R v.4.3.2. Python scripts were run using v.3.10 with polars v.1.20.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Long-read metagenomic sequencing data and RNA-seq data generated in this study are available under NCBI BioProject ID PRJNA1236482. Long-read assemblies from *Enterococcus* isolates sequenced in this study are available at Zenodo (https://doi.org/10.5281/zenodo.15239062)[88]. There, we also include some pre-processed and intermediate source data tables and other files to facilitate reproducibility of custom analyses where appropriate, as well as figure source data files for the original phylogenetic trees, and SRA accession codes for publicly available datasets we used in this study. Accession codes for long-read pathogen isolate datasets from SRA (NCBI long-read assemblies) and for NCBI complete genomes that were used in the IS accounting analysis are also reported in Supplementary Table 2. Assemblies annotated using Bakta[60] (v.1.9) used the Bakta database v.5.1, which is available at Zenodo (https://zenodo.org/records/10522951)[89]. To identify chromosomal contigs, the geNomad[64] (v.1.8) database v.1.6, which is available at Zenodo (https://doi.org/10.5281/zenodo.6994741)[90], was used. GTDB-Tk[57] (v.2.4) release 220 used for phylogenetic analyses is available at https://data.gtdb.aau.ecogenomic.org/releases/release220/. AMRFinder[69] (v.4.0.23) db 2025-07.16.1 used to measure ARGs is available at https://ftp.ncbi.nlm.nih.gov/pathogen/Antimicrobial_resistance/AMRFinderPlus/database/4.0/. PopPUNK[73] (v.2.6.7) v.2 *E. faecium* and *E. faecalis* databases used for clustering are available at https://www.bacpop.org/poppunk-databases/. Source data are provided with this paper.

Code availability

Source code and custom workflows used for analyses in this manuscript are available on GitHub at https://github.com/abehr/IS-evolution, along with a guide explaining how to set up and run the analyses. An archived

snapshot of our codebase can also be found in our project available at Zenodo (https://doi.org/10.5281/zenodo.15239062)[88].

56. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
57. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
58. O'Leary, N. A. et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data* **11**, 732 (2024).
59. Xie, Z. & Tang, H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* **33**, 3340–3347 (2017).
60. Schwengers, O. et al. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb. Genomics* **7**, 000685 (2021).
61. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
62. Minh, B. Q. et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
63. Sereika, M. et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods* **19**, 823–826 (2022).
64. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2024).
65. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
66. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).
67. Tonkin-Hill, G. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
68. Jolley, K. A. & Maiden, M. C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinf.* **11**, 595 (2010).
69. Feldgarden, M. et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.* **11**, 12728 (2021).
70. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
71. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90 K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
72. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
73. Lees, J. A. et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29**, 304–316 (2019).
74. Noll, N., Molari, M., Shaw, L. P. & Neher, R. A. PanGraph: scalable bacterial pan-genome graph construction. *Microb. Genomics* **9**, mgen001034 (2023).
75. Tesson, F. et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.* **13**, 2561 (2022).
76. Néron, B. et al. IntegronFinder 2.0: Identification and analysis of integrons across bacteria, with a focus on antibiotic resistance in *Klebsiella*. *Microorganisms* **10**, 700 (2022).
77. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).
78. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
79. Irber, L. et al. sourmash v4: A multitool to quickly search, compare, and analyze genomic and metagenomic data sets. *J. Open Source Softw.* **9**, 6830 (2024).
80. Harris, S. R. SKA: Split Kmer Analysis toolkit for bacterial genomic epidemiology. Preprint at *bioRxiv* https://doi.org/10.1101/453142 (2018).
81. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
82. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).
83. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinforma. Oxf. Engl.* **37**, 4572–4574 (2021).
84. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
85. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
86. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
87. Blazanin, M. gcplyr: an R package for microbial growth curve data analysis. *BMC Bioinf.* **25**, 232 (2024).
88. Behr, A. & Grieshop, M. Data for *E. faecium* ISL3 project: long-read *Enterococcus* genome resource + preprocessed data for manuscript analyses. *Zenodo* https://doi.org/10.5281/zenodo.15239062 (2026).
89. Schwengers, O. Bakta database. *Zenodo* https://doi.org/10.5281/zenodo.4247252 (2025).
90. Camargo, A. geNomad database. *Zenodo* https://doi.org/10.5281/zenodo.6994741 (2025).

Acknowledgements We are grateful to the patients whose samples made this study possible. We thank T. Andermann and A. Rezvani for generating the HCT patient stool collection and providing clinical expertise regarding the longitudinally sampled HCT patients. We would like to thank S. Cuddihy, A. Moreno and members of the Stanford Clinical Microbiology Laboratory for their work isolating and banking *Enterococcus* from infection samples. We would like to thank Y. Han of M. Steinegger's group for work on IS structural similarity. We would also like to thank H. Tang for her expert statistical guidance in several aspects of this work. We are particularly grateful to J. Wirbel for his frequent counsel and support. We would like to acknowledge the contributions of several additional members of the Bhatt laboratory for their assistance with experiments and feedback during manuscript preparation, especially Y. Pinto, R. Chanin, S. Zlitni and D. Schmidtke. During preparation of this manuscript, we made limited use of large-language-model tools (for occasional prototyping of auxiliary code segments and refining the structure of a small number of sentences); all substantive analyses, interpretations and manuscript content remain the responsibility of the authors. This work used supercomputing resources provided by the Stanford Genetics Bioinformatics Service Center, supported by National Institutes of Health S10 Instrumentation Grant S10OD023452. The Bhatt laboratory is supported by National Institutes of Health R01AI148623 and R01AI143757, and a Stand Up 2 Cancer Grant. A.S.B. is supported by the Allen Distinguished Investigator Award. M.P.G. is supported by Stanford's Medical Scientist Training Program. A.A.B. is supported by an National Science Foundation Graduate Research Fellowship award. This research was also supported, in part, by National Science Foundation grant PHY-2309135 to the Kavli Institute for Theoretical Physics. This paper describes the views of the authors and does not necessarily represent the official views of the National Institutes of Health (USA).

Author contributions A.S.B., M.P.G. and A.A.B. conceived and designed the study. M.P.G., A.A.B., G.Z.M.R. and E.F.B. processed Stanford Health Care *Enterococcus* isolates collected by N.B., G.R.-N. and J.L.S. M.P.G. and A.A.B. designed, implemented and executed IS accounting workflows. M.P.G., A.A.B. and M.M. performed PanGraph structural variant analysis. M.P.G. and A.A.B. performed long-read sequencing of metagenomic samples curated by M.P.G. and B.D. A.A.B. and M.P.G. performed metagenomic assembly. M.P.G. identified *E. faecium* chromosomal contigs and performed clustering. A.A.B. conducted structural variant calling of *E. faecium* chromosomal contigs. M.P.G. and A.A.M. isolated *E. faecium* from p11568 stool. M.P.G., A.A.B. and S.B. performed RNA-seq experiments. J.D.L. and M.P.G. performed folate limitation experiments. A.A.B. led software development, data organization and repository preparation. Funding was acquired by A.S.B. The manuscript was drafted and all figures were prepared by M.P.G. and A.A.B. The manuscript was finalized by A.S.B., M.P.G. and A.A.B.

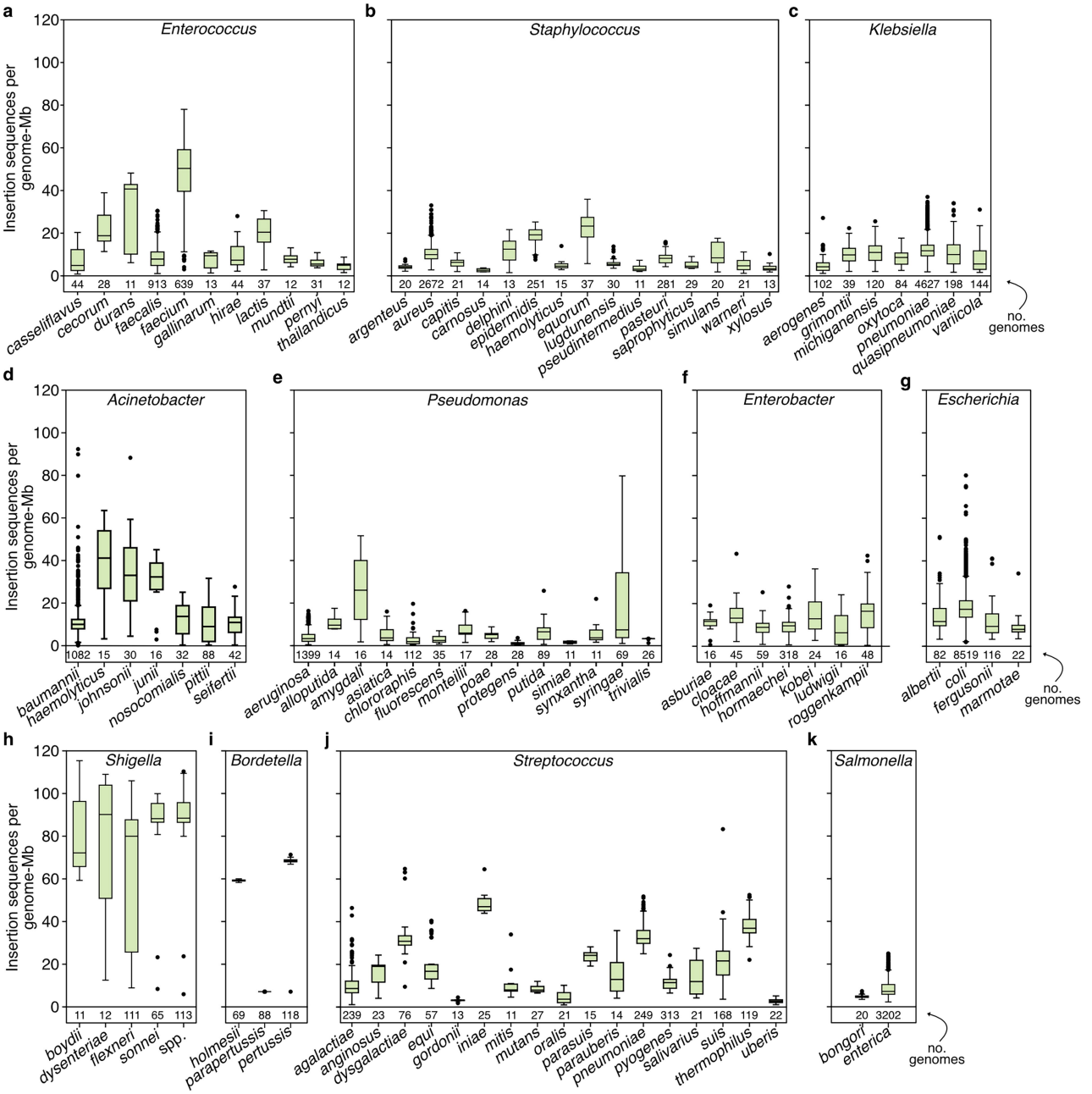
Competing interests The authors declare no competing interests.

Additional information

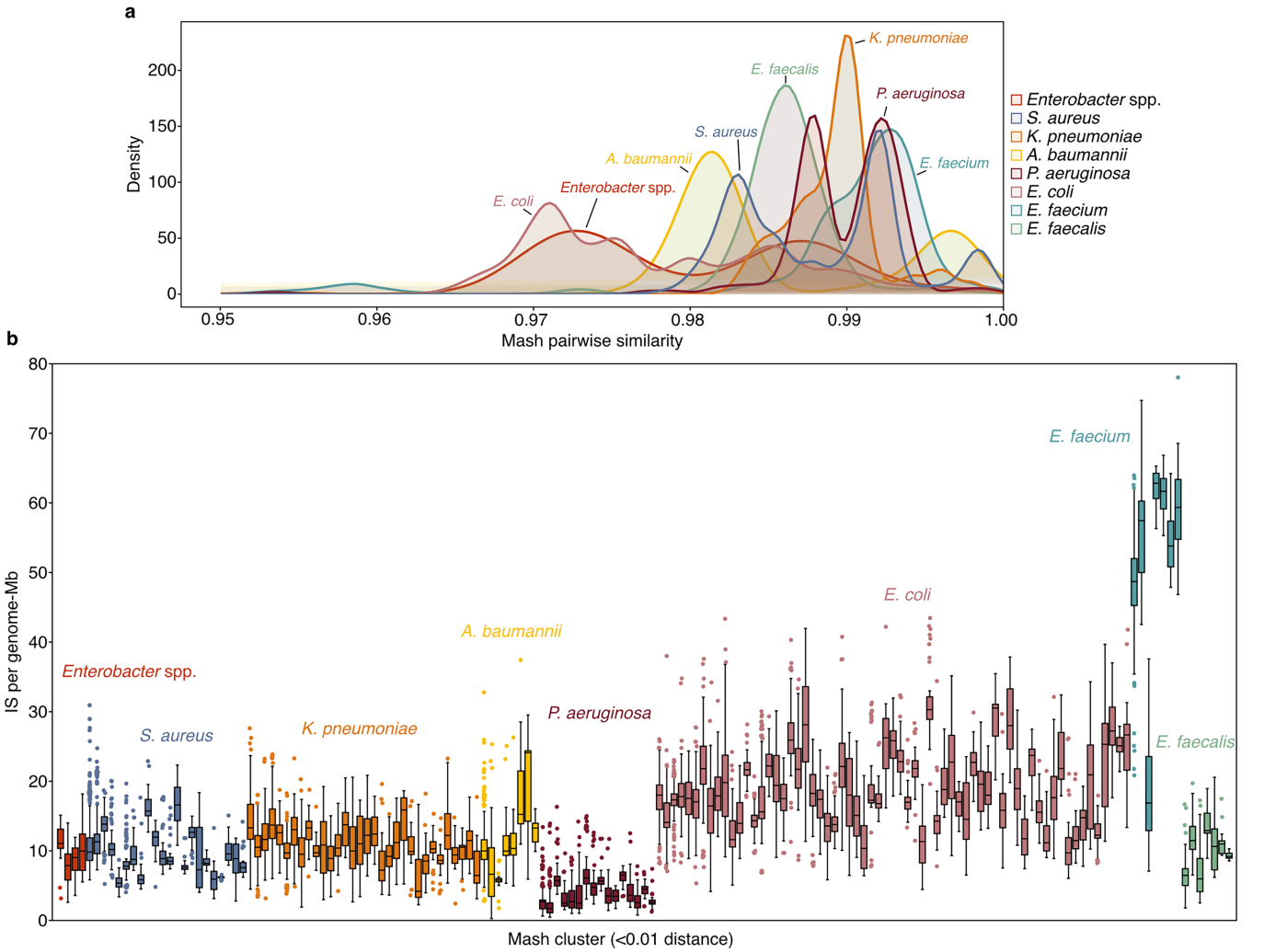
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10373-2.

Correspondence and requests for materials should be addressed to Ami S. Bhatt.

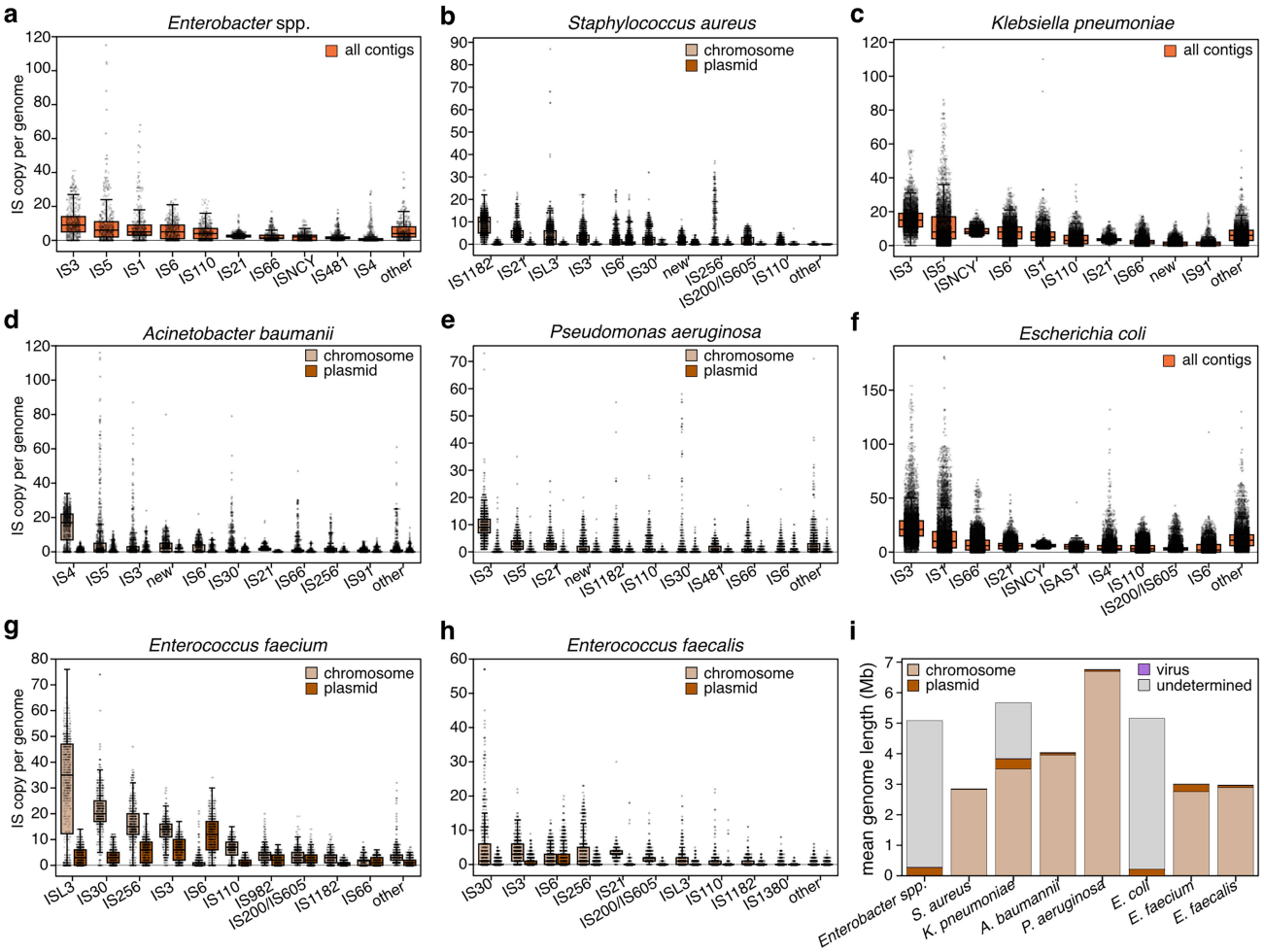
Peer review information *Nature* thanks Kate Baker, Michael Gilmore and Blake Hanson for their contribution to the peer review of this work. Peer reviewer reports are available.



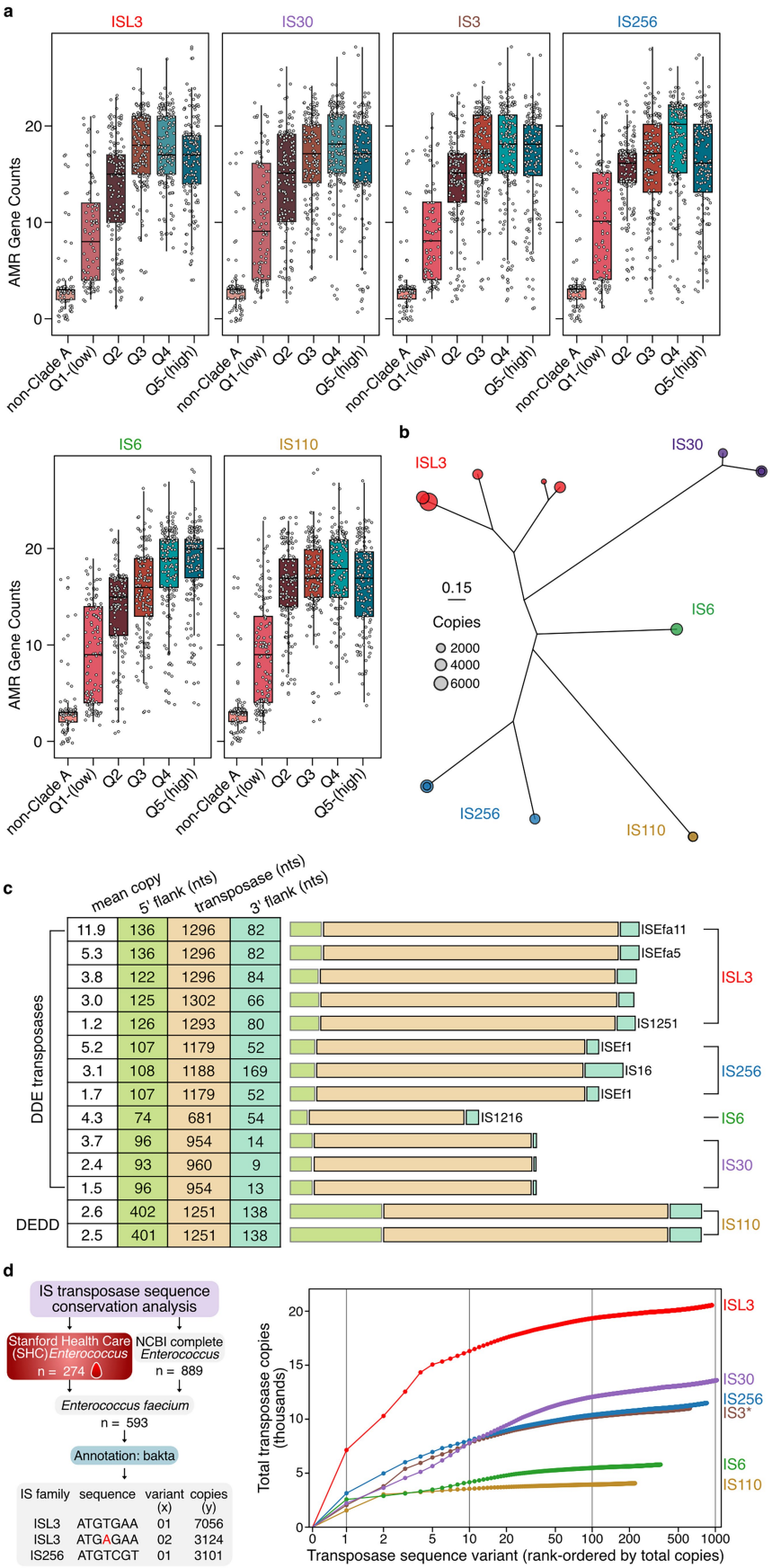
Extended Data Fig. 1 | Total IS abundance across all surveyed species. Data from the NCBI complete genomes and NCBI long-read assemblies passing quality filters are included, as in Fig. 1. Only species with at least 10 genomes defined by GTDB-Tk are included. The total IS counts per species are shown as boxplots. Centre line, median; box, 25th to 75th percentiles; whiskers, 1.5× IQR; individual values outside 1.5× IQR are shown. Species are grouped by genus, with number of genomes per species labelled below each boxplot: **a**, *Enterococcus* **b**, *Staphylococcus* **c**, *Klebsiella* **d**, *Acinetobacter* **e**, *Pseudomonas* **f**, *Enterobacter* **g**, *Escherichia* **h**, *Shigella* **i**, *Bordetella* **j**, *Streptococcus* **k**, *Salmonella*. The y-axis scale is consistent across all panels. Several species of *Shigella*[8] and *Bordetella*[9] serve as positive controls for IS expansion.



Extended Data Fig. 2 | IS copy is elevated across lineages of *E. faecium*. **a**, Mash sketches of each ESKAPEE and *E. faecalis* genome analysed in Fig. 1b were generated. All pairwise Mash distances were calculated within each individual species. The density of pairwise similarities (1-distance) was plotted and coloured by species. **b**, Clusters of at least 25 genomes with all pairwise Mash distances <0.01 were identified total insertion sequence copy per genome normalized by genome assembly length are visualized as boxplots. Centre line, median; box, 25th to 75th percentiles; whiskers, 1.5× IQR; individual values outside 1.5× IQR are shown.

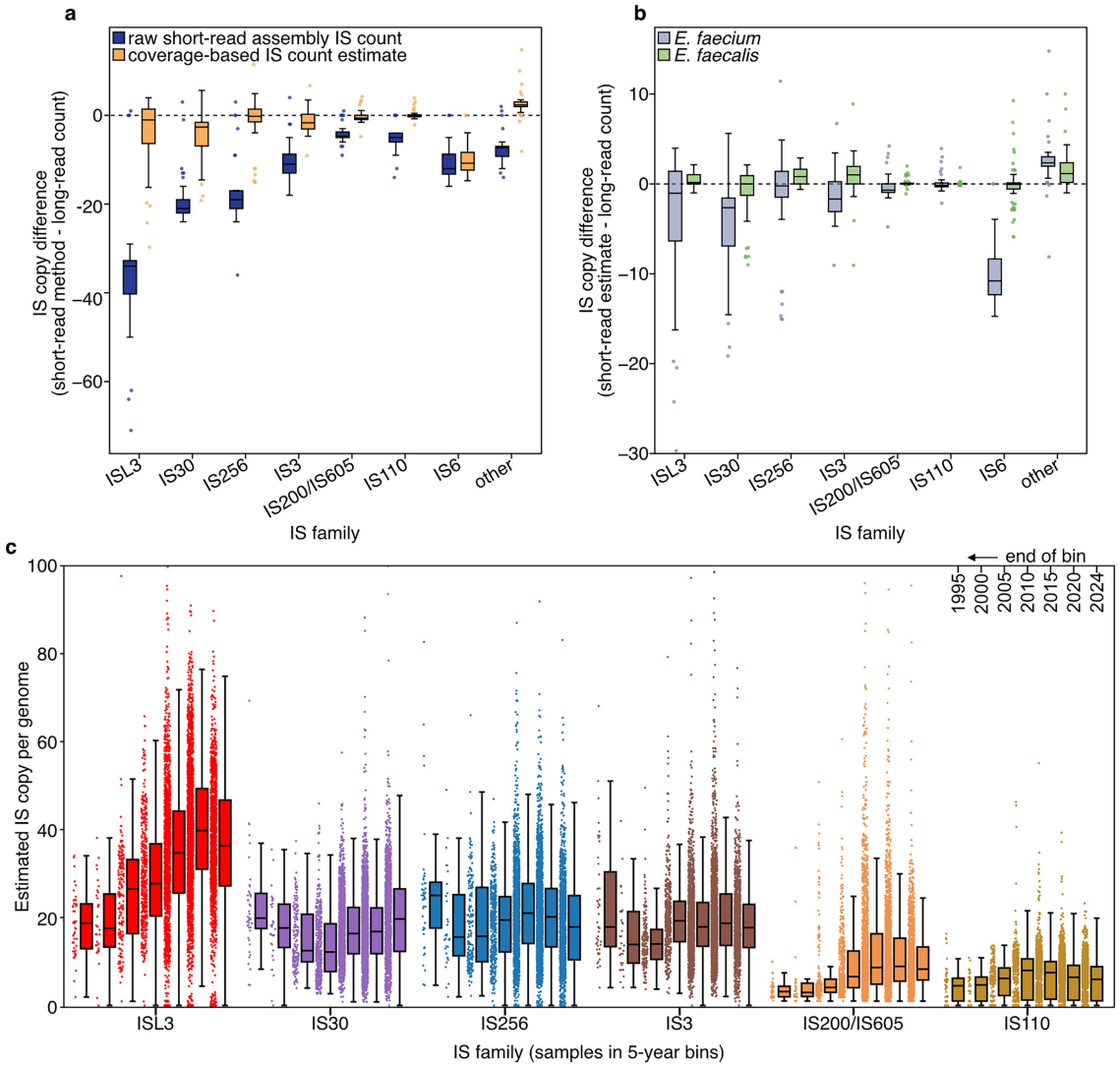


Extended Data Fig. 3 | A survey of IS family abundance in ESKAPEE pathogen genomes. Insertion sequence copy per genome for each insertion sequence family as determined by ISEScan are visualized as boxplots. Centre line, median; box, 25th to 75th percentiles; whiskers, 1.5× IQR. These counts are not normalized by genome length. These data include ESKAPEE taxa from the NCBI complete genomes and NCBI long-read assemblies passing quality filters as in Fig. 1, as well as *Enterococcus faecalis*. When possible, counts were assigned to chromosome or plasmid elements as defined by geNomad. For panels d, g, h, counts from all contigs are accumulated because geNomad chromosome vs. plasmid predictions for genomes in these species were lower-confidence than for other species. **a**, *Enterococcus faecium* (same as Fig. 1c, n = 639) **b**, *Enterococcus faecalis* (n = 917) **c**, *Staphylococcus aureus* (n = 2,672) **d**, *Klebsiella pneumoniae* (n = 4,627) **e**, *Acinetobacter baumannii* (n = 1,082) **f**, *Pseudomonas aeruginosa* (n = 1,399) **g**, *Enterobacter* spp. (n = 536) **h**, *Escherichia coli* (n = 8,530) **i**, mean combined length of all contigs, coloured by geNomad classification (chromosome, plasmid, virus or undetermined) shown for each of the ESKAPEE taxa and *E. faecalis*.

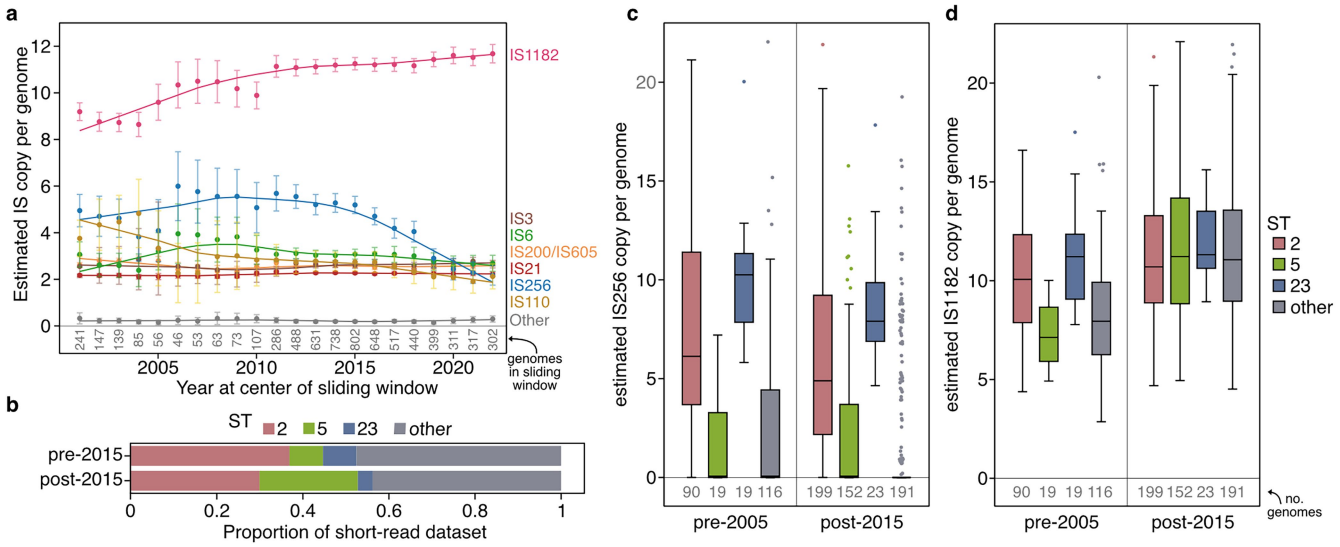


Extended Data Fig. 4 | See next page for caption.

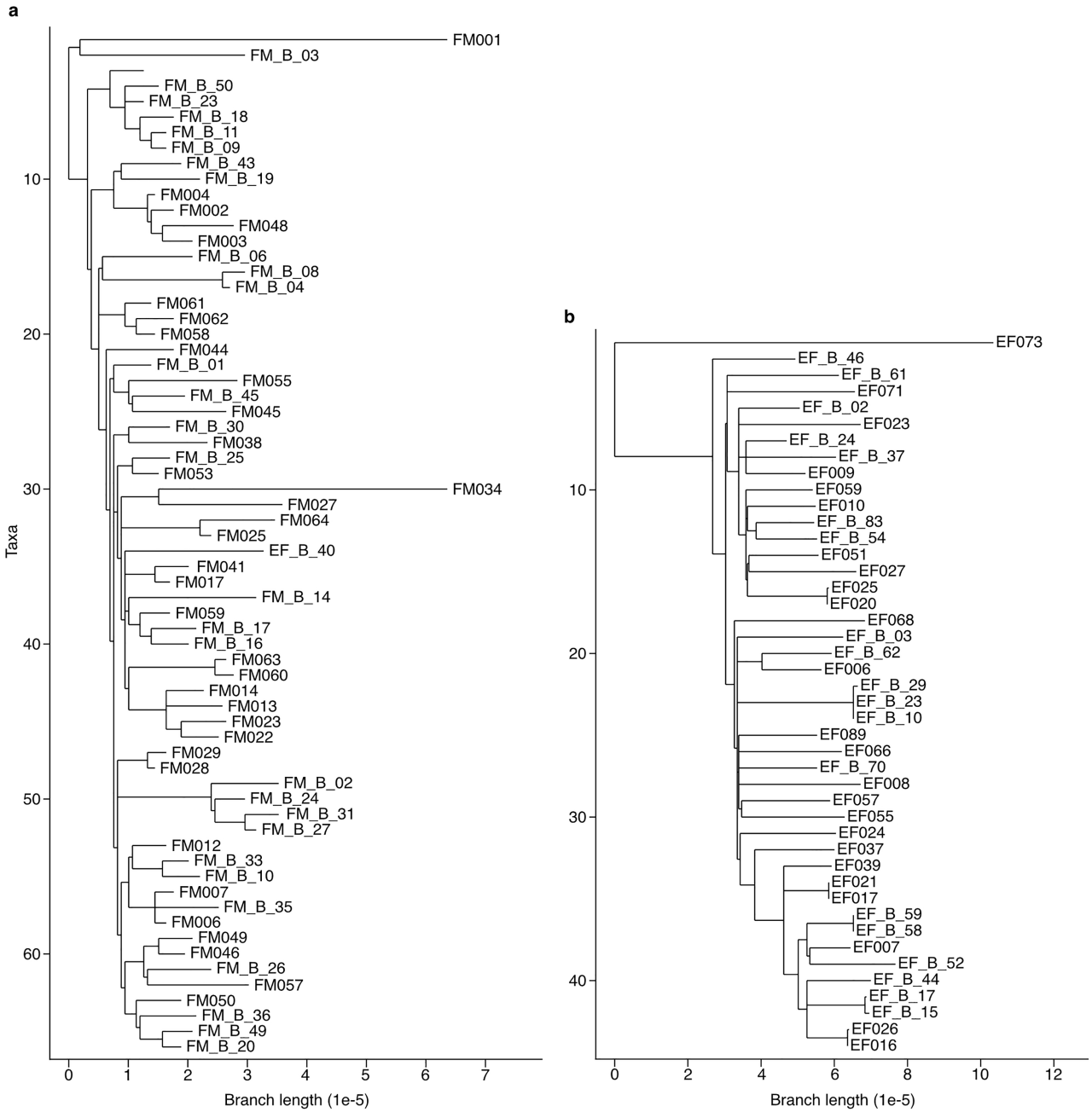
Extended Data Fig. 4 | A small number of IS variants dominate *E. faecium* genomes. AMR gene counts are related with IS abundance. AMRFinder was applied to 785 complete *Enterococcus faecium* genomes (as in Fig. 2a). To avoid multiple predictions per open reading frame, predicted AMR genes were deduplicated per genome by genomic coordinate. IS counts from 720 predicted clade A1/A2 genomes were binned into quintiles (Q1–Q5, lowest to highest IS abundance), with non–clade A1/A2 genomes binned separately. Data are shown as boxplots with individual values overlaid. Centre line, median; box, 25th to 75th percentiles; whiskers, 1.5× IQR. **b**, For each common IS family (Fig. 1c) among 593 high quality *E. faecium* genomes, IS transposase amino acid sequence copies (annotated by Bakta) were collected. Those sequences with >1 mean copy per genome (n = 14) were aligned using MAFFT, and a maximum-likelihood phylogeny was inferred with IQ-TREE2 under the VT + F + G4 substitution model (BIC-selected) with 1,000 ultrafast bootstrap replicates. As IS3 family elements are annotated as split ORFs, they were excluded from this analysis. Branch length reflects expected amino acid substitutions per site. The tree was visualized with ggtree. IS families are indicated by colour and tip size is scaled by total element copy number across genomes. **c**, Mean copy number per genome and length of the 5′-flank (green), transposase (yellow) and 3′-flank (blue) for IS families as in panel b. Transposase chemistry per IS family is indicated. **d**, Coloured per IS family, the total number of transposases (thousands) are plotted against the number of unique transposase sequences (log-scale, visualized by vertical lines) ordered from most to least. Note, IS3* (brown) are often annotated as two separate ORFs, and their count here is not adjusted.



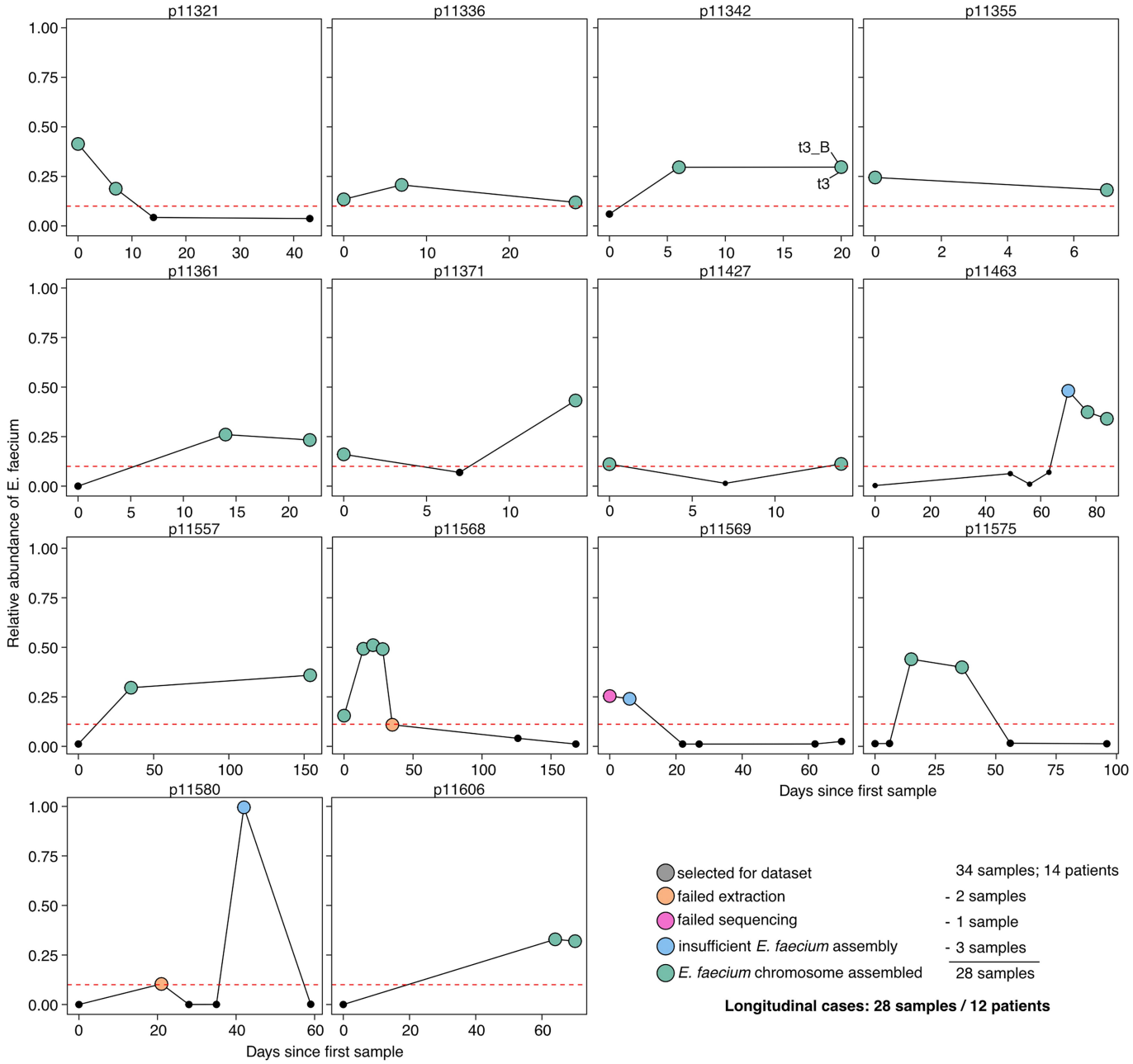
Extended Data Fig. 5 | IS copy estimation from short-read datasets over time. a, *E. faecium* genomes (n = 37 biologically independent samples, SHC long-read *Enterococcus* isolates) were sequenced with both short reads (SR) and long reads (LR). Per–IS family counts were obtained directly from SR SPAdes assemblies (assembly IS count, yellow) or estimated using our coverage-based heuristic as in Fig. 3a (estimated IS count, blue). Differences between LR counts and SR estimates are visualized as boxplots. Centre line, median; box, 25th to 75th percentiles; whiskers, 1.5× IQR; individual values outside 1.5× IQR are shown. **b**, Per–IS family counts were estimated as in Fig. 3a for *E. faecium* (n = 37) and *E. faecalis* (n = 80 biologically independent samples) isolates, visualized as in panel a. **c**, Temporal distribution of estimated IS copy number per genome across 5-year collection bins: 1995 and earlier (n = 45), 1996–2000 (n = 32), 2001–2005 (n = 171), 2006–2010 (n = 396), 2011–2015 (n = 1,411), 2016–2020 (n = 2,169), 2021 and later (n = 1,198). Boxplots are defined as in panel a, with all individual data points shown excepting 36 outlier values > 100 omitted for visualization.



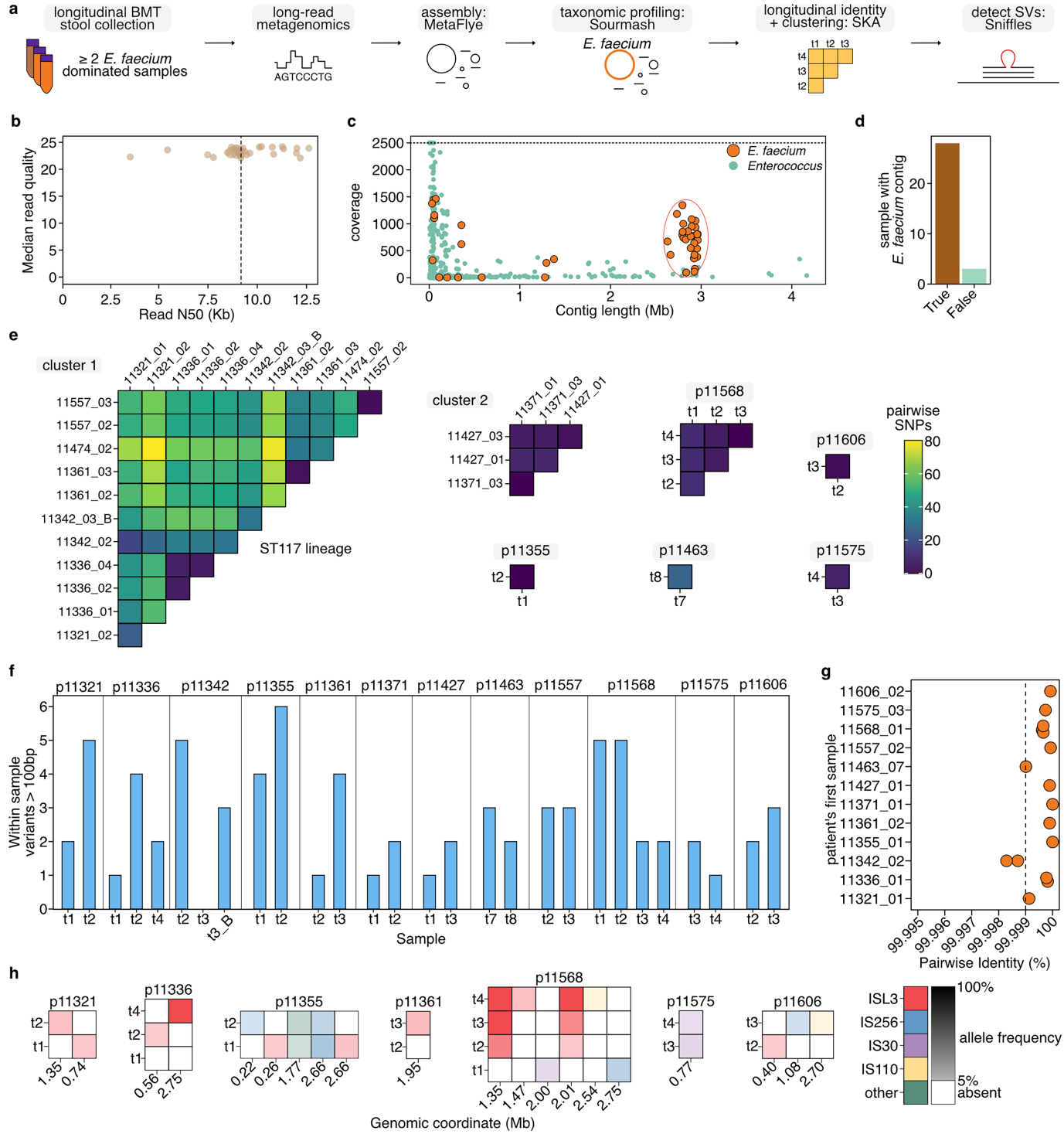
Extended Data Fig. 6 | IS dynamics in healthcare-associated *Staphylococcus epidermidis* lineages over time. a, Mean estimated IS count per genome from 2000–2023 across 1604 *S. epidermidis* genomes, 5-year sliding window, 95% CI = mean ± t·SE; LOWESS trendline smoothing fraction 0.25. All samples from before 2000 are grouped into the first window. **b**, Proportional representation of major ST lineages (ST2, ST5, ST23, and others) in each time bin. **c**, Boxplots of IS256 copy number for isolates with collection date earlier than 2005 (left, n = 224) and later than 2015 (right, n = 665) for ST2, ST5, and ST23, and all other samples. Centre line, median; box, 25th to 75th percentiles; whiskers, 1.5× IQR; individual values outside 1.5× IQR are shown. One outlier value > 23 not shown. **d**, Same as panel c, but for IS1182. Two outlier values > 23 not shown.



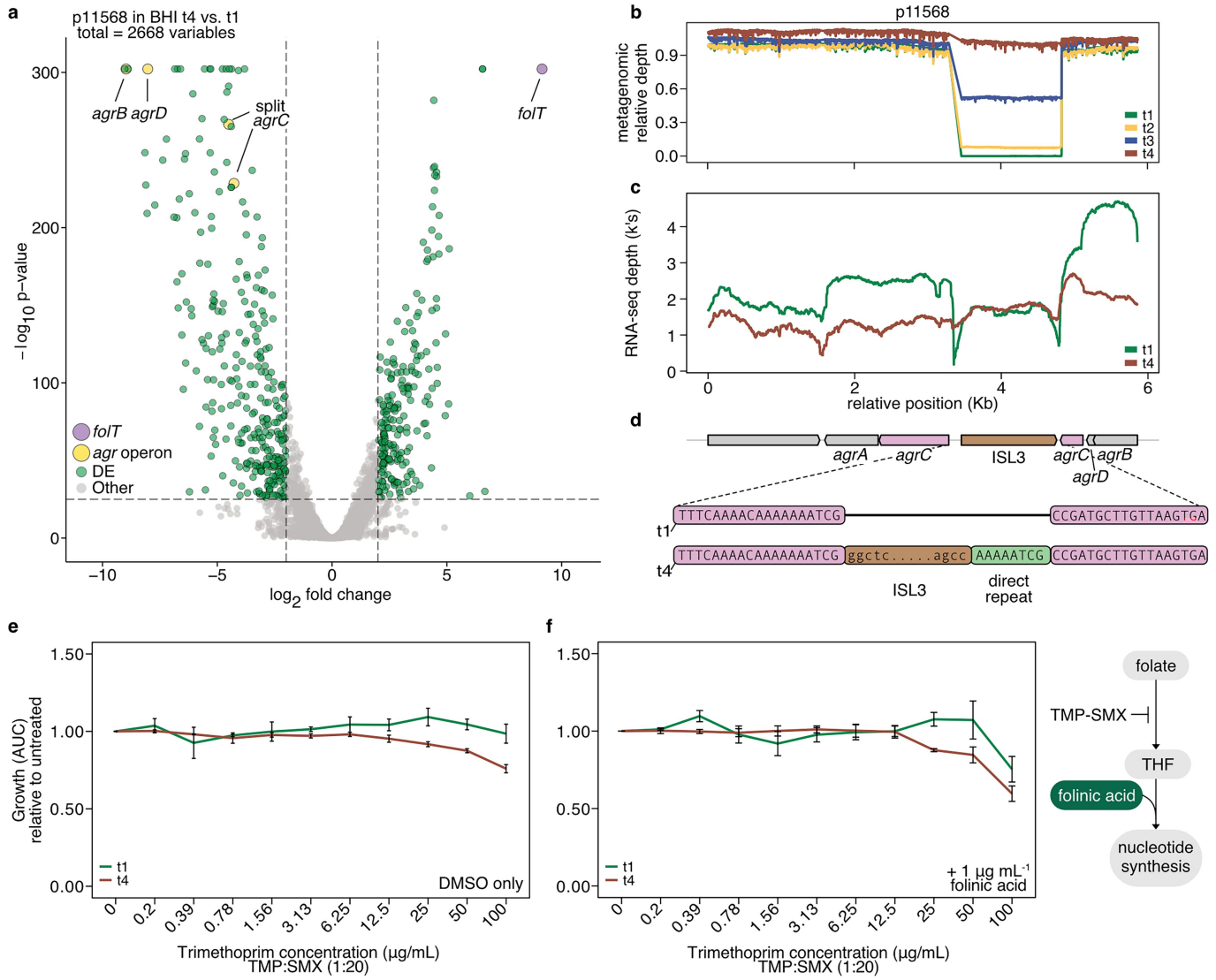
Extended Data Fig. 7 | PanGraph input cluster core genome phylogenetic trees after filtering recombination. a, 66 *E. faecium* genomes making up the SHC ST117 cluster. The polished alignment after filtering out putative recombination regions was 1.587 Mb and the tree was constructed from 865 SNPs. This results in ~8.3 SNPs per genome per Mb. **b**, 44 *E. faecalis* genomes making up the SHC ST179 cluster. The polished alignment after filtering out putative recombination regions was 2.522 Mb and the tree was constructed from 2,355 SNPs. This results in ~21.2 SNPs per genome per Mb. **a**, **b** Branch length is equivalent to the number of substitutions per core-genome alignment length, resulting in a total evolutionary time equivalent to approximately 545 and 934 substitutions per Mb of core-genome alignment for *E. faecium* and *E. faecalis*, respectively.



Extended Data Fig. 8 | Longitudinal Haematopoietic Cell Transplant (HCT) patient cohort for in situ structural variant detection. The relative abundance of *E. faecium* as measured by previous short- or linked-read metagenomics of these samples for each of 14 potential longitudinal cases[59]. Among the 34 samples from 14 patients, *E. faecium* chromosomes could not be assembled due to failed extraction (2 samples), failed sequencing (1 sample) or insufficient assembly (3 samples) for 6 samples. The resulting 28 *E. faecium* chromosomes comprise 12 longitudinal cases. Note, for p11342, time point 3 (t3) was sampled twice within a single day and both were successfully assembled.



Extended Data Fig. 9 | Longitudinal long-read metagenomics of patient stool samples identifies structural variants in otherwise isogenic strains of *E. faecium*. **a**, A workflow for performing longitudinal long-read metagenomic sequencing to identify structural variation against an isogenic background. **b**, Extracted DNA was prepared for long-read sequencing with the Oxford Nanopore Native Barcoding 96-kit (SQK-NBD114.96) and sequenced on an R10.4.1 flow cell. Median read quality was >Q22 for all samples and the median Read N50 is ~9.2 Kb (dashed line). **c**, Extracted contigs longer than 5 kb were taxonomically profiled with Sourmash. Coverage and contig length (Mb) for those contigs predicted to be within g_*Enterococcus* are plotted. *E. faecium* contigs (orange, enlarged) with at least 80x coverage and longer than 2.5 Mb were considered chromosomes for further analysis (orange oval). The y-axis is truncated at 2,500 for visualization; contigs with coverage ≥ 2,500x are plotted at this boundary (dashed line). **d**, 28/31 sequenced samples had an *E. faecium* chromosome. **e**, Pairwise comparison of *E. faecium* chromosome identity as measured by SKA. Clusters were generated with a pairwise SNP threshold of 40. Clusters 1 and 2 contain multiple patients while the other 5 clusters contain samples from a single patient. **f**, Within-sample heterogeneity of 28 *E. faecium* chromosomes. The number of variants per sample longer than 100 bp predicted by Sniffles2 are shown. Samples are grouped by the anonymized patient ID (pXXXXX) and labelled by the time point for that patient. **g**, Pairwise identity of *E. faecium* chromosomes from the same patient as measured by SKA, relative to the first sample. Samples are labelled in the format pXXXXX_timepoint. The identity threshold of 99.999% or 10 SNPs/Mb is indicated with a dashed line. **h**, Among the 11 patients reaching the identity threshold, IS variants predicted by Sniffles2 are shown per patient and arrayed by their patient-specific genomic coordinate (Mb). IS family (colour) and allele frequency (saturation) are visualized.



Extended Data Fig. 10 | ISL3 variants perturb gene expression and an active folate metabolite rescues growth under TMP-SMX treatment. a, A volcano plot showing the relative gene expression of the isolates from t4 vs. t1 (t1 in technical triplicate, t4 in technical quadruplicate). Alignments with MAPQ < 10 were excluded. Features present in at least one replicate in both t1 and t4 were considered (n = 2,668). Adjusted p-value and fold-change (log2) of each feature were calculated by DESeq2. DE genes (green), the *agr* operon (yellow, enlarged) and the *folT* gene (purple, enlarged), were additionally labelled. **b**, the relative depth of long reads aligned to the metagenomic assembled *E. faecium* chromosome of p11568 t4, normalized to the average chromosomal coverage per sample for the ISL3 variant fixing within the *agr* operon. **c**, RNA-seq read coverage (log-scale) of *E. faecium* isolated from p11568 t1 (in triplicate) and t4 (in quadruplicate) stool samples across the same locus as in panel b. **d**, ISL3 structural variant sequence context within the *agr* operon in detail. The position of the ISL3 insertion (brown), the direct repeat (green), and the split *agrC* CDS (purple) are highlighted. **e**, Dose-response curves of *E. faecium* t1 and t4 isolate growth during treatment with dimethylsulfoxide (DMSO) at TMP-SMX solvent concentrations as in Fig. 5e,f. The equivalent concentration of TMP is labelled. **f**, Dose-response curves of *E. faecium* t1 and t4 isolate growth during treatment with TMP-SMX, with all conditions supplemented with 1 μg/mL folinic acid. For panels e and f, growth was measured using the AUC and normalized to the untreated growth condition. Data are the mean ± SE of three independent experiments. The concentration of TMP is labelled (TMP:SMX was prepared in a ratio 1:20). (right) A simplified schematic showing where TMP-SMX and folinic acid act within the folate activation pathway.

nature portfolio

Corresponding author(s): Ami S Bhatt

Last updated by author(s): Jan 20, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Candidate SRA isolate long-read and short-read sequencing datasets were identified using the advanced search interface on NCBI (queries defined in Methods). Metadata for these accessions were downloaded with the SRA Run Selector interface. Raw reads were downloaded using prefetch and fasterq-dump from SRA toolkit v3.0.1. Candidate NCBI 'complete' genome assemblies were identified using the NCBI command line tool ncbi-datasets-cli v16.22.1.
Data analysis	Overall, data analysis was performed using a combination of standard tools/workflows and custom scripts. Standard tools/workflows used are comprehensively documented in the Methods section, and complex custom code is published in our GitHub repository, which can be found here: https://github.com/abehr/IS-evolution and is also referenced in the Code Availability statement. For our Stanford Hospitals Enterococcus isolate collection and for confirming p11568 t1 and t4 genotypes isolated from stool, isolates were analyzed by Plasmidsaurus (Oregon, USA) using their ONT 'Standard Bacterial Genome with DNA extraction' sequencing. Isolate long-read isolate datasets from SRA were assembled with flye v2.9 using the --nano-raw method. Reads from SRA isolate short-read datasets were deduplicated with hts_SuperDeduper v1.3.2, trimmed with trim_galore v0.6.10, and assembled with SPAdes v3.15.5. Assembly stats were generated with QUAST v5.2.0. Genome assemblies from long-read isolate datasets and complete genomes from NCBI were annotated with Bakta v1.9 (DB v5.1), and IS elements were annotated with ISEScan v1.7.2.3. Genome taxonomy was verified with gtdb-tk v2.4 release 220 with the de_novo_wf. Plasmid contigs were predicted from assemblies with geNomad (v1.8.0, v1.6 DB) with the end-to-end workflow. Pairwise mash (v2.3) distance was estimated to define sub-species lineages among ESKAPEE pathogens. AMRFinder v4.0.23 with DB 2025-07.16.1 was used to predict antibiotic resistance genes in high-contiguity Enterococcus genomes. Identification of transposase flanks included nucleotide multiple sequence alignment with MAFFT (v7.520). Amino acid multiple sequence alignments of translated transposase CDS were done with MAFFT (v7.526), phylogeny was inferred with IQ-

TREE (v2.2.0) and visualized with ggtree (v3.10.1).
A core pangenome was generated with panaroo (v1.5.2) and visualized with IQ-TREE (v2.2.0).
LOWESS smoothing for the were ISL3 expansion timeline was performed with the statsmodels (v0.14) Python module.
PopPUNK (v2.6.7, v2 E. faecium and E. faecalis PopPUNK databases) was used to identify clusters of highly related genomes within our SHC collections as input to Pangraph (full details in the Methods section).
The "Structural genome evolution" workflow published by Molari et al. was adapted to work on our cluster and used to build pangenome graphs and junction sub-graphs for each the E. faecium ST117 and E. faecalis ST179 cluster. The resulting graphs were further analyzed to produce the results shown in our paper and outputs published as Supplementary Data with custom scripts published in our GitHub repository. Circos plots based on these results were generated using the Circos module from dash_bio v1.0.2. The gene disruption and intergenic distance analysis was performed using custom scripts published in our GitHub repository.
Data analysis of the longitudinal long-read metagenomic data is detailed in the Methods section. Briefly, we used Dorado (v0.7.3) for basecalling, NanoPlot (v1.41.6) to evaluate sequencing quality, Flye (v2.9.2) for assembly, Sourmash (v4.8.11) for taxonomic profiling, and SKA (v1.0) to assess relatedness of assembled E. faecium chromosomes. For assessing structural variation within and across these chromosomes, we used ngmlr (v0.2.7) to align reads against chromosomal assemblies, and Sniffles (v2.6.0) to call structural variants. Further analysis was performed using custom scripts published in our GitHub repository.
RNA was sequenced with Novogene's Prokaryotic Total RNA sequencing service (NovaSeq PE150, 3G raw data per sample). Raw RNA-seq reads were trimmed with Trim Galore (v0.6) and aligned to the chromosome with Bowtie2 (v2.5.4); feature count matrices were generated with bedtools (v2.27.1), and differential expression analysis was performed with DESeq2 (v1.42.1) in R. Growth curves from experiments of p11568 t1 and t4 isolates were analyzed in gcplyr v1.11.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Long-read metagenomic sequencing data and RNA-sequencing data generated in this study are available under NCBI BioProject ID PRJNA1236482. Long-read assemblies from Enterococcus isolates sequenced in this study can be downloaded from Zenodo at https://doi.org/10.5281/zenodo.15239062. There, we also include some pre-processed and intermediate source data tables and other files to facilitate reproducibility of custom analyses where appropriate, as well as figure source data files for the original phylogenetic trees, and SRA accession codes for publicly-available datasets we used in this study. Accession codes for long-read pathogen isolate datasets from SRA (NCBI long-read assemblies) and for NCBI complete genomes that were used in the IS accounting analysis are also reported in Supplementary Table 2. Annotation of assemblies with Bakta59 (v1.9) used database v5.1 available at https://zenodo.org/records/10522951. For identification of chromosomal contigs, geNomad63 (v 1.6) database v1.8 was used and is available at https://zenodo.org/records/14874465. GTDB-Tk56 (v2.4) release 220 used for phylogenetic analyses is available at https://data.gtdb.aau.ecogenomic.org/releases/release220/. AMRFinder68 (v4.0.23) db 2025-07.16.1 used to measure antibiotic resistance genes is available at https://ftp.ncbi.nlm.nih.gov/pathogen/Antimicrobial_resistance/AMRFinderPlus/database/4.0/. PopPUNK72 (v2.6.7) v2 E. faecium and E. faecalis databases used for clustering are available at https://www.bacpop.org/poppunk-databases/. Source data are provided with this paper.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	NA
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. Sample size calculation is useful for scenarios when not all samples can be included. We collected all Enterococcus bloodstream isolates from our hospital over a specified time window where we report descriptive statistics.
Data exclusions	Candidate SRA long-read isolate datasets and NCBI 'complete' genomes were identified using search criteria described in detail in the Methods section, and then filtered according to quality control criteria described in the Methods section and implemented by scripts published in our GitHub repository. We excluded several samples from our Stanford Hospitals Enterococcus isolate collection after taxonomic verification and standard quality control performed by Plasmidsaurus; this is detailed in the Methods section. Samples with more than five contigs were excluded from the PanGraph analysis. For our longitudinal stool metagenomic samples: from the initial set of 34 samples nominated by criteria laid out in the Methods section, we excluded two samples from sequencing due to low DNA concentration and purity values, and one subsequently due to low sequencing depth. We excluded three more from downstream analysis that did not yield a sigle E. faecium chromosomal contig.
Replication	For RNA-seq assays, we repeated experiments in biological quadruplicate. In the comparison between p11568 t1 and t4, one replicate for t1 failed during sequencing/alignment. For E. faecium growth and antibiotic dose-response assays, we repeated experiments in technical duplicate in three independent experiments; all demonstrated the same trend. To validate the trendline of IS copy number over time, confidence intervals were generated via 1000 bootstrap replicates subsampling 25% of the data with replacement for each replicate.
Randomization	No randomization was performed for in vitro assays as inclusion in a group was determined by the bacterial genotype. We controlled for confounding variables by treating all samples in exactly the same way and with the same media.
Blinding	Blinding was not done as the results of sequencing experiments, standardized analysis pipelines and in vitro laboratory experiments are objective measurements that are not influenced by observers/participants. These types of experiments and analyses make up the results in this manuscript.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*