Emergence of Epidemic Multidrug-Resistant Enterococcus faecium from Animal and Commensal Strains

Lebreton, Francois

2013


http://hdl.handle.net/10138/166982
https://doi.org/10.1128/mBio.00534-13

Downloaded from Helda, University of Helsinki institutional repository.
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.
Emergence of Epidemic Multidrug-Resistant Enterococcus faecium from Animal and Commensal Strains

François Lebreton, Willem van Schaik, Abigail Manson McGuire, Paul Godfrey, Allison Griggs, Varun Mazumdar, Jukka Corander, Lu Cheng, Sakina Saif, Sarah Young, Qiandong Zeng, Jennifer Wortman, Bruce Birren, Rob J. L. Willems, Ashlee M. Earl, Michael S. Gilmore

Departments of Ophthalmology, Microbiology and Immunobiology, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts, USA; The Broad Institute, Cambridge, Massachusetts, USA; Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, The Netherlands; Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

F.L., W.V.S., and A.M.M. contributed equally to this article.

ABSTRACT Enterococcus faecium, natively a gut commensal organism, emerged as a leading cause of multidrug-resistant hospital-acquired infection in the 1980s. As the living record of its adaptation to changes in habitat, we sequenced the genomes of 51 strains, isolated from various ecological environments, to understand how E. faecium emerged as a leading hospital pathogen. Because of the scale and diversity of the sampled strains, we were able to resolve the lineage responsible for epidemic, multidrug-resistant human infection from other strains and to measure the evolutionary distances between groups. We found that the epidemic hospital-adapted lineage is rapidly evolving and emerged approximately 75 years ago, concomitant with the introduction of antibiotics, from a population that included the majority of animal strains, and not from human commensal lines. We further found that the lineage that included most strains of animal origin diverged from the main human commensal line approximately 3,000 years ago, a time that corresponds to increasing urbanization of humans, development of hygienic practices, and domestication of animals, which we speculate contributed to their ecological separation. Each bifurcation was accompanied by the acquisition of new metabolic capabilities and colonization traits on mobile elements and the loss of function and genome remodeling associated with mobile element insertion and movement. As a result, diversity within the species, in terms of sequence divergence as well as gene content, spans a range usually associated with speciation.

IMPORTANCE Enterococci, in particular vancomycin-resistant Enterococcus faecium, recently emerged as a leading cause of hospital-acquired infection worldwide. In this study, we examined genome sequence data to understand the bacterial adaptations that accompanied this transformation from microbes that existed for eons as members of host microbiota. We observed changes in the genomes that paralleled changes in human behavior. An initial bifurcation within the species appears to have occurred at a time that corresponds to the urbanization of humans and domestication of animals, and a more recent bifurcation parallels the introduction of antibiotics in medicine and agriculture. In response to the opportunity to fill niches associated with changes in human activity, a rapidly evolving lineage emerged, a lineage responsible for the vast majority of multidrug-resistant E. faecium infections.
the genomes of 51 strains, sampling all areas of the existing multilocus sequence type (MLST) phylogeny (Fig. 1).

RESULTS
Phylogenomic reconstruction of *E. faecium* divergence. We determined the nucleotide sequences of the genomes of 51 *E. faecium* strains of different MLST types (see Table S1 in the supplemental material), which were obtained from diverse ecological environments (see Fig. S1 in the supplemental material) on five continents, and isolated over the last 60 years (Fig. S1). A single nucleotide polymorphism (SNP)-based phylogenetic tree, which compared these strains to each other and to an additional 22 strains from GenBank (Table S1), was generated based on variation in 1,344 shared single-copy orthologous groups (orthogroups) (Fig. 2). This tree confirmed the deep divide between clades (clades A and B) (13, 16). Most (5/7) strains isolated from the feces of nonhospitalized humans cluster in clade B. We were able to resolve the epidemic hospital strains (clade A1) from a mixed group of animal strains and sporadic human infection isolates (clade A2). This clade structure was independently recapitulated based on cluster analysis of (i) shared gene content (Fig. S2) and (ii) gene synteny (Fig. S3).

Clade A1 strains account for the vast majority of human infection (Fig. 2) and include sequence types (STs) from the clonal complex 17 (CC17) genogroup (e.g., sequence type 17 [ST17], ST117, and ST78 [18]) associated with hospital ward outbreaks around the globe (see Table S1 in the supplemental material). Interestingly, the three clade A1 strains of animal origin are from pet dogs, consistent with known links between hospital strains and household pets (19). Two strains (EnGen0002 and 1_231_408) possess hybrid genomes, consisting of a background genome of clade A1, into which 195 kb to 740 kb DNA from a clade B donor have recombined (Fig. S4).

To determine whether the calculated mutation rate differences reflected historic events or whether they are still experimentally detectable, the rate of mutation to fosfomycin resistance was measured for 10 randomly selected strains from each clade. Resistance was verified for stability by passage in the absence of selection, followed by retesting. Clade A1 strains yielded spontaneous fosfomycin-resistant variants at a rate about an order of magnitude higher than strains of either clade A2 or clade B (Fig. 3), paralleling the results of BEAST analysis. Therefore, mutation

FIG 1  goeBURST analysis of 2,273 *E. faecium* entries in the *E. faecium* MLST database (http://efaecium.mlst.net), which can be grouped into 773 sequence types (STs) (brown circles), based upon MLST relatedness. STs included in this study are highlighted in purple.
rates for each clade inferred by BEAST were used to estimate the
time of divergence between clades A1, A2, and B. This placed
the time of the initial split between clade A and clade B at 2,776 ±
818 years ago and that between clade A1 and clade A2 at 74 ±
30 years ago (Fig. 2).

**Gene content differences.** Gene gain and loss make funda-
mental contributions to new habitat adaptation and the emer-
gence of new lineages (24). Strains from clade A1 were found to
have significantly larger overall average genome size (2,843 ± 159
genes; 2.98 ± 0.15 Mb) than strains of either clade A2 (2,597 ±
153 genes; 2.75 ± 0.14 Mb) or clade B (2,718 ± 120 genes; 2.84 ±
0.1 Mb) (Fig. 4A), indicating that perpetuating cycles of infection
and survival in the hospital are associated with acquisition of new
functions. Clade A1 strains also have larger core genomes (1,945
genes) than strains of clade A2 (1,724 genes) or clade B (1,805
genes), which is consistent with a very recent emergence of this

![RAxML SNP-based tree based on the concatenated alignments of DNA sequences of 1,344 single-copy core genes in 73 *E. faecium* genomes. Bootstrapping was performed with 1,000 replicates. The origins of the strains are indicated. The dates for the split between the clades, estimated by a BEAST analysis, are indicated (ya, years ago). The infectivity score reflects the number of strains of a particular ST, in the MLST database, isolated from infection. The clades are color coded as follows: clade B in dark blue, clade A1 in red, and clade A2 in gray.](https://mbio.asm.org/doi/fig/10.1128/mBio.00534-13)
lineage (i.e., little time for divergence between strains to occur) (Fig. 4C). In contrast, the pan-genome of clade A2 is larger (6,343 genes) than those of clade A1 and B (5,663 and 5,551 genes, respectively) (Fig. 4B), which is consistent with the diverse origins of strains from this clade. In comparison to other opportunists, the *E. faecium* genome is relatively open (see Fig. S5 in the supplemental material).

Previously, the genomes of hospital strains of the sister species, *Enterococcus faecalis*, were found to differ from commensal organisms largely as the result of mobile element acquisition (13), which was associated with the absence of CRISPR (clustered regularly interspaced short palindromic repeat) protection (25). It was, therefore, of interest to determine the extent to which mobile elements drove the divergence of *E. faecium* clades. Mobile elements were identified using PHAST (26) for phages, SIGI-HMM (27) for genomic islands, and BLAST for repA orthologs in plasmid-related contigs (28). Clade A1 was found to be enriched in mobile elements, including plasmids (5.4 ± 1.9 plasmids/genome in clade A1, compared to 2.7 ± 2.2 and 1.5 ± 1.1 plasmids/genome in clade A2 and B strains, respectively), integrated phages (1.6 ± 0.9 phages/genome, compared to 0.7 ± 0.7 and 0.9 ± 0.8 phages/genome in clade A2 and B strains, respectively) and other genomic islands (36 ± 26 kb of island-associated sequence/genome, compared to 14 ± 10 and 17 ± 11 kb of island-associated sequence/genome in clade A2 and B strains, respectively) (Fig. 4D). Because the genome sequences generated in the present study were of high quality, yielding a small number of scaffolds

![FIG 3](image_url) Frequency of fosfomycin resistance was determined in triplicate for 10 randomly selected strains from each *E. faecium* clade (clade A1 [red], A2 [gray], and B [dark blue]). Each symbol represents the average value for one strain, and the clade average ± standard deviation (error bars) for the 10 strains per clade are indicated.

![FIG 4](image_url) (A) Genome size comparison for *E. faecium* clade A1 (red), A2 (gray), and B (dark blue). (B and C) Pan-genome (B) and core genome (C) are shown for increasing values of the number of sequenced *E. faecium* genomes within each clade. Circles represent the number of new or core genes present when a particular genome is added to each subset. Black bars represent median values. The curve for the estimation of the size of the *E. faecium* pan-genome for each clade is a least-squares power law fit through medians. The size of the core genome within each clade was estimated by fitting an exponential curve through medians. (D) Heat map showing the enrichment in genetic mobile elements in *E. faecium* genomes within each clade (clade A1 [red], A2, [gray], and B [light blue]). Horizontal boxes represent strains, which are ordered within clades as in Fig. 2 (rotated 90°). The aggregate length (kb) of islands was used to compare content in each clade (ranging from 4 kb to 99 kb; median, 17 kb), whereas the numbers of putative plasmids (ranging from 0 to 9; median, 3) or phage elements (ranging from 0 to 4; median, 1) are represented. The heat map reflects the 10th percentile (light gray), 50th percentile (medium gray), and 90th percentile (black). The “×” symbol in a box indicates genome sequence for which the length of genomic islands could not be determined using the SIGI-HMM algorithm (27).
Emergence of Multidrug-Resistant Enterococcus faecium

FIG 5 Summary of clade-specific antibiotic resistance genes, insertion sequences (IS), and select defenses against horizontal gene transfer. Each box represents a strain, arranged by clade as shown in Fig. 2. The “X” symbol in a box indicates genome sequence with an assembly quality that precluded identification of the indicated feature. An asterisk in a box indicates hybrid genomes that contain CRISPR-cas on recombined fragments. CRISPR and type IV restriction-modification (RM) systems are included in the miscellaneous (Misc.) category.

(see https://olive.broadinstitute.org/projects/work_package_1/downloads), we were able to quantify and determine the rate of occurrence and location of IS elements. IS element occurrence ranges from a low of 2.6 per Mbp (clade B strain EnGen0047) to a high of 50.7 IS elements per Mbp (clade A2 strain EnGen0024). Three IS elements (ISEnfa3, ISSpm10, and IS16) are highly enriched in clade A1 and are found outside this clade only in a single clade A2 strain (EnGen0024) and the clade A1/B hybrid strain, EnGen0002 (Fig. 5). On average, strains of clade A1 harbored a total of 391 kb of mobile element DNA, and clade A harbored an average of 332 kb. Clade B strains contained an average of 340 kb of mobile element DNA.

To identify functional differences and remaining differences in gene content not restricted to mobile elements, we next identified orthogroups present in ≥80% of genomes of one clade but in ≤20% of strains from a comparator (see Table S2 in the supplemental material). Contiguous groups of genes were identified and associated with the mobile elements identified above where possible. To begin to understand the ecological forces that led to the initial bifurcation between clades A and B, we identified genes occurring in most clade A (A1 plus A2) strains but that were rare in clade B and vice versa. We found 66 orthogroups enriched at the level of ≥80% in clade A and ≤20% in clade B and 138 orthogroups enriched in clade B versus clade A (Table S2). Genes enriched in clade A strains largely occurred in 12 clusters of contiguous genes (cluster 2 [C2], C8, C10, C11, C12, C17, C19, C20, C21, C22, C23, and C24), with 8 clusters occurring in identifiable mobile elements. Cluster 10, 11, 12, and 24 genes encode functions related to altered carbohydrate utilization (Table 1 and Table S2). Cluster 19 genes include ABC transporters putatively related to antibiotic transport. Other genes enriched in clade A strains, with predicted roles in adapting to different habitats, include genes encoding a putative membrane-bound metalloprotease in cluster 17 that likely confers resistance to a cognate bacteriocin (29), and an LPXTG-anchored collagen adhesin in cluster 21 that may relate to colonization and niche selection (30). Individual genes showing an enrichment in clade A versus clade B strains include a putative choloylglycine bile hydrolase related to that known to be important in the pathogenesis of Listeria infection (31), which may enable E. faecium to colonize regions of the intestine more proximal to the bile duct.

Genes representing 138 orthogroups were found to be enriched in clade B strains compared to clade A strains. These largely occur in 24 clusters of contiguous genes but this time with few signatures of mobile elements. Gene groups C33, C35, C37, C43, C44, C45, C51, and C54 and a single gene (EfmE980_2866) have predicted roles in carbon metabolism, highlighting the differential use of carbohydrates by strains of each clade (Table 1; see Table S2 in the supplemental material). Cluster 50 encodes a cysteine-containing DnaJ-like chaperone, adjacent to a putative metallo-β-lactamase class protein that is likely to be involved in the homeostasis of glutathione pools (since these commensal strains of E. faecium do not inactivate β-lactams), involved in maintenance of protein structure. A main driver of clade divergence, therefore, appears to stem from residence in different ecological environments that have selected for the systematic exchange of phosphotransferase system (PTS) systems, with strains of clade A acquiring new PTS systems on mobile elements and deleting obsolete PTS systems from the clade B chromosome.

Interestingly, cluster 39, which is enriched in clade B, contains four genes that are predicted to form an agr-like quorum-sensing system (32), along with another Mga-type regulator that may connect quorum sensing to carbohydrate utilization (Table 1; see Table S2 in the supplemental material) (33). Unexpectedly, cluster 53, with an apparent 98-amino-acid secretion target (EfmE980_2510), which also is enriched in clade B, appears to encode a type VII secretion system. Both agr (32) and type VII secretion systems (34, 35) have been studied for their contribution to infection pathogenesis, but the pattern of differential presence observed here highlights potentially important roles in commensalism as well.

It was also of interest to examine differential gene presence in clades A1 and A2. In hospital epidemic clade A1, 48 genes were identified as differentially present, with 37 genes occurring in 6
distinct clusters associated with mobile elements (Table 1; see Table S2 in the supplemental material). Interestingly, the split between clades A1 and A2 is also associated with the gain of pathways for carbohydrate utilization. Clade A1 strains acquired an apparent mobile element of 13 genes (C6 [Table 1]) encoding enzymes for uptake and utilization of fructose, sorbose, and mannose. This appears to be functionally related to a cluster (C36) that earlier was lost from clade B by strains of clade A. C6 is known to play an important role in GI tract colonization following antibiotic treatment (36). It is interesting that clade A1 recovered this ability, and this observation suggests that it may relate to human colonization.

Cluster C16 is also differentially enriched in clade A1 and contributes to carbohydrate utilization. No orthogroups were enriched in clade A2 versus clade A1.

### Table 1

<table>
<thead>
<tr>
<th>Cluster</th>
<th>A vs B</th>
<th>A1 vs B</th>
<th>A2 vs B</th>
<th>B vs A</th>
<th>A1 vs A2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Putative function of the cluster or gene of interest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>PTS system, N-acetylglucosamine-specific</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>PTS system, glucitol/sorbitol-specific</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Alternate pathways for glycolysis and gluconeogenesis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Starch, xylose and sucrose utilization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>ABC transporter and regulatory proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>ABC transporter of unknown function</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>ABC transporter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Bacteriocin self-immunity protease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Surface proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Hexapaptide transferase, LysR substrate binding domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Putative toxin-antitoxin system</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>PTS system, Glucose/mannose and GlcNAc, ManNAc and Neu5Ac</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>PTS system, Lactose/Cellobiose specific</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Enterocin A immunity, Class II bacteriocin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>ABC transporter of unknown function</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>IS66 family transposase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Putative toxin-antitoxin system IS605- and IS200-like</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>PTS system, mannose/fructose/sorbose specific</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Glycosyl hydrolase, Sugar uptake systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Phage integrase and excisionase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>PTS system, sucrose/amylose specific</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>PTS system, fructose/sorbose specific</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>PTS system, Lactose/Cellobiose specific</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>PTS system associated Lactose/Cellobiose/maltose</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>Exopolysaccharide biosynthesis, glycosyltransferase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Mga regulators</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>Mga regulator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Differentially occurring clusters of genes associated with chromosomal DNA (black), putative ICE elements (integrative and conjugative elements) (dark gray), plasmids (medium gray), or phages (light gray). Clusters functionally associated with carbohydrate uptake and utilization are indicated in blue type. No genes are differentially enriched in the genomes of strains in clade A2 compared to clade A1. HTH, helix-turn-helix.
We identified additional genes that show enrichment in clade A1 compared to clade B. Gene clusters 1, 3, and 27 putatively encode proteins for PTS systems and enzymes for the interconversion and metabolism of lactose/cellobiose, glucose, mannose, N-acetylneuraminic acid, N-acetylmannosamine, and other sialic acids. Clusters 1 and 27 are associated with mobile elements. Cluster 18 (C18), which is also enriched in clade A1 compared to clade B, encodes a three-gene operon for a class II bacteriocin that may be a colonization factor (Table 1; see Table S2 in the supplemental material).

Bifurcation of clade A parallels the proliferation of resistance. To understand the role that antibiotics played as a driver of clade formation, we examined the differential presence of resistance genes (see Table S3 in the supplemental material). Two resistance genes [aac(6’)-Ib conferring resistance to kanamycin and bacA conferring bacitracin resistance] are part of the core E. faecium genome. The ubiquitous presence of aac(6’)-Ib has been observed before and contributes to the intrinsic resistance of E. faecium to several aminoglycosides (37). The bacA gene may be responsible for intrinsic resistance to bacitracin observed among E. faecium (38). Seven strains analyzed were isolated in the 1950s and 1960s, allowing for the identification of genes associated with some of the earliest known acquired resistances to occur in E. faecium. Strains EnGen0025, EnGen0027, EnGen0031, EnGen0032, and E1636 were isolated between 1957 and 1965; these strains fall into clade A2. Each of these strains also possesses the fusA fusidic acid resistance gene. Additionally, strains EnGen0025, EnGen0027, EnGen0031, and E1636 possess the msrC gene, which confers erythromycin resistance. Strain EnGen0025 additionally acquired the aminoglycoside resistance genes ant(6’)-la (conferring resistance to streptomycin) and aph(3’)-III (conferring resistance to several aminoglycosides, including neomycin and gentamicin B), ermB, and tetM. As shown in Fig. 2, this strain (the fifth strain from the top of clade A2) is closely related to the clade A1 branch point and presumably the clade A1 founder.

Other resistances exhibit clear clade specificity (Fig. 5; see Table S3 in the supplemental material). Vancomycin resistance is completely absent from clade B. Vancomycin resistance occurs mainly in clade A1 but also occurs in clade A2. Aminoglycoside resistance genes ant(6’)-la and aph(3’)-III are completely absent from clade B strains, but they occur in most clade A1 isolates. Interestingly, in clade B, the msrC resistance gene correlates perfectly with the presence of a CRISPR element. We have not found prior mention of the occurrence of several resistance genes in E. faecium, including the aadD cassette, which confers resistance to tobramycin and kanamycin, in a single genome (strain EnGen0035). We also observed genes lnuB, ermG, and ermT (that likely confer various degrees of resistance to the macrolides-lincosamides-streptogramin B [MLS] class of antibiotics), tetC (conferring resistance to tetracycline), and fosB (conferring resistance to fosfomycin) in E. faecium.

Clade structure is reflected in E. faecium genome organization. The Aus0004 genome possesses a previously identified 683-kb inversion around the replication termination site (17). Similar inversions appear to have occurred several times independently (since boundaries were not strictly identical) in strains of clades A1 and A2 (i.e., in strains EnGen0007 and EnGen0025), but not in strains of clade B (see Fig. S6 in the supplemental material). This inversion is bounded by different phases in different strains, and it appears that the recombination responsible for this rearrangement occurred within the phage sequence. Larger inversions in other areas of clade A1 and A2 genomes were also observed, including a 1.2-Mbp inversion in both EnGen0046 and EnGen0049, and again appear to be driven by recombination within phages present at the boundaries. Most genome rearrangements observed in E. faecium can be linked to the occurrence of mobile genetic elements at the boundaries. Select novel rearrangements were arbitrarily verified by PCR, and the accuracy of assembly was verified in each case.

In addition to mediating inversions and recombinations, introduction and proliferation of IS elements in a bacterial population can facilitate adaptation to new niches as the result of obsolete gene inactivation (1). We identified 133 instances of IS element-mediated gene inactivation in E. faecium (see Table S4 in the supplemental material). The number of IS-mediated gene inactivation events was highest in clade A1 genomes and lowest in clade B strains. In clade A1 strains, we found a strong enrichment for disruption of a core gene encoding a putative major facilitator superfamily (MFS) transporter (EFAU004_02447 in strain AU50004) (Table S4).

Since compromised defense was associated with the evolution of hospital epidemic strains of E. faecalis (25), it was of interest to examine more closely the relationship between the presence of a CRISPR-Cas system and mobile element content. We therefore examined the 73 E. faecium genome sequences studied for the presence of CRISPR-cas using CRISPRfinder (39). Only 7 E. faecium genomes carried cas genes (Fig. 5), and in 5 of these (strains Com12, EnGen0002, EnGen0056, 1_141_733, and 1_231_408), a CRISPR array could be readily identified immediately downstream. In strains 1_231_408 and EnGen0056, where spacers could be matched to known genes, one was derived from a phage that is a common lysogen in E. faecium genomes (present in 39 out of 73 genomes). Interestingly, this phage is absent from these 2 genomes, suggesting CRISPR-Cas functionality. Notably, all strains that carry cas genes are either found in a distinct subgroup within clade B or are hybrid strains 1_231_408 and EnGen0002 that acquired the cas genes and its associated CRISPR-locus from the clade B parent (see Fig. S4 in the supplemental material). Apart from the CRISPR defense, we observed a gene encoding a putative type IV methyl-directed restriction enzyme in strains of both clade B and A2, but not in clade A1 genomes (Fig. 5).

Evidence of varying selection in genomes from each clade. We examined polymorphisms in shared genes to detect genes under particularly strong selection in the different habitats occupied by strains of each clade. Because of the clade structure, we used a tree-based approach (40) to compare the ratios of nonsynonymous to synonymous base changes (dN/dS ratio). We removed potentially confounding (41) recombined fragments using BRAT-NextGen (21). Genes under positive selection were identified when the dN/dS ratio in the clade of interest (foreground) was observed to be significantly higher than the dN/dS ratio in the comparator genomes (background) (see Table S4B in the supplemental material). No genes were found to be under positive selection in clade B compared to clades A, A1, and A2, likely reflecting the fact that clade B strains had long-fixed beneficial mutations in this particular niche before the emergence of the A clade. Only four genes were found to be under differential positive selection pressure in clade A compared to clade B, two of which were annotated as having roles in amino acid transport and metabolism (Table S4B). Interestingly, in strains of the hospital-adapted clade
A1, a penicillin binding protein transpeptidase and the D-alanyl-D-alanine ligase were under differential positive selection compared to strains of both clades A2 and clade B (Table S4B). Finally, an MFS transporter involved in carbohydrate transport and metabolism in clade A1 and an N-acetylglucosamine transferase in clade A2 were found to be under positive selection pressure, providing independent support for the importance of differential carbohydrate utilization as a determinant of clade structure, as inferred from gene gain/loss patterns described above.

**DISCUSSION**

Speciation results from expansion into new ecological niches and subsequent isolation from the founder population (42) and is accompanied by changes in the genome stemming from mutation, recombination (43), and horizontal gene transfer (44). All of these processes have contributed to the current population structure of *E. faecium* and its emergence as a leading multidrug-resistant hospital pathogen.

Quantification of mutation rates for strains in each *E. faecium* clade allowed us to estimate that the first bifurcation in the *E. faecium* population took place approximately 3,000 years ago, substantially sooner than previously suggested (16). Although it is difficult to know the ecological drivers of this split with precision, the timing suggests it relates to increasing insulation between the flora of humans and animals, which likely stemmed from increased urbanization, increased domestication of animals providing restricted and specialized diets (45, 46), and increasing use of hygienic measures (47, 48). This bifurcation was associated with a wholesale loss and replacement of carbohydrate utilization pathways, mediated largely by acquisition on mobile elements by strains of clade A. Many of the clade B pathways lost by clade A strains relate to the utilization of complex carbohydrates from dietary sources, and the pathways lost were replaced by pathways on mobile elements associated with the utilization of amino sugars, such as those occurring on epithelial cell surfaces and in mucin, suggesting a possible shift from a lifestyle dependent mainly on host diet (clade B) to one increasingly dependent on host secretions (clade A). In addition to carbohydrate utilization pathways, there was a substantial shift in genes encoding Mga-type helix-turn-helix regulators, which in *Streptococcus pyogenes* connect expression of niche-specific genes with carbohydrate metabolism (33).

The second split in the *E. faecium* population, the split between clade A1 and clade A2, appears to have occurred approximately 75 years ago, coinciding precisely with the introduction of antibiotics in both clinical medicine and agriculture. However, this split may not have been directly driven by the usage of antibiotics, as antibiotics are used both in farming and in human medicine. The ability to rapidly acquire new traits on mobile elements, including carbohydrate utilization pathways as well as resistance to antibiotics, appears to be an intrinsic trait of clade A1 and clade A2. Although clade A1 strains now cause the vast majority of infections (Fig. 2), early clinical isolates from the 1950s and 1960s do not cluster in clade A1. The earliest isolation of a strain associated with an MLST type occurring in clade A1, occurred in 1982 (49). That isolate already possessed high-level resistance to gentamicin and carried the *esp* gene.

Interestingly, we found that the recently emergent hospital-adapted clade A1 is hypermutable, as reflected in the inferred rate of mutation in the genomes, and experimentally. Hypermutation in Gram-negative bacteria has been linked to the emergence of antibiotic-resistant lineages that are pathogenic to humans (50–52). In Gram-positive bacteria, hypermutating populations of pathogenic *Streptococcus pneumoniae* and *Staphylococcus aureus* have been observed (53, 54). In *E. faecium*, polymorphisms in *mutS* and *mutL* (which encode DNA mismatch repair proteins) have been noted (55), but the polymorphisms are not associated with differential mutation rates in different clades. Higher mutation rates have been associated with microbes recently experiencing a host switch (e.g., *Mycoplasma gallisepticum*, 0.8 × 10^-8 to 1.2 × 10^-8 substitutions per site per year [61]) and with the emergence of pathogenic lineages (52), possibly including *E. faecium* strains of the CC17 genogroup (56). It appears that the epidemic hospital clade A1 emerged because of its ability to acquire mobile elements, its ability to utilize carbohydrates of nondiary origin, and its hypermutability.

Previously, the average nucleotide identity of eight *E. faecium* strains was determined to range between 93.5 and 95.6% when comparing strains from clades A and B (13), and clade A and B strains would be considered to be distinct species by existing criteria (57, 58). The identification of hybrid clade A1/B strains (strains EnGen0002 and 1_231_408) show that the ecological niches of human-infecting hospital strains and human commensal strains do occasionally overlap. The emergence of the distinct clade structure in *E. faecium* parallels anthropogenic changes in urbanization and animal domestication and, more recently, the introduction of antibiotics into agriculture and medicine. The net effect of these forces is the emergence of a rapidly evolving lineage, which has crossed a degree of divergence usually associated with speciation.

**MATERIALS AND METHODS**

**Bacterial strains.** Strains selected for genome analysis were drawn from those representing diverse points within the known phylogenetic structure, as determined by MLST (Fig. 1), and are listed in Table S1 in the supplemental material. DNA was purified from each *E. faecium* strain as described before (13) for DNA sequence analysis. Methods for DNA sequencing, genome assembly, and bioinformatic analysis are provided in Supplemental Methods at https://olive.broadinstitute.org/projects/work_package_1/downloads, along with details of the genome sequences.

**SUPPLEMENTAL MATERIAL**

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00534-13/-/DCSupplemental.

Figure S1, JPG file, 0.5 MB.
Figure S2, JPG file, 1.4 MB.
Figure S3, JPG file, 1.5 MB.
Figure S4, JPG file, 2.6 MB.
Figure S5, JPG file, 0.7 MB.
Figure S6, JPG file, 6.2 MB.
Table S1, DOCX file, 0.1 MB.
Table S2, PDF file, 0.6 MB.
Table S3, PDF file, 0.1 MB.
Table S4, DOCX file, 0.1 MB.

**ACKNOWLEDGMENTS**

This project was funded in part by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract HHSN27200900018C. Portions of this work were also supported by NIH/NIAID grants AI083214 (Harvard-Program on Antibiotic Resistance), and AI072360. W.V.S. and R.J.L.W. were supported by the European Union Seventh Framework Programme (FP7-HEALTH-2011-single-stage) “Evolution and Transfer..."
of Antibiotic Resistance” (EvoTAR) under grant agreement number 282004.

We acknowledge Lucia Alvarado and Clint Howarth for data submissions, Susanna Hamilton and Sinead Chapman for project management, Chris Desjardins for helpful discussions, and Matthew Laird for help with IslandViewer.

ADDENDUM IN PROOF

Following submission we were made aware that others recently described a split, between human and bovine populations of S. aureus, datable by BEAST analysis, to approximately 5,000 years ago (L. A. Weinert, J. J. Welch, M. A. Suchard, P. Lemey, A. Rambaut, and J. R. Fitzgerald, Biol Lett. 8:829-832, 2012).

REFERENCES


