Insertion Sequence–Driven Diversification Creates a Globally Dispersed Emerging Multiresistant Subspecies of *E. faecium*

Helen L. Leavis[1,2*], Rob J. L. Willems[1], Willem J. B. van Wamel[1], Frank H. Schuren[2], Martien P. M. Caspers[2], Marc J. M. Bonten[1]

1 Eijkman–Winkler Institute for Medical Microbiology, Infectious Diseases and Inflammation, University Medical Center Utrecht, Utrecht, The Netherlands, **2** TNO Quality of Life, Department of Microbiology, Zeist, The Netherlands

***Enterococcus faecium,* an ubiquous colonizer of humans and animals, has evolved in the last 15 years from an avirulent commensal to the third most frequently isolated nosocomial pathogen among intensive care unit patients in the United States. *E. faecium* combines multidrug resistance with the potential of horizontal resistance gene transfer to even more pathogenic bacteria. Little is known about the evolution and virulence of *E. faecium,* and genomic studies are hampered by the absence of a completely annotated genome sequence. To further unravel its evolution, we used a mixed whole-genome microarray and hybridized 97 *E. faecium* isolates from different backgrounds (hospital outbreaks ($n = 18$), documented infections ($n = 34$) and asymptomatic carriage of hospitalized patients ($n = 15$), and healthy persons ($n = 15$) and animals ($n = 21$)). Supported by Bayesian posterior probabilities (PP = 1.0), a specific clade containing all outbreak-associated strains and 63% of clinical isolates was identified. Sequencing of 146 of 437 clade-specific inserts revealed mobile elements ($n = 74$), including insertion sequence (IS) elements ($n = 42$), phage genes ($n = 6$) and plasmid sequences ($n = 26$), hypothetical ($n = 58$) and membrane proteins ($n = 10$), and antibiotic resistance ($n = 9$) and regulatory genes ($n = 11$), mainly located on two contigs of the unfinished *E. faecium* DO genome. Split decomposition analysis, varying guanine cytosine content, and aberrant codon adaptation indices all supported acquisition of these genes through horizontal gene transfer with IS*16* as the predicted most prominent insert (98% sensitive, 100% specific). These findings suggest that acquisition of IS elements has facilitated niche adaptation of a distinct *E. faecium* subpopulation by increasing its genome plasticity. Increased genome plasticity was supported by higher diversity indices (ratio of average genetic similarities of pulsed-field gel electrophoresis and multi locus sequence typing) for clade-specific isolates. Interestingly, the previously described multi locus sequence typing–based clonal complex 17 largely overlapped with this clade. The present data imply that the global emergence of *E. faecium,* as observed since 1990, represents the evolution of a subspecies with a presumably better adaptation than other *E. faecium* isolates to the constraints of a hospital environment.**

Citation: Leavis HL, Willems RJL, van Wamel WJB, Schuren FH, Caspers MPM, et al. (2007) Insertion sequence–driven diversification creates a globally dispersed emerging multiresistant subspecies of *E. faecium*. PLoS Pathog 3(1): e7. doi:10.1371/journal.ppat.0030007

Introduction

Once not recognized as clinically relevant microorganisms, enterococci currently are the third most frequently isolated nosocomial pathogen from intensive care unit patients in the United States [1]. The emergence of enterococci as nosocomial pathogens in the 1990s was associated with a gradual replacement of *Enterococcus faecalis* by *Enterococcus faecium* and an epidemic rise of vancomycin-resistant *E. faecium* [2]. In Europe, though, vancomycin-resistant enterococcus (VRE) initially was only found to colonize healthy individuals, and nosocomial VRE outbreaks have only recently begun to emerge. This epidemiological difference between the US and Europe presumably resulted from massive bioindustrial avoparcin usage in Europe, which created a VRE reservoir among farm animals with spillover via the food chain to consumers [3–8]. Abundant antibiotic use in hospitals, most notably of vancomcyin and cephalosporins, was the presumed cause of VRE emergence in US hospitals [9].

The emergence of VRE as a nosocomial pathogen in countries with polyclonal endemicity seems irreversible, despite enforced hygiene measures and restricted antibiotic prescription policies [10]. Nevertheless, despite unsuccessful eradication, a sustained reduction in prevalence rates has been reported [11]. Also, successful control of monoclonal outbreaks in countries with low VRE prevalence has been reported [12]. Recent reports on the transfer of vancomycin resistance from enterococci to methicillin-resistant *Staph-*

Editor: Frederick M. Ausubel, Harvard Medical School, United States of America

Received August 30, 2006; **Accepted** November 30, 2006; **Published** January 26, 2007

Abbreviations: CAI, Codon Adaptation Index; CC, clonal complex; CGH, comparative genomic hybridization; COG, cluster of orthologous genes; *esp, enterococcal surface protein;* FDR, false discovery rate; g, genomic; GC, guanine cytosine; IS, insertion sequence; MLST, multi locus sequence typing; ORF, open reading frame; p, plasmid; PAI, pathogenicity island; PFGE, pulsed-field gel electrophoresis; PP, Bayesian posterior probabilities; RSCU, relative synonymous codon usage; SD, standard deviation; SDA, split decomposition analysis; VRE, vancomycin-resistant enterococcus

* To whom correspondence should be addressed. E-mail: hleavis@umcutrecht.nl

Author Summary

Whole-genome sequencing has become instrumental in investigating the genome contents of bacteria. However, there is enormous diversity within bacterial populations, and annotation of multiple genomes is costly and elaborate. For investigating diversity and phylogeny within bacterial species, comparative genomic hybridization is an attractive alternative that may provide fundamental insights into the factors (genes) distinguishing bacterial subpopulations. *Enterococcus faecium,* a worldwide emerging nosocomial pathogen usually resistant to multiple antibiotics, causes infections in immunocompromised patients. Using comparative genomic hybridization of 97 *E. faecium* strains isolated from different epidemiological niches worldwide, a subpopulation of *E. faecium* strains was identified that was associated with invasive infections and hospital outbreaks. Approximately 13% of the *E. faecium* pangenome was highly specific for this subpopulation, and, based on phylogenetic clustering, it should be considered a subspecies. We hypothesize that extensive variation within specific functional genes and high prevalence of mobile elements, mostly insertion sequence elements, contributed to the success of this genetic subset in its competition with other enterococci in hospital settings, creating a novel globally dispersed nosocomial subspecies. These findings fully confirmed previous phylogenetic studies based on multi locus sequence typing that had also revealed a genetic subset of *E. faecium,* clonal complex 17. Identification of genes specific for clonal complex 17 is a first step in elucidating how global spread and adaptation to the hospital environment of this emerging nosocomial pathogen has occurred.

ylococcus aureus [13–16] stressed the need to better understand molecular epidemiology, as well as transmissibility and virulence of enterococci, to control further spread and develop treatment and eradication strategies. Yet, little is known about the virulence and pathogenesis of *E. faecium*. Apart from antibiotic resistance genes, only the *enterococcal surface protein (esp)* gene and the *hyaluronidase* gene have been epidemiologically associated with infections and documented hospital outbreaks [17–20]. The *esp* gene is contained on a putative pathogenicity island (PAI), but functional studies on any of these *E. faecium* genes have not been performed yet [18]. Previously, we described the population structure of *E. faecium* with multi locus sequence typing (MLST), relying on the variation in silent mutations in short sequences from seven housekeeping genes [21,22]. This population shows host specificity, and a globally present hospital subpopulation, clonal complex (CC) 17, is responsible for most outbreaks and colonization of hospitalized patients [22]. Apart from linkage with the putative PAI, not much is known about the gene content of this subset and whether, based on gene content, a similar population can be characterized.

Microarray-based comparative genomic hybridization (CGH) has provided novel insights into the diversity and adaptability of several bacterial populations, such as the relevance of lateral gene transfer and recombination, which both result in mosaic genome structures in *Helicobacter pylori* [23,24], *Salmonella* species [25], *Escherichia coli* [26,27] and *S. aureus* [28–30]. In addition, CGH has been used to study evolution and to decipher bacterial virulence and host specificity [31–33]. In this respect, CGH has major advances over more conventional genotyping methods, as it also provides insights into the core genome and accessory genes, which may help to further disclose gross signatures of niche differentiation. Almost all CGH studies originate from PCR-based arrays of amplified open reading frames (ORFs) derived from one or multiple annotated sequenced strains, sometimes completed with additional genes not present in the sequenced strains. This approach, though expensive, ascertains coverage of a whole genome. Unfortunately, this approach is not possible for *E. faecium,* as there is no complete annotated genome sequence. Moreover, the partially sequenced, but still not annotated, *E. faecium* DO strain doesn't contain the putative PAI, one of the few known gene clusters associated with virulence and epidemicity. For all these reasons, a different approach is necessary for a broad and detailed genomic analysis of *E. faecium*.

In the present study, we performed comparative phylogenomics to study the genome composition population-wide as well as population dynamics of *E. faecium* using a mixed whole-genome array constructed from a shotgun library of nine strains from different ecological and genetic backgrounds, including the sequenced *E. faecium* DO strain. DNA–DNA hybridizations of 97 epidemiologically and genotypically different isolates to the array identified a distinct, globally dispersed clade containing all epidemic isolates and the majority of clinical isolates. Isolates within this clade harbored a large content of accessory genes mainly concentrated on two contigs in *E. faecium* DO. Furthermore, hybridization data revealed high rates of recombination and deletion resulting in mosaic-structured genomic regions. Insertion sequence (IS) elements were predicted to be prominent loci in the bifurcation of this clade, and probably have played a major role in adaptation and diversification of hospital-associated *E. faecium* strains belonging to this clade.

Results

Array Evaluation

In total, 3,474 spots (n(genomic[g]) = 2,727 , n(plasmid[p]) = 712, n(extra spotted genes) = 35) met the quality criteria (see Materials and Methods) and were included in this study. Since the microarray consists of a mixture of nine strains, only the genomic coverage of core genes can be determined. The total nucleotide and gene detection coverage of the *E. faecium* core genome was estimated to be 80% and 93%, respectively, using algorithms by Akopyants et al. and Moore et al., respectively [34,35]. Obviously, the microarray coverage of strain-unique accessory genes will be lower.

The detected Cy5/Cy3 ratios of the 3,474 spots were subjected to log2 transformation and GACK normalization (see Materials and Methods). Duplo hybridizations of seven isolates clustered as nearest neighbors in hierarchical clustering, each with 92% identical GACK values (98% for binary output) in spot profile, indicating that microarray results were highly reproducible. Log2-transformed data is available in Dataset S1. Validation of six ORFs located on ten inserts by Southern hybridization followed the presence and absence of spots-transformed array data (unpublished data; see Materials and Methods).

Among a selection of the accessory genome of 437 clone inserts, a subselection of 146 inserts (explained elsewhere in this section) was partially sequenced for further analysis. Sequence alignments revealed 16 redundant genes from 42 inserts and seven redundant plasmid loci from 27 inserts (represented by >100 base pair–overlap in two to four spots).

Eight of these 16 genes (27 of 42 inserts) were present in multiple copies in the *E. faecium* DO genome; therefore, actual redundancy among the libraries is limited.

Comparative Phylogenomic Analysis Based on Origin

Phylogenomic analysis with the microarray data using a Bayesian-based phylogenic method identified a distinct clade (Bayesian posterior probabilities [PP] = 1.0) containing all epidemic isolates ($n = 18$), 63% of clinical isolates ($n = 22/35$), 33% of hospital surveillance isolates ($n = 5/15$), no community survey isolates, 7% ($n = 1/15$) of animal isolates, and 0% ($n = 0/3$) of environmental isolates (Figure 1 and Table 1). *E. faecium* DO (E1794 in Table 1) is also contained in this clade. Throughout the rest of the article this clade is referred to as *hospital clade,* and hospital-associated *E. faecium* strains belonging to this clade are referred to as *hospital clade strains*. The bifurcation was supported by complete linkage transversal hierarchical clustering with GACK-transformed data of graded output, and by maximum parsimony analysis on binary GACK-transformed data (Table 1). With the latter two techniques, only one isolate was clustered differently. Internal branching within and outside this specific clade was less reliable (mostly PP = 0.50). The identification of a distinct hospital clade indicates the successful evolution of an *E. faecium* clone that adapted to its niche and diversified.

Gene Composition of the *E. faecium* Pangenome and Identification of Genes Specific to the *E. faecium* Hospital Clade

The *E. faecium* core genome defined as genomic spots present or divergent (GACK > -0.50) in each of the strains consisted of 65% of all genomic spots ($n(g) = 1,772$) (Figure 2). The clone inserts of 35 randomly selected spots were PCR-amplified, partially sequenced, and blasted to GenBank; these inserts encoded 37 (partial) genes (Table S1). Twenty-seven inserts showed highest similarity with 30 different *E. faecium* DO genes located on different contigs, two with two different *E. faecium* DO sequences with no corresponding ORF, and six with 12 *E. faecalis* V583 genes (Table S1). Because of array design (random shearing), more than one (partial) gene could be located on one insert. Assigned functions by clusters of orthologous genes (COGs) defined genes to be involved in basic cell function (Table S1).

Among all other spots ($n(g) = 955$, $n(p) = 712$, n(extra spotted genes) = 35) representing the accessory genome, 437 spots ($n(g) = 165$, $n(p) = 261$, n(extra spotted genes) = 11) were $\geq 80\%$ specific for and significantly associated with the hospital clade (χ^2 test followed by false discovery rate [FDR] correction ($p < 0.01$)) (Figure 2). The sensitivities for presence of these spots in clade-specific strains varied from 20% to 98%, indicating that some spots were present in almost all isolates belonging to the hospital clade, while other inserts were only present in a small subset (Table 2). The inserts from a selection of 146 spots ($n(g) = 86$, $n(p) = 60$) (criteria in Materials and Methods) were partially sequenced and blasted in GenBank for significant similarity. Sequencing revealed 175 ORFs with varying similarity to genes present in the *E. faecium* DO genome (131 ORFs located on 104 inserts), on *E. faecium* plasmids (21 ORFs on 19 inserts), in *E. faecalis* V583 (six ORFs on eight inserts), and in other bacterial species (17 ORFs) (see Table 2). Nine sequences showed no significant similarity at all (Table 2). Furthermore, 11 separately spotted PCR products were identified as hospital clade–specific. Hospital clade–specific sequences were identical or similar to genes from 13 different COGs (Table 2). By far the largest COG, group L consists of 80 ORFs (46% of 175 ORFs) and contains genes involved in DNA replication, recombination, and repair. This group mainly comprises IS elements and transposases ($n = 42$) and plasmid DNA sequences ($n = 26$). COG groups R and S, representing genes with a general function prediction and unknown function, respectively, are the second most prominent COG groups and include 55 (31%) of the ORFs. ORFs identical or similar to genes encoding metabolic pathways (COG G, E, F, H, P), and to proteins involved in cell wall and membrane biogenesis (M) and transcription (K), occurred less frequently (17, 9, and 12 times, respectively). Among all hospital clade–enriched ORFs, eight inserts represented five different antibiotic resistance genes (streptomycine adenylyltransferase; aminoglycoside phosphotransferase, which is similar to aph(3′)-III; chloramphenicol O-acetyltransferase; an aminoglycoside-streptothricin resistance cassette [*aadE* and *sat4* from the *aadE-sat4-aphA* cluster], and an aminoglycoside resistance cassette [*aac(6′)-Ie-aph(2″)-Ia* and *aac(6′)-Ie-aph(2″)-Ia2*]). Eleven ORFs were identical or highly similar to six different (putative) phage genes. Thirty-four of the 261 ORFs that originated from the plasmid library (58%) were similar to gene sequences on two enterococcal plasmids, pEFNP1 and pKQ10.

At least three different hybridization patterns could be recognized among hierarchical clustering of these plasmid-specific inserts, indicating existence of at least three different pKQ10/pEFNP1 variants.

Recombination and Genomic Mosaicism

Four different approaches were used to identify recombination in the hospital clade strains. First, we studied patterns of presence and absence of hospital clade–specific genes within gene clusters on the bacterial chromosome. Second, reticular networks were identified with split decomposition analysis (SDA). Third and fourth, guanine cytosine (GC) content and the Codon Adaptation Index (CAI) of hospital clade–specific genes were determined.

Overall, 94 of the partially sequenced inserts that were hospital clade–specific were identical or similar to *E. faecium* DO genes located on 23 to 27 different contigs. Uncertainty in the number of contigs was explained by the presence of multiple copies of the same genes on different contigs, which are predominantly transposases. Most hospital clade–specific genes were dispersed over the different contigs (mostly one gene per contig), but three contigs evidently represented hotspots for hospital–clade specific genes. Twenty-five genes with similarity to the hospital clade–specific ORFs were located on *E. faecium* DO contig 658, 16 genes were located on contig 656, and nine on contig 653. Clustering of hospital clade–specific accessory genes on the genome may indicate that these gene clusters have been acquired through horizontal gene transfer and recombination.

Phylogenetic analysis of hospital clade–specific genes within contigs 658 and 656 using SDA revealed networked structures consisting of nine parallelograms in contig 656 (n(strains) = 13) and eleven parallelograms in contig 658 (n(strains) = 9), with reasonable bootstrap values for contig 656 (99.9, 96.3, 96.2, 86.7, 86.5, 65.0, 64.7, 63.4, and 23.9), high bootstrap values for contig 658 (100), and high fit for both

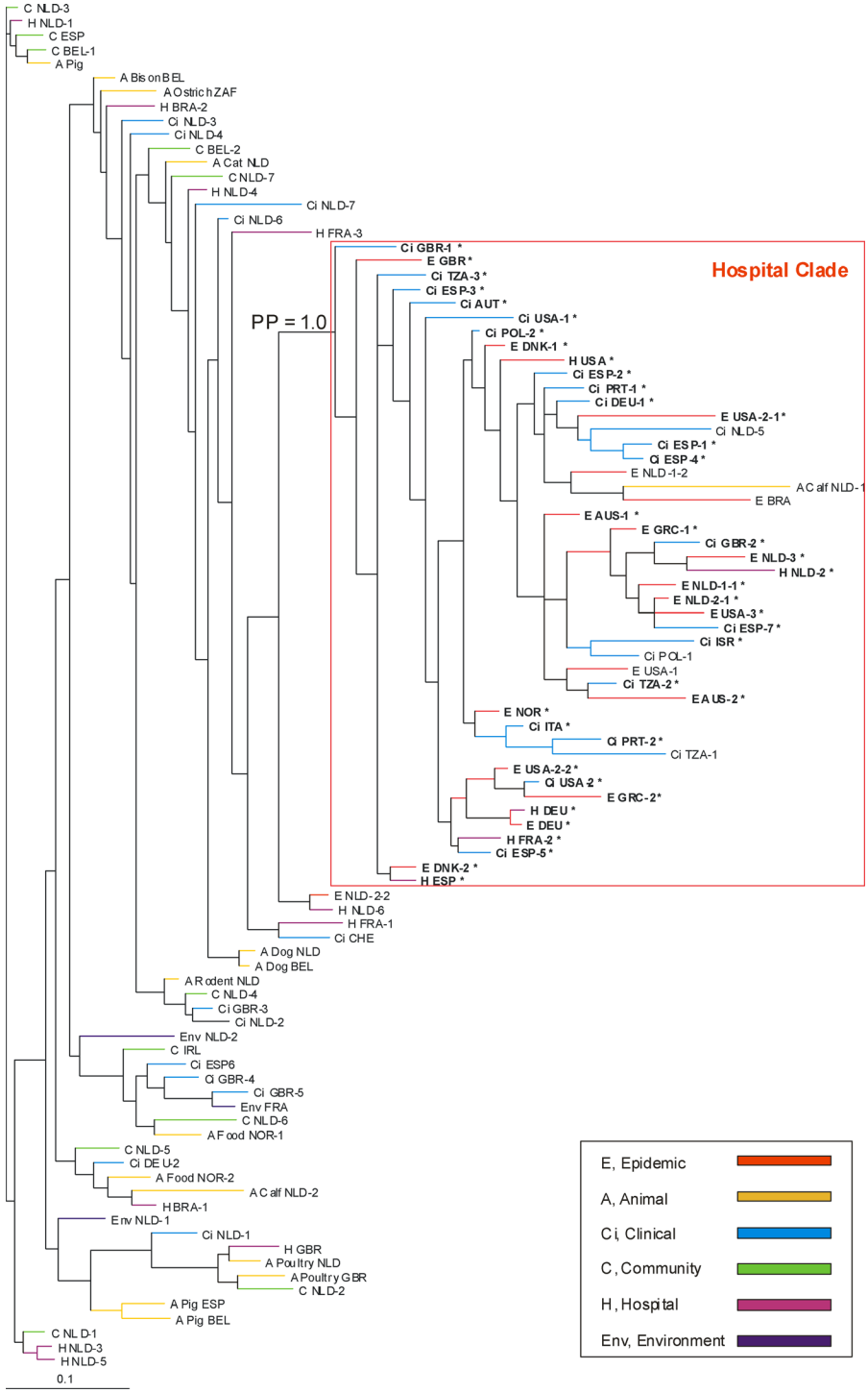

Figure 1. A Bayesian Phylogenomic Relationship of Strains Associated with Different Ecological Niches

The Bayesian posterior probability (PP) supports internal branch robustness. PP = 1.0 represents 100% of all phylogenies showing a given topology. Strains are designated at the end of branches and are colored according to the ecological niche from which the *E. faecium* strain was isolated. Strains in bold indicated with an asterisk are part of CC17.

AUS, Australia; AUT, Austria; BEL, Belgium; BRA, Brazil; CHE, Switzerland; DEU, Germany; DNK, Denmark; ESP, Spain; FRA, France; GBR, Great Britain; GRC, Greece; IRL, Ireland; ISR, Israel; ITA, Italy; NLD, Netherlands; NOR, Norway; POL, Poland; PRT, Portugal; TZA, Tanzania; USA, United States of America; ZAF, Republic of South Africa.

doi:10.1371/journal.ppat.0030007.g001

contigs (100) (Figure 3). These findings demonstrate the frequent occurrence of recombination events. Several clade-specific genes showed a different GC content than the 36% to 40% found in the rest of the genome (mean 37.9%; http://genome.ornl.gov/microbial/efae): the GC content of five of 16 genes within contig 656 ranged from 34% to 43%, whereas for contig 658, the GC content for eight of 25 genes ranged from 27% to 42% (Figure 4). These findings suggest foreign

origin of genes acquired through lateral gene transfer. Inserts that map on the same contig grouped in different clusters based on hierarchical clustering (unpublished data). This indicates that physically linked genes were differentially present in different *E. faecium* isolates, which is also highly suggestive of genomic mosaicism and recombination events. This is further illustrated in more detail for contigs 658 and 656 in Figures 3 and 4, respectively.

Besides genomic mosaicism, difference in codon usage can support foreign acquisition of genes. The mean CAI in *E. faecium* core genes was 0.65 (95% confidence interval: 0.62 and 0.68). Five of these genes were all located on contig 595. CAI values of all genes on contig 595 were calculated in comparison with the calculated mean CAI of the *E. faecium* core, since the whole contig probably contains only *E. faecium* DO core genes. CAI values were even significantly higher than the *E. faecium* core CAI ($p = 0.001$, *t*-test), indicating that the calculated mean CAI based on a small number of core genes might be underestimated (Figure S1 and Table S2). Nevertheless, CAI distribution among the contig 656 hospital clade–associated genes and contig 658 hospital-associated genes was significantly lower than mean CAI core genes ($p < 0.001$, *t*-test). However, local variations in CAI in hospital clade–specific gene clusters were substantial (Figure S1). This is exemplified by the CAI of genes belonging to the previously identified putative PAI (Table S3). In this island, the CAI of the *esp* gene (0.71) was higher than the CAI of the other genes (CAI: 0.52–0.62) (Table S3). In conclusion, CAI differences between genes expressed at high and low level as described for *E. coli* [36] were less pronounced. The relatively high CAI value of *E. faecium* core genes suggests that the codon usage of *E. faecium* is shaped towards optimal codon usage irrespective of cellular demands, but that the translational apparatus of the bacterium handled (part of) recruited DNA less efficiently than core DNA. Deviating CAI values and GC percentages of single genes implicate relatively recent acquisition through lateral gene transfer. In summary, these results in *E. faecium* DO support recent acquisition and recombination of accessory DNA, as defined in this study.

Character Evolution: Identification of Genes Specific to the Hospital Clade

Three inserts, all identical to the IS*16* transposase gene (mostly annotated in *E. faecium* DO as mutator-like transposase), were identified with 98% sensitivity and 100% specificity and validated by Southern blotting as the most clade-predictive locus present in the hospital clade (Figure 5 and Table S4). Multiple copies of this transposase were present in *E. faecium* DO, though not always likewise annotated. Two complete copies were present in contigs 658 and 646 with 100% and 97% sequence similarity, respectively. Contigs 630, 625, and 546 contained only the right-end side, and contigs 654 and 613 only the left-end side of the gene. IS*16,* part of the IS*256* family, is flanked by nonidentical inverted repeats, with the right inverted repeat resembling a −35 promoter region [37,38]. Sequences of 14 high-ranking hospital clade predictive inserts (≥94% predicted presence and ≤5% presence in the hospital and non–hospital clade ancestral strain, respectively) revealed ≥98% similarity with genes encoding a transposase belonging to the IS*30* family (*tra8* gene) (annotation *E. faecium* DO: integrase with catalytic region) (n(inserts) = 7); an extracellular solute-binding protein, a glycosyl hydrolase, and a conserved hypothetical protein (which are all located on contig 638); chloramphenicol O-acetyltransferase; ROK (Repressor, ORF, Kinase); a phage terminase; and a phage portal protein (Table S4).

Apart from the hospital clade–unique inserts, IS elements were also prominent among hospital clade–enriched inserts. Approximately 30% of all sequenced inserts were similar to genes encoding five additional different types of transposases/IS elements (transposase IS*IIIA*/IS*1328*/IS*1533*; transposase IS*110*/IS*116*/IS*902,* transposase IS*3*/IS*911,* transposase IS*256,* transposase IS*204*/IS*1001*/IS*1096*/IS*1165,* and transposase IS*66*).

Character tracing predicted that the hospital clade acquired certain genes, like the putative PAI, only after initial branching. Absence of the variant *esp* gene in the ancestor of the hospital clade was 98% likely. Other genes acquired after development of the hospital clade include plasmid-derived genes, membrane proteins ($n = 6$), genes involved with carbohydrate transport and metabolism ($n = 7$), transcription-related genes ($n = 6$), defense mechanism genes, aminoglycoside resistance cassettes, and several solitary genes ($n = 5$) not belonging to the same COG.

Hospital Clade Genome Rearrangements

The observation that IS elements were most specific and abundant for the hospital clade suggests that the acquisition of IS elements increased genome plasticity and the propensity of acquiring further adaptive mechanisms, thus facilitating adaptation to the hospital environment. In general, the genetic variability of isolates that are evolutionary-linked, e.g., the hospital clade isolates, is expected to be less than the genetic variability of isolates that belong to different evolutionary lineages, like all the different non–hospital clade isolates. This difference, however, can be mitigated if specific mechanisms, like the enrichment of IS element, enhance the genetic variability of hospital clade isolates. To compare genetic variation between strains within the hospital clade to variation between all other strains, the genetic similarity among isolates was determined by pulsed-field gel electrophoresis (PFGE), the outcome of which is affected by genome rearrangements, and MLST, which is not influenced by genome rearrangements. As expected, the average genetic similarities among 21 evolutionary-linked hospital clade isolates, based on MLST, was higher (60%) than that among 23 non–hospital clade isolates (26.7%) (Tables S5 and S6). However, the PFGE-based average genetic similarity among the hospital clade strains was 53.6 % ± 13.2 standard deviation (SD), comparable to the average genetic similarity among the non–hospital clade strains (54.9 % ± 12.8 SD; not significant, *t*-test) (Table S7). The resulting recombination/diversity indices of 1.12 ± 0.72 SD for the hospital clade and 0.51 ± 0.50 SD for the non–hospital clade strains supported frequent genome rearrangements in the hospital clade ($p < 0.001$; *t*-test).

Discussion

Using a mixed whole-genome microarray, we have identified with three different phylogenetic algorithms—Bayesian-based phylogenic analysis, maximum parsimony analysis, and hierarchical clustering—a globally dispersed *E. faecium* clade

Table 1. Strains Used for Hybridization

Category	Epidemiology	Country	Strain Identity Code	Year	MLST		PAI[a]	vanR[a]	ampR[a]	Strain	Bayes	Max Pars	Hier Clust
					CC	ST							
Epidemic ($n = 18$)	E	AUS	1	1998	17	17	1	1	1	E0510	H	H	H
	E	AUS	2	2000	17	173	0	1	0	E1760	H	H	H
	E	BRA		1998	non 17	114	1	1	1	E1679	H	H	H
	E	DEU		2002	17	78	0	1	1	E1644	H	H	H
	E	DNK	1	unknown	17	17	1	0	1	E1716	H	H	H
	E	DNK	2	unknown	17	18	0	0	1	E1717	H	H	H
	E	GRC	1	2000	17	16	1	1	0	E1441	H	H	H
	E	GRC	2	1999	17	65	1	1	0	E1435	H	H	H
	E	NLD	1–1	2000	17	16	1	1	0	E0734	H	H	H
	E	NLD	1–2	2002	17	18	0	1	1	E1652	H	H	H
	E	NLD	2–1	2000	17	16	1	1	1	E0745	H	H	H
	E	NLD	3	1998–1999	17	16	1	1	0	E0470	H	H	H
	E	NOR		1999	17	17	0	0	0	E1340	H	H	H
	E	GBR		1992	17	18	1	1	1	E0013	H	H	H
	E	USA	1	1994	non 17	20	1	1	1	E0300	H	H	H
	E	USA	2–1	1995	17	17	1	1	1	E0155	H	H	H
	E	USA	2–2	1995	17	16	1	1	0	E0161	H	H	H
	E	USA	3	2001	17	16	1	1	0	E1132	H	H	H
Clinical ($n = 35$)	Ci	AUT		1998	17	78	1	0	0	E1263	H	H	H
	Ci	CHE		1997	non 17	25	0	0	1	E1250			
	Ci	DEU	1	1998	non 17	130	0	0	1	E1283			
	Ci	DEU	2	1998	17	17	0	0	1	E1284	H	H	H
	Ci	ESP	1	1995	17	18	0	0	0	E1734	H	H	H
	Ci	ESP	2	1997	17	18	1	0	1	E1467	H	H	H
	Ci	ESP	3	1997	17	18	0	0	0	E1500	H	H	H
	Ci	ESP	4	1997	17	18	0	0	0	E1737	H	H	H
	Ci	ESP	5	1998	17	17	1	1	0	E1463	H	H	H
	Ci	ESP	6	1999	non 17	74	0	0	1	E1499			
	Ci	ESP	7	2001	17	16	1	0	1	E1735	H	H	H
	Ci	GBR	1	1997	17	17	0	1	0	E0380	H	H	H
	Ci	GBR	2	2000	17	16	1	0	0	E1391	H	H	H
	Ci	GBR	3	2000	non 17	84	0	0	0	E1403			
	Ci	GBR	4	2000	non 17	100	0	0	1	E1421			
	Ci	GBR	5	2000	non 17	94	0	0	0	E1423			
	Ci	ISR		1997	17	80	0	1	1	E0333	H	H	H
	Ci	ITA		1997	17	17	1	0	0	E1292	H	H	H
	Ci	NLD	1	1950	non 17	67	0	0	0	E1620			
	Ci	NLD	2	1959	non 17	86	0	0	0	E1621			
	Ci	NLD	3	1960	non 17	22	0	0	0	E1623			
	Ci	NLD	4	1961	non 17	22	0	0	0	E1625			
	Ci	NLD	5	1961	non 17	106	0	0	1	E1636	H	H	H
	Ci	NLD	6	1995	non 17	22	0	1	0	E0073			
	Ci	NLD	7	1995	non 17	21	0	1	1	E0125			
	Ci	NLD	2–2	2000	non 17	50	0	1	0	E0772			
	Ci	POL	1	1998	non 17	99	1	0	1	E1172	H	H	H
	Ci	POL	2	1998	17	17	1	0	0	E1302	H	H	H
	Ci	PRT	1	1998	17	132	0	0	0	E1307	H	H	H
	Ci	PRT	2	1998	17	17	1	0	0	E1308	H	H	H
	Ci	TZA	1	unknown	non 17	169	0	0	0	E1721	H	H	H
	Ci	TZA	2	unknown	17	132	0	0	1	E1728	H	H	H
	Ci	TZA	3	unknown	17	18	0	0	0	E1731	H	H	H
	Ci	USA	1	1991	17	18	0	0	1	E1794	H	H	H
	Ci	USA	2	2001	17	16	1	0	0	E1360	H	H	H
Hospital ($n = 15$)	HS	BRA	1	2000	non 17	110	0	0	1	E1674			
	HS	BRA	2	2001	non 17	111	0	0	0	E1675			
	HS	DEU		2002	17	78	1	1	1	E1643	H	H	H
	HS	ESP		2001	17	18			0	E1850	H	H	H
	HS	FRA	1	1986	non 17	25	0	1	1	E0005			
	HS	FRA	2	1997	17	17	0	1	1	E0321	H	H	H
	HS	FRA	3	1997	non 17	79	0	1	1	E0322			
	HS	GBR		1992	non 17	146	0	1	1	E0027			
	HS	NLD	1	1995	non 17	6	0	1	0	E1149			
	HS	NLD	2	1998	17	16	1	1	0	E1147	H	H	H
	HS	NLD	3	2000	non 17	5	0	1	1	E0802			
	HS	NLD	4	2000	non 17	21	0	1	0	E0849			
	HS	NLD	5	2000	non 17	5	0	1	1	E1316			
	HS	NLD	6	2002	non 17	50	0	1	0	E1554			

Table 1. Continued.

Category	Epidemiology	Country	Strain Identity Code	Year	MLST		PAI[a]	vanR[a]	ampR[a]	Strain	Bayes	Max Pars	Hier Clust
					CC	ST							
	HS	USA		2001	17	117	1	1	0	E1133	H	H	H
Community ($n = 11$)	C	BEL	1	1996	non 17	6	0	1	1	E1764			
	C	BEL	2	1996	non 17	136	0	1	0	E1766			
	C	ESP		2000	non 17	101	0	1	1	E1485			
	C	IRL		2001	non 17	163	0	0	0	E1590			
	C	NLD	1	1996	non 17	147	0	1	1	E0060			
	C	NLD	2	1996	non 17	82	0	1	1	E0128			
	C	NLD	3	1996	non 17	6	0	1	0	E0135			
	C	NLD	4	1998	non 17	54	0	0	1	E1002			
	C	NLD	5	1998	non 17	42	0	0		E1039			
	C	NLD	6	1999	non 17	94	0	0	1	E0980			
	C	NLD	7	2000	non 17	32	0	1	0	E1071			
Environment ($n = 3$)	Env	FRA		1985	non 17	172	0	0	1	E1759			
	Env	NLD	1	1981	non 17	68	0	0	0	E1628			
	Env	NLD	2	1981	non 17	69	0	0	1	E1630			
Animal ($n = 15$)	A Bison	BEL		1994	non 17	21	0	0	0	E1573			
	A Calf	NLD	1	1996	non 17	1	0	1	1	E0172	H		
	A Calf	NLD	2	1996	non 17	4	0	1	1	E0211			
	A Cat	NLD		1996	non 17	21	0	1	0	E0466			
	A Dog	BEL		1995	non 17	27	0	1	0	E1574			
	A Dog	NLD		1996	non 17	27	0	1	0	E0463			
	A Food	NOR	1	1956	non 17	76	0	0	1	E1607			
	A Food	NOR	2	1964	non 17	70	0	0	0	E1619			
	A Ostrich	ZAF		2001	non 17	159	0	0	1	E1576			
	A Pig	BEL		2001	non 17	6	0	0	0	E1781			
	A Pig	ESP		unknown	non 17	137	0	0	1	E0685			
	A Pig	NLD		1996	non 17	6	0	1	0	E0144			
	A Poultry	GBR		1992	non 17	9	0	1	1	E0045			
	A Poultry	NLD		1997	non 17	8	0	1	0	E0429			
	A Rodent	NLD		1959	non 17	104	0	0	1	E1622			

[a]1 indicates presence, 0 indicates absence.

A, animal; ampR, ampicillin resistance; AUS, Australia; AUT, Austria; Bayes, Bayesian analysis; BEL, Belgium; BRA, Brazil; C, community surveillance; CHE, Switzerland; Ci, clinical; DEU, Germany; DNK, Denmark; E, epidemic; Env, environmental; ESP, Spain; FRA, France; GBR, Great Britain; GRC, Greece; H, hospital clade; Hier Clust, hierarchical clustering; HS, hospital surveillance; IRL, Ireland; ISR, Israel; ITA, Italy; Max Pars, maximum parsimony analysis; NLD, Netherlands; NOR, Norway; POL, Poland; PRT, Portugal; ST, sequence type; TZA, Tanzania; USA, United States of America; vanR, vancomycin resistance; ZAF, Republic of South Africa.

doi:10.1371/journal.ppat.0030007.t001

among an epidemiologically well-characterized strain collection. This clade is highly specific for nosocomial outbreaks and infections, and 146 clade-specific genes were identified, which were located scattered over 23 to 27 different contigs. Three contigs appeared to be hotspots of these clade-specific genes and were characterized by extensive genomic mosaicim. CAI values and GC content of hospital clade–specific genes on these contigs were slightly different from the rest of the genome. This may indicate rapid adoption of these genes to the translational apparatus of *E. faecium* or acquisition from closely related species. Among hospital clade–specific genes, IS elements were identified as the most predictive loci for this clade. Besides lateral gene transfer, these IS elements might have facilitated extensive genome rearrangements in the hospital clade as shown by PFGE results, a technique used previously to demonstrate heterogeneity in location of transposon-mediated transconjugation in enterococci [39]. These genomic events could have contributed to the transition of an avirulent commensal to a nosocomial pathogenic *E. faecium* subspecies.

A random shotgun library of nine *E. faecium* genomes, selected upon different MLST profiles, was used to investigate population dynamics and genome content of 97 *E. faecium* isolates. In the absence of a finalized annotation of an *E. faecium* genome, mixed whole-genome microarray technology offers the optimal tool for comparative genomics. Nevertheless, data interpretation is a potential limitation of microarrays generated upon cloned random fragments rather than gene-specific primers, since multiple gene fragments may be present per insert [40]. This technical restriction stresses the need for validation. Since confirmatory hybridizations matched array data, we consider our microarray suitable for genomic comparisons of isolates.

MLST of *E. faecium* previously identified host specificity and a globally dispersed subpopulation named CC17, which was associated with hospital outbreaks and infections, and which had apparently replaced the more heterogeneous enterococcal bacterial population within hospitals [22,41,42]. Intriguingly, our microarray phylogenetic analyses, based on genome-wide presence and absence of genes, was comparable to our findings obtained by MLST. Two outbreak-related strains, which were not considered part of CC17 by MLST, clearly belonged to the hospital clade in CGH. Although these strains were evolutionary unrelated to CC17 based on allelic profiles of housekeeping genes, they probably acquired hospital clade–specific genes by horizontal gene transfer.

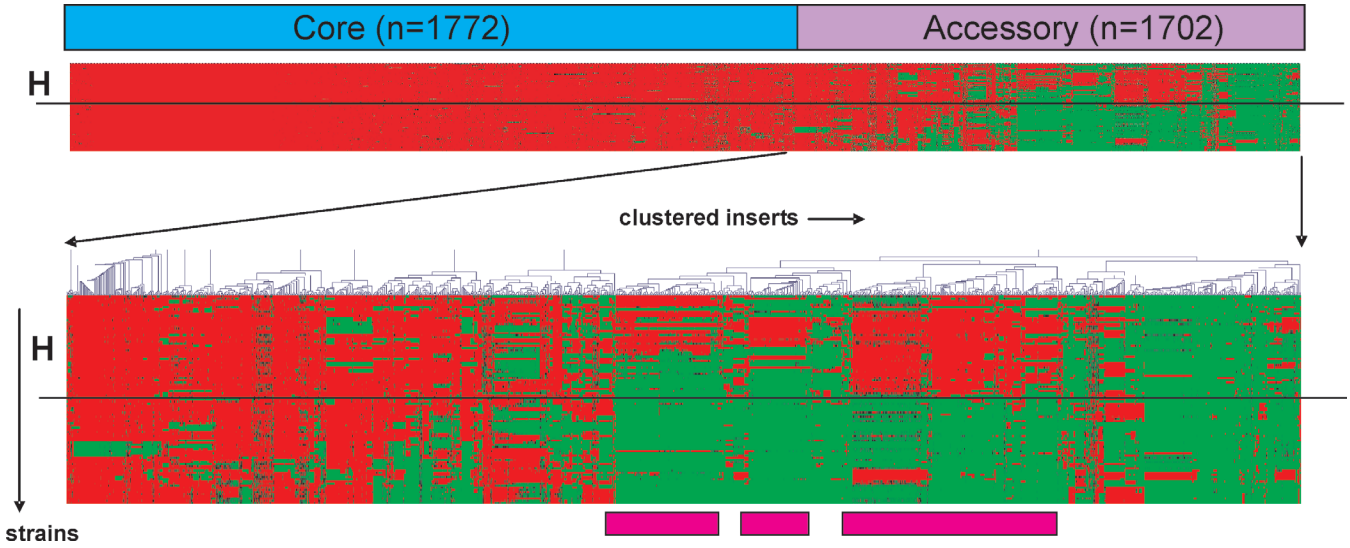

Figure 2. Complete Linkage Hierarchical Clustering of *E. faecium* Inserts with Pearson Correlation and Euclidian Distance, Zoom on Accessory Insert

The hierarchical clustering of inserts visualizes composition of the microbial pangenome: the core genome and accessory genes. The lower panel of the figure represents a zoomed-in accessory genome, as indicated by the vertical and diagonal arrows in the middle of the figure. Presence of an insert is indicated in red and absence of an insert in green. Inserts are clustered on the *x*-axis, and strains are presented on the *y*-axis (indicated by a vertical arrow left of the lower panel). Horizontal straight lines separate the hospital clade strains (H) from the non-hospital clade strains. The core genome and accessory genes are indicated by a blue and purple bar, respectively. Hospital clade–associated clustered inserts are indicated with pink bars.

doi:10.1371/journal.ppat.0030007.g002

Other *E. faecium* MLST clonal complexes were not identified by CGH, probably reflecting high recombination frequencies. Congruence between array distance trees and MLST trees has also been described for *S. aureus* and *Neisseria meningitidis* [43,44].

IS elements contributed prominently to the hospital clade–specific genes. In general, mobile genetic elements, such as ISs, transposons, phages, and genomic islands, are common components of microbial genomes and are driving forces for novel genotypic and phenotypic variants. Transposition of IS elements may disrupt genes, but may also activate downstream genes [37] and fine tune gene expression through transposition mediated genome inversions. Some IS elements contain a −35 promoter-like sequence in their terminal, which may result in formation of a functional promoter [37,38]. Mobile elements are often flanked by IS elements which facilitate recombination and mobilization. The most prominent marker indicative for the hospital clade was IS*16,* which is present in multiple copies on different contigs of the *E. faecium* DO genome. In the unfinished annotation of *E. faecium* DO, IS*16* is designated as a mutator type transposase, sharing high similarity with similar tranposases in other bacterial species and in maize [45]. This IS element seems to possess all transposing capacities: IS*16* (*i*) was already identified in *E. faecalis* as flanking part of Tn*1547* [46], (*ii*) inserted in the *vanY* gene of Tn*1546,* resulting in a VanD-like phenotype in a *vanA* genotype vancomycin-resistant *E. faecium* [47], and (*iii*) contains a −35 promoter-like sequence. In addition, IS*16* was found in multiple copies in several enterococcal strains [48] and in pRUM, an *E. faecium* plasmid containing a postsegregational killing system [49]. Our findings demonstrate that IS*16* can be used as a specific marker of the hospital clade genotype.

Enrichment of particular IS elements in the genome of bacterial (sub)species has been documented previously. In *S. epidermidis,* IS*256* is present in multiple copies in clinical strains, where it might induce genome flexibility of multiresistant, biofilm-forming isolates [33]. *Shigella* species are enriched with 300 to 700 copies of IS elements [50], and *Bordetella pertussis* is significantly enriched in IS elements compared with *Bordetella bronchiseptica,* from which *B. pertussis* is thought to have been evolved [51,52]. The observation that IS elements are among the most specific and abundant hospital clade–specific inserts suggests that acquisition of these elements has contributed to the ecological success of this clade in the hospital environment, through enhanced genome plasticity. A higher diversity index for the hospital clade compared with that of the non–hospital clade demonstrates higher levels of genetic variability, which could have resulted from enrichment with IS elements. IS elements may not only affect gene expression and enhance genome plasticity, but may also increase the propensity of acquiring further adaptive mechanisms. All of these IS element–induced events may have been pivotal for *E. faecium* to adapt and survive in highly competitive niches, such as the hospital environment.

The structure and genetic makeup of the *E. faecium* pangenome shows many similarities with the sequenced *E. faecalis* V583 [53]. In concordance with the more than 25% mobile elements in V583, 50 (putative) IS elements were contained among the 146 identified *E. faecium* hospital clade–specific genes. One of the most prominent V583 IS elements, IS*256,* was also found among these. V583 plasmids seem to be complex mosaic structures compared with similar plasmids in GenBank [53]. The presence of multiple plasmid remnants and the three resident plasmids in V583 was hypothesized as being important for genome plasticity [53]. Our data for contigs 656 and 658 provide more evidence for this concept in *E. faecium*.

Highly prevalent hospital clade–specific genes represent

Table 2. Sequenced *E. faecium* Hospital Clade–Specific Genes Sorted by Functional COG Groups

Category	Functional Group	Spot ID	Sensitivity	Specificity	FDR Rank	FDR	COG	Contig	EfaeDRAFT Gene Number	Gene	COG Description	Species	E-Value	Similarity (%)[a]
***E. faecium* DO genes**	**Information storage and processing**	EfmP.003D11	0.35	0.98	507	1.46E-03	K	587	584	Helix-turn-helix motif	COG1476		0	99
		Efm.036C07	0.67	0.9	249	7.17E-04	K	653	2193	BRO (prophage antirepressor)	COG3617		0	98
		Efm.029D04	0.8	0.8	242	6.97E-04	K	653	2195	Helix-turn-helix motif			0	99
		Efm.033D03	0.74	0.82	293	8.43E-04	K	653	2195	Helix-turn-helix motif			0	99
		Efm.032A05	0.8	0.84	205	5.90E-04	K	656, 557[b]	2498[b]	Conserved hypothetical protein (predicted transcriptional regulator)	COG2378			97
		Efm.030H12	0.3	0.98	554	1.59E-03	K	658	2711	Regulatory protein, GntR:Bacterial regulatory protein, GntR	PhnF		−10	{65 (77)}
		Efm.028E04	0.91	0.92	60	1.73E-04	L	573	441	Transposase, IS*204*/IS*1001*/IS*1096*/IS*1165* family	COG3464		0	99
		Efm.031G03	0.96	0.94	35	1.01E-04	L	573, 658, 616, 527, 655, 594, 637, 580[b]	441[b]	Transposase, IS*204*/IS*1001*/IS*1096*/IS*1165* family	COG3464			99
		Efm.023B01	0.46	0.84	655	1.89E-03	L	587	588	Relaxase/mobilization nuclease domain			0	99
		EfmP.009F01	0.26	1	544	1.57E-03	L	587	590	Integrase, catalytic region; IS*30* family	COG3316		−48	85
		EfmP.002H11	0.5	0.98	316	9.10E-04	L	587	592	Mobilization			−100	90
		Efm.021E07	0.63	0.8	496	1.43E-03	L	590, 630, 658, 613, 562, 545, 547, 638[b]	633[b]	Integrase, catalytic region; IS*30* family	Tra8		0	98
		Efm.036C08	0.54	0.98	251	7.23E-04	L	609	901	Phage integrase	XerD		0	99
		EfmP.004B11	0.85	0.88	105	3.02E-04	L	624	1151	Integrase, catalytic region; IS*30* family	Tra8		0	99
		Efm.002F08	0.54	1	231	6.65E-04	L	630	1268	Transposase IS*4;* IS*4* family	COG3385		0	96
		EfmP.009H08	0.91	0.8	114	3.28E-04	L	636, 614, 613[b]	1413[b]	Integrase, catalytic region; IS*30* family	Tra8		0	99
		Efm.024F11	0.89	0.86	84	2.42E-04	L	636	1413	Integrase, catalytic region; IS*30* family	Tra8		0	99
		Efm.023D12	0.72	0.96	128	3.68E-04	L	637, 656, 501, 638, 658, 582, 541[b]	1465[b]	IS*66* ORF2-like; IS*66* transposase family	COG3436		0	98
		EfmP.008D12	0.52	0.94	390	1.12E-03	L	639	1502	Integrase, catalytic region; IS*30* family	COG3316		−135	99
		Efm.009G07	0.91	0.98	34	9.79E-05	L	646, 658, 590, 529, 599, 618, 594, 657[b]	1772[b]	Transposase, IS*204*/IS*1001*/IS*1096*/IS*1165* family	COG3464		0	99
		Efm.028F11	0.39	0.9	628	1.81E-03	L	649	1882	Transposase	COG0675		0	95
		EfmP.005D09	0.39	0.88	669	1.93E-03	L	649	1882	Transposase	COG0675		−89	100
		Efm.013B06	0.96	0.96	30	8.64E-05	L	653	2160	Transposase IS*3*/IS*911;* IS*3* family	COG2963		0	97
		EfmP.012C03	0.98	0.94	13	3.74E-05	L	654, 635, 614, 655, 658, 639, 570[b]	2261[b]	Transposase, IS*IIIA*/IS*1328*/IS*1533* family; transposase IS*110*/IS*116*/IS*902* family	COG3547			97
		EfmP.011C05	0.35	0.98	508	1.46E-03	L	654	2310	IS*16* (transposase, mutator type) (annotation: replication initiator factor); IS*256* family			−9	{25 (55)}
		Efm.019F12	0.98	1	2	5.76E-06	L	654	2310	IS*16* (transposase, mutator type) (annotation: replication initiator factor); IS*256* family			0	99
		Efm.036D10	0.76	0.8	295	8.49E-04	L	655	2316	Integrase, catalytic region; IS*30* family	Tra8		−79	100
		EfmP.012G10	0.85	0.86	123	3.54E-04	L	655, 656, 590, 606, 527, 658[b]	2316[b]	Integrase, catalytic region; IS*30* family			−180	89
		EfmP.001F10	0.91	0.96	40	1.15E-04	L	655, 599, 551, 648[b]	2344[b]	Integrase, catalytic region; IS*30* family	Tra5		−85	99
		Efm.006H03	0.65	0.96	190	5.47E-04	L	656	2435	C5 cytosin-specific DNA methylase	Dcm		0	98

Table 2. Continued.

Category	Functional Group	Spot ID	Sensitivity	Specificity	FDR Rank	FDR	COG	Contig	EfaeDRAFT Gene Number	Gene	COG Description	Species	E-Value	Similarity (%)[a]
		Efm.024B03	0.65	0.96	189	5.44E-04	L	656	2435	C5 cytosin-specific DNA methylase	Dcm		0	99
		Efm.006H03	0.65	0.96	190	5.47E-04	L	656	2437	Helix-turn-helix motif: replication initiation factor	COG2946		0	98
		Efm.011A04	0.67	0.96	174	5.01E-04	L	656	2437	Helix-turn-helix motif: replication initiation factor	COG2946		0	99
		Efm.011E10	0.67	0.96	173	4.98E-04	L	656	2437	Helix-turn-helix motif: replication initiation factor (putative phage replication protein RstA)	COG2946		0	100
		Efm.019B03	0.67	1	110	3.17E-04	L	656, 645, 610[b]	2438[b]	RNA-directed DNA polymerase (reverse transcriptase):HNH endonuclease	COG3344		0	99
		EfmP.006H07	0.78	0.88	168	4.84E-04	L	656	2449	Plasmid recombination enzyme			−7	{22 (39)}
		Efm.031C04	0.91	0.94	55	1.58E-04	L	656, 609[b]	2464[b]	Transposase IS*66;* IS*66* family	COG3436		0	99
		EfmP.010A12	0.74	0.88	215	6.19E-04	L	656, 590, 658, 655, 638, 647, 636, 614, 613, 630[b]	2485[b]	Integrase, catalytic region; IS*30* family	Tra8		0	99
		EfmP.011C03	0.87	0.86	97	2.79E-04	L	656, 647, 638, 636, 621, 614, 613, 606, 630, 562, 527[b]	2490[b]	Integrase, catalytic region; IS*30* family	Tra8		−71	98
		EfmP.010F07	0.87	0.86	98	2.82E-04	L	656, 647, 638, 636, 614, 613, 630[b]	2490[b]	Integrase, catalytic region; IS*30* family	Tra8		0	98
		Efm.022F03	0.89	0.92	64	1.84E-04	L	658	2597	Transposase, IS*204*/IS*1001*/IS*1096*/IS*1165* family	COG3464		0	97
		Efm.028A12	0.91	0.96	39	1.12E-04	L	658	2597	Transposase, IS*204*/IS*1001*/IS*1096*/IS*1165* family	COG3464		0	99
		EfmP.002G09	0.78	0.88	169	4.86E-04	L	658	2597	Transposase, IS*204*/IS*1001*/IS*1096*/IS*1165* family	COG3464		−80	98
		Efm.004F07	0.59	0.94	285	8.20E-04	L	658	2605	Integrase, catalytic region	COG3316		−67	98
		Efm.001E04	0.76	0.96	91	2.62E-04	L	658	2606	Replication initiator A, N-terminal				99
		Efm.019H06	0.85	0.8	193	5.56E-04	L	658	2610	DNA-directed DNA polymerase	DinP		0	98
		Efm.036F05	0.83	0.9	96	2.76E-04	L	658	2610	DNA-directed DNA polymerase	DinP			98
		Efm.004B11	0.98	1	3	8.64E-06	L	658, 625, 630	2633[b]	IS*16* (annotation: transposase, mutator type); IS*256* family	COG3328		0	100
		Efm.033B03	0.83	1	58	1.67E-04	L	658	2633	IS*16* (annotation: transposase, mutator type); IS*256* family	COG3328		0	99
		Efm.028B02	0.98	0.94	16	4.61E-05	L	658, 655, 654[b]	2641[b]	Transposase, IS*IIA*/IS*1328*/IS*1533* family; transposase IS*110*/IS*116*/IS*902* family	COG3547		0	98
		Efm.028D10	0.98	0.94	15	4.32E-05	L	658, 655, 654, 635[b]	2641[b]	Transposase, IS*IIA*/IS*1328*/IS*1533* family; transposase IS*110*/IS*116*/IS*902* family	COG3547		0	98
		EfmP.003D10	0.85	0.82	167	4.81E-04	L	658, 655, 656, 647, 636, 624, 614, 589, 531, 545[b]	2669[b]	Integrase, catalytic region; IS*30* family	Tra8		−128	99
		Efm.018E04	0.78	0.8	265	7.63E-04	L	658, 655, 656, 647, 638, 636, 590, 614, 613, 630[b]	2669[b]	Integrase, catalytic region; IS*30* family	Tra8		−89	98
		Efm.007H02	0.93	0.98	23	6.62E-05	L	658	2680	Transposase, IS*204*/IS*1001*/IS*1096*/IS*1165* family	COG3464			100

Table 2. Continued.

Category	Functional Group	Spot ID	Sensitivity	Specificity	FDR Rank	FDR	COG	Contig	EfaeDRAFT Gene Number	Gene	COG Description	Species	E-Value	Similarity (%)[a]
		EfmP.012A12	0.93	0.92	54	1.55E-04	L	658, 655, 637, 580, 616, 594, 573, 527[b]	2681[b]	Transposase, IS*204*/IS*1001*/IS*1096*/IS*1165* family	COG3464		−83	98
		EfmP.003D10	0.85	0.82	167	4.81E-04	L	658, 655, 656, 647, 636, 624, 614, 589, 531, 545[b]	2689[b]	Integrase, catalytic region; IS*30* family	Tra8		−145	99
		EfmP.009B02	0.85	0.82	165	4.75E-04	L	658, 655, 656, 636, 614, 589, 531[b]	2689[b]	Integrase, catalytic region; IS*30* family	Tra8		−112	100
		Efm.007E12	0.91	0.9	63	1.81E-04	L	658	2696	Transposase IS*3*/IS*911;* IS*3* family	COG2963			98
		Efm.035D02	0.85	0.86	124	3.57E-04	L	658	2716	Integrase, catalytic region; IS*30* family	Tra8		0	99
		EfmP.007E08	0.87	0.86	100	2.88E-04	L	658, 638, 590[b]	2716[b]	Integrase, catalytic region; IS*30* family	Tra8		0	99
		Efm.024E08	0.98	1	1	2.88E-06	L	646	1769 + 1770	IS*16* (transposase, mutator type) (annotation: hypothetical protein + IS*256*, transposase); IS*256* family			0	99
		Efm.031H06	0.91	0.94	57	1.64E-04	L	658, 653[b]	2691/2695[b]	Integrase, catalytic region; IS*30* family			0	99
	Cellular processing	Efm.024B03	0.65	0.96	189	5.44E-04	D	656	2434	Cell division FtsK/SpoIIIE protein	FtsK		0	99
		Efm.024B05	0.54	1	229	6.59E-04	V	582	531	Glyoxalase/bleomycin resistance protein/ dioxygenase			0	98
		EfmP.003D11	0.35	0.98	507	1.46E-03	V	587	585	Bacteriocin, lactococcin 972			0	99
		Efm.009H02	0.57	1	212	6.10E-04	M	630	1272	NLP/P60	Spr		0	96
		Efm.015D06	0.63	0.98	182	5.24E-04	M	630	1272	NLP/P60	Spr		−136	100
		Efm.015D06	0.63	0.98	182	5.24E-04	M	630	1273	Putative membrane protein			−136	100
		Efm.009F11	0.57	1	213	6.13E-04	M	656	2430	Surface protein from Gram-positive cocci, anchor region	COG4932		0	99
		Efm.020A11	0.74	0.98	86	2.48E-04	M	656	2430	Surface protein from Gram-positive cocci, anchor region	COG4932		0	99
		Efm.027E10	0.7	0.96	141	4.06E-04	M	656	2430	Surface protein from Gram-positive cocci, anchor region	COG4932		0	100
		Efm.032F07	0.7	0.96	140	4.03E-04	M	656	2430	Surface protein from Gram-positive cocci, anchor region	COG4932		0	99
	Metabolism	Efm.009B08	0.3	0.96	610	1.76E-03	G	501	52	Phosphoglucomutase/phosphomannomutase	ManB		0	99
		Efm.025H03	0.48	1	290	8.35E-04	G	581	526	Glycoside hydrolase family 1	BglB		0	99
		Efm.024B05	0.54	1	229	6.59E-04	G	582	532	Glycoside hydrolase family 32	SacC		0	98
		Efm.017C07	0.54	1	230	6.62E-04	G	582	540	PTS system, glucose-like IIB component	PtsG		0	95
		Efm.033C04	0.54	1	227	6.53E-04	G	582	541	Glycoside hydrolase family 32	SacC		0	97
		Efm.024F04	0.93	1	8	2.30E-05	G	638	1477	Glycosyl hydrolase family 88			0	98
		Efm.047B09	0.93	1	7	2.01E-05	G	638	1481	Extracellular solute-binding protein	UgpB		0	96
		Efm.024H08	0.39	1	409	1.18E-03	G	656	2468	PTS system mannose/fructose/sorbose family IId component	ManZ		0	99
		Efm.022A03	0.33	0.98	520	1.50E-03	G	658	2713	Tagatose 1,6-diphosphate aldolase, Gram positive	LacD		−91	{65 (77)}
		Efm.024F11	0.89	0.86	84	2.42E-04	G	658	2715	PTS lactose/cellobiose IIC component	CelB		−152	99
		Efm.019D08	0.74	0.94	133	3.83E-04	E	594	662	Conserved hypothetical protein, histidinol phosphatase and related hydrolases of the PHP family	HIS2		−24	{41 (64)}

Table 2. Continued.

Category	Functional Group	Spot ID	Sensitivity	Specificity	FDR Rank	FDR	COG	Contig	EfaeDRAFT Gene Number	Gene	COG Description	Species	E-Value	Similarity (%)[a]
		Efm.021E04	0.28	1	518	1.49E-03	F	625	1163	GMP synthetase C terminal, GMP synthetase, N terminal	GuaA		0	99
		Efm.005E04	0.78	0.8	264	7.60E-04	H	656	2455	Putative methyltransferase	UbiE		0	100
		Efm.028E04	0.91	0.92	60	1.73E-04	P	573	440	Haloacid dehalogenase-like hydrolase:E1-E2 ATPase-associated region	MgtA		0	99
		Efm.017B07	0.83	0.82	199	5.73E-04	P	640	1539	Hypothetical protein	NrfD		0	92
	Combination	Efm.032D11	0.54	1	228	6.56E-04	K, T	582	537	Sigma-54 factor, interaction region	PspF		0	94
		Efm.028A02	0.93	0.92	51	1.47E-04	K, G	647	1787	ROK (Repressor, ORF, Kinase)	NagC		0	97
		Efm.022F03	0.89	0.92	64	1.84E-04	T, K	658	2596	Response regulator receiver:transcriptional regulatory protein, C-terminal	OmpR		0	97
	AB resistance	Efm.029D12	0.8	0.84	206	5.93E-04	J	656	2452	Aminoglycoside phosphotransferase	Aph		0	99
		Efm.032A05	0.8	0.84	205	5.90E-04	J	656, 557[b]	2452	Aminoglycoside phosphotransferase	Aph		0	97
		Efm.005E04	0.78	0.8	264	7.60E-04	V	656	2454	Streptomycin adenylyltransferase			0	100
		Efm.033B03	0.83	1	58	1.67E-04	V	658	2634	Chloramphenicol O-acetyltransferase	COG4845		0	99
	Poorly characterized	Efm.021F10	0.67	0.96	175	5.04E-04	R	630	1273	Putative membrane protein			−149	90
		Efm.031C07	0.63	1	131	3.77E-04	R	630	1273	Putative membrane protein			0	99
		Efm.037D10	0.65	0.98	148	4.26E-04	R	630	1273	Putative membrane protein			0	99
		Efm.036H02	0.52	1	246	7.08E-04	R	645	1720	Phosphoesterase PHP, N-terminal	COG0613		−94	{99 (99)}
		Efm.021F12	0.76	0.98	76	2.19E-04	R	650	1966	Phage terminase	COG4626		0	97
		Efm.048D04	0.76	0.98	73	2.10E-04	R	650	1966	Phage terminase	COG4626		0	99
		Efm.036C08	0.54	0.98	251	7.23E-04	S	609	900	Conserved hypothetical protein			0	99
		Efm.021E04	0.28	1	518	1.49E-03	S	625	1162	Conserved hypothetical protein			0	99
		Efm.034D12	0.52	1	247	7.11E-04	S	632	1325	Conserved hypothetical protein	COG5495		0	98
		Efm.023D12	0.72	0.96	128	3.68E-04	S	637, 656, 501, 638, 658, 582, 541[b]	1464	Hypothetical protein			0	98
		Efm.024F04	0.93	1	8	2.30E-05	S	638	1476	Conserved hypothetical protein	COG4289		0	98
		Efm.021F12	0.76	0.98	76	2.19E-04	S	650	1965	Phage protal protein HK97	COG4695		0	97
		Efm.048D04	0.76	0.98	73	2.10E-04	S	650	1965	Phage portal protein HK97	COG4695		0	99
		Efm.013B06	0.96	0.96	30	8.64E-05	S	653	2159	Hypothetical protein (previous annotation: methyl accepting chemotaxis protein)			0	97
		Efm.033D03	0.74	0.82	293	8.43E-04	S	653	2194	Hypothetical protein			0	99
		Efm.029D04	0.8	0.8	242	6.97E-04	S	653	2196	Protein of unknown function DUF 955			0	99
		Efm.033D03	0.74	0.82	293	8.43E-04	S	653	2196	Protein of unknown function DUF 955			0	99
		Efm.036C07	0.67	0.9	249	7.17E-04	S	653	2244	Hypothetical protein			0	98
		Efm.036C07	0.67	0.9	249	7.17E-04	S	653	2247	Hypothetical protein			0	98
		Efm.013B06	0.96	0.96	30	8.64E-05	S	653	2249	Hypothetical protein			0	97
		Efm.004G03	0.98	0.9	37	1.07E-04	S	655	2336	Conserved hypothetical protein				
		Efm.020A11	0.74	0.98	86	2.48E-04	S	656	2431	Conserved hypothetical protein			0	99
		Efm.006H03	0.65	0.96	190	5.47E-04	S	656	2436	Conserved hypothetical protein			0	98
		EfmP.001E07	0.37	0.9	657	1.89E-03	S	658	2601	Hypothetical protein			−11	{76 (87)}
		EfmP.002H07	0.54	0.98	258	7.43E-04	S	658	2602	Hypothetical protein			0	82
		EfmP.002H08	0.67	0.96	177	5.09E-04	S	658	2602	Hypothetical protein			0	91
		EfmP.008H01	0.26	0.98	614	1.77E-03	S	658	2602	Hypothetical protein			−17	{59 (81)}
		EfmP.011C02	0.54	0.92	402	1.16E-03	S	658	2602	Hypothetical protein			0	95
		EfmP.001D07	0.41	1	367	1.06E-03	S	658	2603	Hypothetical protein			−179	94

Table 2. Continued.

Category	Functional Group	Spot ID	Sensitivity	Specificity	FDR Rank	FDR	COG	Contig	EfaeDRAFT Gene Number	Gene	COG Description	Species	E-Value	Similarity (%)[a]
		EfmP.001G04	0.5	0.98	309	8.89E-04	S	658	2603	Hypothetical protein			−179	94
		EfmP.002H07	0.54	0.98	258	7.43E-04	S	658	2603	Hypothetical protein			0	82
		EfmP.002H08	0.67	0.96	177	5.09E-04	S	658	2603	Hypothetical protein			0	91
		EfmP.003B08	0.46	0.92	506	1.46E-03	S	658	2603	Hypothetical protein			−148	94
		EfmP.006H08	0.61	0.96	234	6.74E-04	S	658	2603	Hypothetical protein			−179	92
		EfmP.011C02	0.54	0.92	402	1.16E-03	S	658	2603	Hypothetical protein			0	95
		EfmP.006H07	0.78	0.88	168	4.84E-04	S	658	2604	Hypothetical protein			0.008	{80 (90)}
		Efm.035H04	0.43	0.9	563	1.62E-03	S	658	2604	Hypothetical protein			−38	97
		EfmP.008H07	0.7	0.98	118	3.40E-04	S	658	2604	Hypothetical protein			−38	97
		Efm.004F07	0.59	0.94	285	8.20E-04	S	658	2604	Hypothetical protein			−67	98
		Efm.027B09	0.54	0.96	302	8.69E-04	S	658	2604	Hypothetical protein			−168	87
		EfmP.001D07	0.41	1	367	1.06E-03	S	658	2604	Hypothetical protein			−179	94
		EfmP.001G04	0.5	0.98	309	8.89E-04	S	658	2604	Hypothetical protein			−179	94
		EfmP.002H08	0.67	0.96	177	5.09E-04	S	658	2604	Hypothetical protein			0	91
		EfmP.003B08	0.46	0.92	506	1.46E-03	S	658	2604	Hypothetical protein			−148	94
		EfmP.003H02	0.57	0.98	239	6.88E-04	S	658	2604	Hypothetical protein			−168	87
		EfmP.010A02	0.39	1	411	1.18E-03	S	658	2604	Hypothetical protein			−133	96
		Efm.019H06	0.85	0.8	193	5.56E-04	S	658	2609	Lin2822			0	98
		EfmP.004D01	0.72	0.98	107	3.08E-04	S	658	2612	Hypothetical protein			−39	98
		EfmP.010G03	0.72	0.98	106	3.05E-04	S	658	2612	Hypothetical protein			0	99
		Efm.019H06	0.85	0.8	193	5.56E-04	S	658	2721	Hypothetical protein			0	98
	No homologous ORF	Efm.026H08	0.93	0.86	65	1.87E-04		646, 529[b]		between 1771 and 1772[b]			−40	96
		EfmP.007G12	0.96	0.8	78	2.25E-04		606		downstream 850 (last ORF on contig)			0	99
		EfmP.011F01	0.78	0.86	202	5.81E-04		630		between 1276 en 1267			−40	97
		EfmP.009F01	0.26	1	544	1.57E-03		587		between 589 and 590			−48	85
		Efm.019F12	0.98	1	2	5.76E-06		654		Upstream 2310			0	98
		Efm.015C03	0.96	0.9	49	1.41E-04		580, 527, 616, 594, 658, 655, 573[b]		downstream 523 (last ORF on contig)[b]			0	100
Homologues other than *E. faecium* DO	**Information storage and processing**	Efm.021E07	0.63	0.8	496	1.43E-03	K			Phosphosugar-binding RpiR family, transcriptional regulator	EF0692	*E. faecalis* V583	−141	{83 (93)}
		Efm.024E09	0.63	0.82	474	1.36E-03	L			IS*256*, pTEF1/2/3; IS*256* family	EF2632	*E. faecalis* V583	0	99
		EfmP.002D12	0.67	0.84	345	9.93E-04	L			IS*256;* IS*256* family	EF2632	*E. faecalis* V583	0	99
		EfmP.008H10	0.65	0.84	403	1.16E-03	L			IS*256*, pTEF1/2/3; IS*256* family	EF2632	*E. faecalis* V583	−123	99
		EfmP.004E03	0.63	0.84	434	1.25E-03	L			IS*256* transposase; IS*256* family	SH0339	*Streptococcus haemolyticus* JCSC1435	0	99
		Efm.011H03	0.67	0.88	283	8.15E-04	L			Phage integrase	GKP42	*Geobacillus kaustophilus* HTA426	−23	{47 (60)}
		Efm.024C11	0.61	1	163	4.69E-04	L			Putative transposase	glr3859	*Gloeobacter violaceus PCC* 7421	−13	{37 (60)}
		EfmP.001F02	0.2	1	639	1.84E-03	L			pEFNP1		*E. faecium* N15	−25	82
		EfmP.002H02	0.35	1	464	1.34E-03	L			pEFNP1		*E. faecium* N15	−108	100

Table 2. Continued.

Category	Functional Group	Spot ID	Sensitivity	Specificity	FDR Rank	FDR	COG	Contig	EfaeDRAFT Gene Number	Gene	COG Description	Species	E-Value	Similarity (%)[a]
		EfmP.004E09	0.41	1	387	1.11E-03	L			pEFNP1		*E. faecium* N15	−120	{100 (100)}
		EfmP.004G12	0.5	0.98	310	8.92E-04	L			pEFNP1		*E. faecium* N15	−41	88
		EfmP.005B01	0.41	0.98	443	1.28E-03	L			pEFNP1		*E. faecium* N15	−68	{71 (82)}
		EfmP.007B02	0.35	0.96	537	1.55E-03	L			pEFNP1		*E. faecium* N15	−76	92
		EfmP.007D07	0.37	0.98	493	1.42E-03	L			pEFNP1		*E. faecium* N15	−21	83
		EfmP.007G06	0.41	1	356	1.02E-03	L			pEFNP1		*E. faecium* N15	−123	96
		EfmP.011F07	0.59	0.98	223	6.42E-04	L			pEFNP1		*E. faecium* N15	−176	96
		EfmP.011G12	0.43	0.96	461	1.33E-03	L			pEFNP1		*E. faecium* N15	−13	86
		EfmP.008D12	0.52	0.94	390	1.12E-03	L			pEFNP1		*E. faecium* N15	−15	85
		Efm.029H01	0.57	0.96	268	7.71E-04	L			pEFNP1		*E. faecium* N15	−13	88
		Efm.035H04	0.43	0.9	563	1.62E-03	L			pEFNP1		*E. faecium* N15	−45	81
		EfmP.001D06	0.54	0.88	478	1.38E-03	L			pEFNP1		*E. faecium* N15	−22	62
		EfmP.001D05	0.46	0.96	432	1.24E-03	L			pKQ10		*E. faecium* pKQ10	−127	98
		EfmP.003G08	0.5	0.98	319	9.18E-04	L			pKQ10		*E. faecium* pKQ10	−7	79
		EfmP.005B01	0.41	0.98	443	1.28E-03	L			pKQ10		*E. faecium* pKQ10	−15	79
		EfmP.006A06	0.41	0.96	481	1.38E-03	L			pKQ10		*E. faecium* pKQ10	0	97
		EfmP.007D07	0.37	0.98	493	1.42E-03	L			pKQ10		*E. faecium* pKQ10	−17	92
		EfmP.005A10	0.43	0.98	404	1.16E-03	L			MOB, pTX14–1		*B. thuringiensis* pTX14–1	−10	{41 (65)}
		EfmP.003E03	0.35	1	465	1.34E-03	L			pTX14–1		*B. thuringiensis* pTX14–1	−14	89
		EfmP.003H02	0.57	0.98	239	6.88E-04	L			pTX14–1		*B. thuringiensis* pTX14–1	−3	82
		Efm.029H01	0.57	0.96	268	7.71E-04	L			pTX14–1		*B. thuringiensis* pTX14–1	−3	82
		EfmP.011C05	0.35	0.98	508	1.46E-03	L			pEFNP1		*E. faecium* N15	−40	81
		EfmP.001D06	0.54	0.88	478	1.38E-03	L			pER371		*Streptococcus thermophilus* ST371	−31	93
		EfmP.007D01	0.26	1	543	1.56E-03	L			Unknown	ORFA	*Lactococcus lactis* NCDO 275 pCI2000	−9	{42 (53)}
	Cellular processing	Efm.041B04	0.72	0.96	127	3.66E-04	M			Cell wall surface anchor family protein	EF1896	*E. faecalis* V583	0	96
		EfmP.012B02	0.5	0.96	348	1.00E-03	M			M protein	emm	*Streptococcus dysgalactiae* SS957C	−2	{31 (48)}
	Metabolism	Efm.024H03	0.43	0.92	519	1.49E-03	G			Putative endo-arabinase	CsacDRAFT_0054	*Caldicellulosiruptor saccharolyticus* DSM 8903	−59	{40 (58)}
		Efm.019D08	0.74	0.94	133	3.83E-04	H		Pfl_1686	ABC transporter, periplasmic substrate-binding protein	AfuA	*Pseudomonas fluorescens* PfO-1	−20	{60 (68)}
	Combination	Efm.026H08	0.93	0.86	65	1.87E-04	K, T			RNA polymerase sigma-70 factor, ECF subfamily	EF3180	*E. faecalis* V583	−40	{50 (70)}
	Poorly characterized	Efm.032H11	0.83	0.82	200	5.76E-04	R			Hypothetical protein, predicted transcriptional regulator	BCE0370	*Bacillus cereus* ATCC 10987	−2	{48 (62)}
		Efm.007D04	0.65	0.96	188	5.41E-04	R			Hypothetical protein	EF1877	*E. faecalis* V583	0	87
		EfmP.011B08	0.57	0.94	322	9.27E-04	S			YttA	YttA	*Bacillus licheniformis*	−3	{31 (51)}
		EfmP.008B02	0.48	0.98	334	9.61E-04	S			Unnamed protein product	KLLA0F15367g	*Kluyveromyces lactis* NRRL Y-1140	−2	{34 (52)}
		Efm.006E01	0.7	0.96	142	4.09E-04	S			Hypothetical protein	PG0456	*Photobacterium profundum* SS9	−8	{48 (66)}
		Efm.006E01	0.7	0.96	142	4.09E-04	S			Hypothetical protein	PG0457	*Porphyromonas gingivalis* W83	−8	{35 (60)}

Table 2. Continued.

Category	Functional Group	Spot ID	Sensitivity	Specificity	FDR Rank	FDR	COG	Contig	EfaeDRAFT Gene Number	Gene	COG Description	Species	E-Value	Similarity (%)[a]
		Efm.020F07	0.39	0.98	466	1.34E-03	S		Orf 37	Hypothetical protein	traE	*E. faecium* pHT beta	−89	[96 (98)}
	No homology	Efm.043G03	0.48	0.92	487	1.40E-03								
		Efm.046F02	0.65	0.98	147	4.23E-04								
		EfmP.008D11	0.59	0.94	287	8.26E-04								
PCR-amplified virulence and antibiotic resistance genes		Ent-aac6-Ie-aph2-Ia (1)	0.7	0.84	328	9.44E-04	J			Aminoglycoside acetylase phosphotransferase	aac6-Ie-aph2-Ia			
		Ent-aac6-Ie-aph2-Ia (2)	0.39	0.94	587	1.69E-03	J			Aminoglycoside acetylase phosphotransferase	aac6-Ie-aph2-Ia			
		Ent-uveE	0.61	1	160	4.61E-04	V			Putative sigma factor; ORF1	uve2	*E. faecium* E300		
		Ent-sat4	0.8	0.84	203	5.84E-04	V			Streptogramine resistance gene	sat4			
		Ent-Aph3-III	0.8	0.84	204	5.87E-04	V			Aminoglycoside phosphotransferase Aph3-IIIa	Aph3-III			
		Ent-aadE	0.8	0.82	217	6.25E-04	V			Aminoglycoside 6-adenylyltransferase	aadE			
		Ent-esp	0.61	1	159	4.58E-04	M			Putative enterococcal surface protein; ORF3	esp	*E. faecium* E300		
		Ent-nox	0.61	1	162	4.66E-04	R			Putative NADH oxidase; ORF4	nox	*E. faecium* E300		
		Ent-pep1F	0.61	1	161	4.63E-04	V			Peptidoglycan hydrolase; autolysin; ORF5		*E. faecium* E300		
		Ent-pRIE2988F	0.59	0.98	222	6.39E-04	K, T			Transcriptional regulator; ORF2	araC	*E. faecium* E300		
		Ent-txe-axe	0.76	0.94	111	3.20E-04	S, D			Toxin and antitoxin	txe and axe	*E. faecium* pRUM		

[a]Nucleotide similarity is given in percent, protein identity is given between parentheses in percent, and protein similarity is given between curly brackets in percent.
[b]For insert sequences equally similar to genes on more than one contig, all contigs are shown; gene details are only given for the first contig.
D, cell division and chromosome partitioning; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme metabolism; J, translation ribosomal structure and biogenesis; K, transcription; L, DNA replication, recombination, and repair; M, cell membrane biogenesis; P, inorganic ion transport and metabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanisms; V, defense mechanisms.
doi:10.1371/journal.ppat.0030007.t002

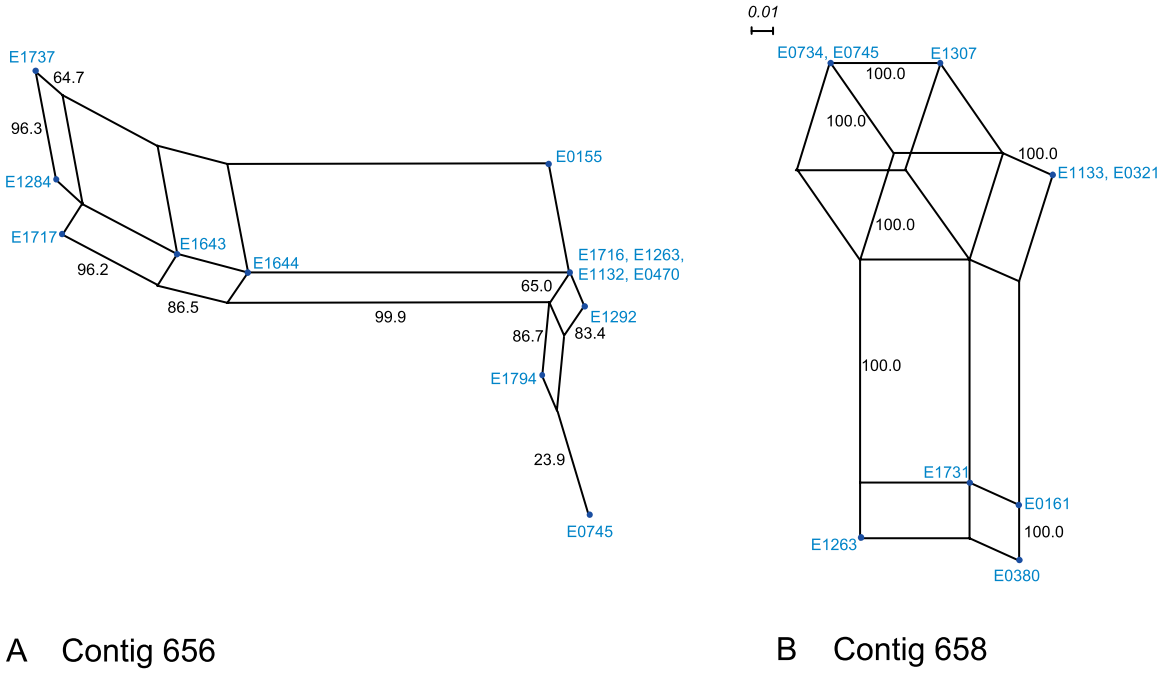

Figure 3. SDA of Hospital Clade–Specific Genes on Contigs 656 and 658 among *E. faecium* Strains

SDA of hospital clade–specific genes on contig 656 of 13 hospital clade strains (A) and hospital clade–specific genes on contig 658 of ten hospital clade strains (B). The nodes represent strains (E numbers correspond to Table 1) and are depicted as blue circles. The scale bar represents Hamming distance. Numbers at the edges represent the percent bootstrap support of the split obtained after 1,000 replicates. Paralellograms indicate recombinatory events. Fit in both graphs is 100%.

doi:10.1371/journal.ppat.0030007.g003

antibiotic resistance genes, present in 80%–83% of the hospital clade isolates, and genes involved in carbohydrate metabolism, present in 89%–93% of hospital clade isolates, in addition to IS elements. Interestingly, we found the *aadE-sat4-aphA-3* aminoglycoside streptothricine resistance gene cluster highly associated with the hospital clade (present in 84% of hospital clade isolates) and not linked to a specific IS element. In a previous study, this gene cluster had been identified as part of a Tn*5405*-like structure in 70.1% of isolates [54]. The Tn*5405*-like structure is present in *E. faecium* DO, with IS*1182* annotated as EfaeDRAFT__2457 cassette chromosome recombinase B1 on contig 656, and ORFs X and Y (EfaeDRAFT__2456 and 2455) and the *aadE-sat4-aphA-3* cluster (EfaeDRAFT__2452, 2453 and 2454) located directly upstream. In contrast to our data, this gene cluster was found in German multiresistant animal and sewage isolates, as well as in human hospitalized patients and outpatients [55]. Most hospital clade–specific hypothetical proteins and other genes, including the putative PAI genes, were identified in only a smaller subset, reflecting acquisition after initial branching of the hospital clade. Previously, the PAI has also been identified only in a fraction (approximately 60%) of hospital outbreak and infection strains [22]. Frequent recombination resulted in genomic mosaicism, as exemplified by contigs 656 and 658. Variation in the combination of accessory genes might supply the organism with a varying armamentarium to colonize or infect the host and escape the immune system.

Our study is unique in that it identifies a single phylogenetic hospital clade of strains with epidemic potential in hospitals *and* with more than 100 genes specific to this clade. Among the 97 strains included in this study, a core genome of 65% of the inserts was identified, while 13% of spots were highly associated with the hospital clade. Genetic subpopulations strongly associated with virulent or epidemic potential have not been identified for many other species. Recently, a nonlivestock-associated clade that contained all infectious *Campylobacter jejuni* isolates and associated genetic factors was identified using an approach similar to ours [56]. In addition, Howard et al. confirmed by comparative phylogenomics three distinct clusters of *Yersinia enterocolitica* composed of a nonpathogenic clade, a low pathogenic clade, and a high pathogenic clade [57]. In contrast, population genetics of *S. aureus* repeatedly failed to identify virulence factors associated with enhanced virulence or a subpopulation adapted to the hospital, though overrepresentation of invasive isolates in certain (sub)clusters was identified [28,30,58]. In another study, factors associated with *S. epidermidis* invasiveness were identified, but a phylogenetic analysis failed to distinguish invasive isolates from controls [33]. Nevertheless, evolution of a pathogenic species from a less pathogenic species has been documented before (i.e., *Shigella* species from *E. coli* [59], and *Bacillus anthracis* from *Bacillus cereus* [60]). A quantum leap of evolution in these bacteria occurs when new genes are acquired en bloc via horizontal gene transfer by plasmids and bacteriophages [61]. Acquisition of these genes enables the pathogen to colonize a new niche, and new selective constraints lead to progressive adaptation of the organism by pathoadaptive mutations [62]. When aligning the enterococcal plasmid pRUM to *E. faecium* DO, the plasmid was highly similar to parts of contig 658, including the toxin–anititoxin system. This may reflect integration of a plasmid in the *E. faecium* genome. An important element of pathoadaptive evolution is selection of "black holes": inactivation to pseudogenes or loss of genes,

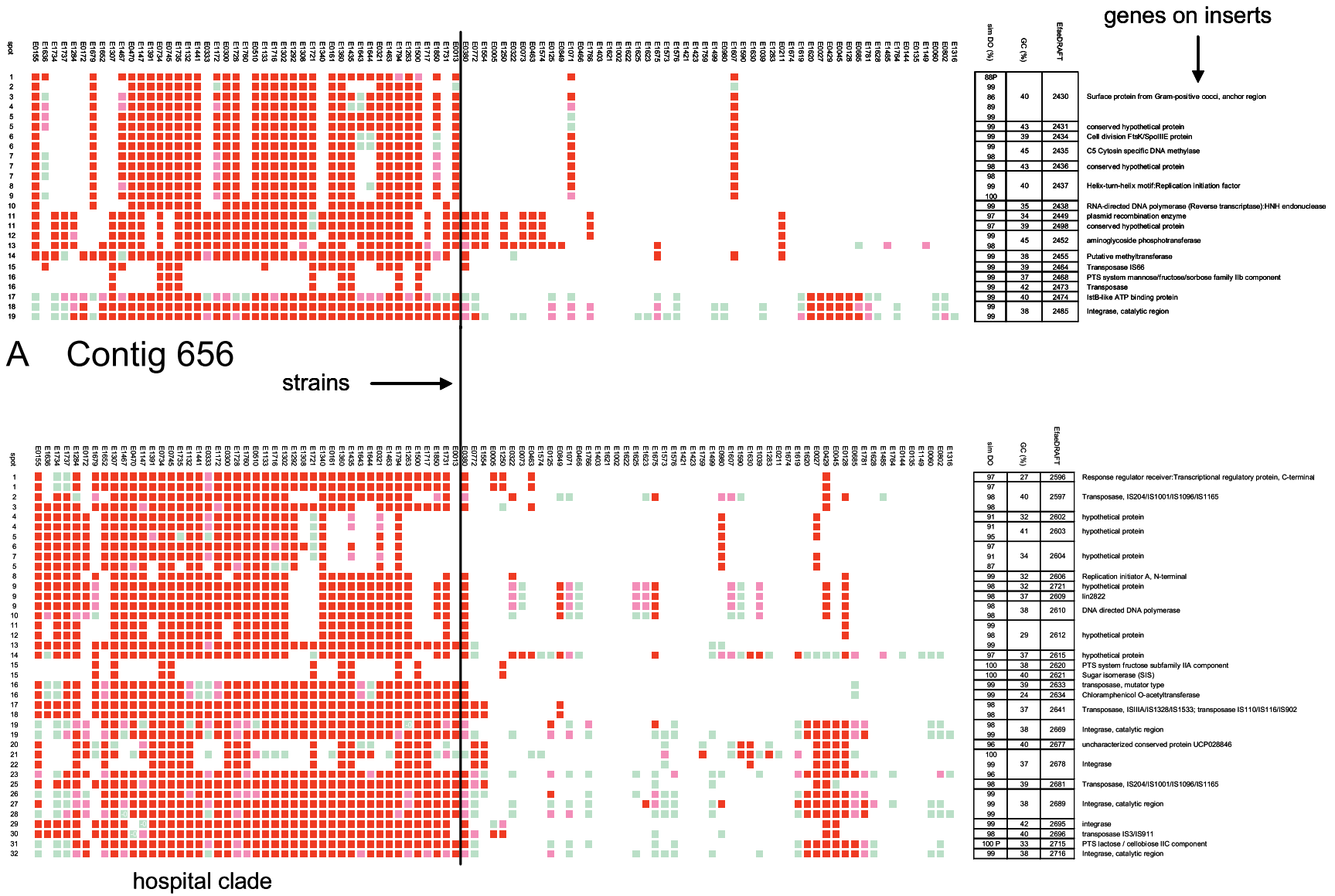

Figure 4. Presence and Absence of Hospital Clade–Specific Genes on Contigs 656 and 658 among *E. faecium* Strains

Presence and absence of hospital clade–specific genes on contig 656 (A) and 658 (B) among *E. faecium* strains. Strains are given on the *x*-axis (E numbers correspond to Table 1). Hospital clade strains are shown on the left panel and are separated from non–hospital clade strains by a vertical line. Hospital clade–specific genes with similarity to *E. faecium* DO genes, gene numbers, and GC content are given on the *y*-axis on the right side, in presumed order. Spot number is given on the *y*-axis on the left side; more than one gene can be located in one insert. A highly certain present gene (plus 0.5) is indicated with a red square, an absent or highly diverging gene (minus 0.5) with absence of a square. Slightly deviating genes (equal to or larger than 0, and smaller than 0.5) are indicated with a pink square, deviating genes (larger than minus 0.5, and smaller than 0) with a green square.

doi:10.1371/journal.ppat.0030007.g004

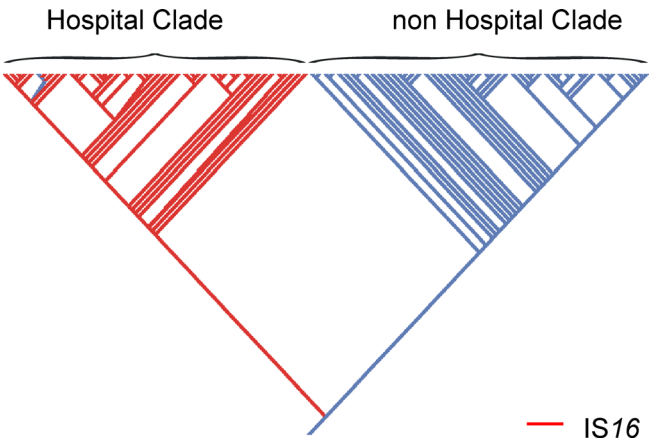

Figure 5. Distribution of *IS*16 among *E. faecium* Strains

Maximum likelihood–based gene analysis for determining the distribution of individual IS*16* throughout the phylogenetic tree is shown. Strains in which IS*16* is present are colored red; strains in which IS*16* is absent are colored blue. Strains in the hospital clade all contain IS*16* with one exception. IS*16* is absent in all strains in the non–hospital clade.

doi:10.1371/journal.ppat.0030007.g005

which leads to genome decay [62]. The inactivated and lost genes are often antivirulence genes [62]. This DNA may still be present in the nonpathogenic ancestor. Our results didn't show direct evidence for complete loss of many genes specific for hospital clade strains; however, identification of pseudogenes with this approach is not possible. The fact that hospital clade strains are only rarely found outside the hospital indicates reduced fitness of these strains in the ancestral niche, which possibly reflects antagonistic pleiotropy in which hospital specialization is detrimental in other niches [63]. From these evolutionary insights, one might conclude that worldwide emergence of the *E. faecium* hospital clade represents, in fact, evolution of a novel hospital-adapted subspecies from a nonpathogenic (commensal) *E. faecium* ancestor, which succeeded in competing with *E. faecalis* as causative agent of hospital infections.

Materials and Methods

Bacterial strains and nucleic acid extraction. The 97 bacterial isolates used in this study originated from different documented epidemiological niches: 18 mostly monoclonal, hospital epidemics; 35 clinical sites representing invasive human infections, including *E. faecium* DO (E1794 in Table 1); 15 hospital surveys, representing asymptomatic carriage of hospitalized patients; 11 community surveys, representing asymptomatic carriage in healthy subjects; three environmental isolates; and 15 animals in 21 countries on six continents (Table 1) [20,41,64–72]. These strains were a subset of a large, genotypically well-characterized collection, which represented the global *E. faecium* population. The subset was selected based on differences in geographic locations, hosts, and sequence types. All strains were cultured on tryptic soy agar sheep blood plates at 37 °C. DNA for labeling (see below) was prepared from cell suspensions by bead-beating and chloroform phenol extraction. DNA for Southern blots was isolated according to the manufacturer's instructions with DNeasy Tissue kit (Qiagen, http://www.qiagen.com). The shotgun library was created from nine strains from different epidemiologic and genetic backgrounds according to MLST analysis (Table 3) [21]. In order to prevent overrepresentation of inserts from high copy plasmids, plasmid DNA was separated from chromosomal DNA according to Willems et al. [73] with the modification that clumping high molecular chromosomal DNA, cured from plasmid DNA, was captured with a glass capillary. Plasmids were prepared from the library strains with the QIAprep Spin Miniprep Kit according to the manufacturer's instructions (Qiagen) and amplified with the Templi-Phi Amplification Kit (Amersham, http://www.amersham.com).

Microarray fabrication. Equal amounts of chromosomal DNA from nine genetically diverse *E. faecium* strains were mixed to create a shotgun library as described by Borucki et al. [74] (Table 3). The same procedure was repeated for the plasmid DNA preparations. Briefly, 10 μg of pooled DNA (equal amounts from each strain) was sonicated (Branson 250/450 Sonifier, 6-mm microtip; http://www.sonifier.com), and fragments of approximately 0.8–1.2 kb for genomic DNA and 1.2–1.7 kb for plasmid DNA were gel isolated, extracted (Qiaquick columns, Qiagen), and end-repaired (DNA Terminator End Repair Kit; Lucigen Corporation, http://www.lucigen.com). End-repaired fragments were ligated to pSMART-HC-Kan (Clone-SMART, Lucigen), and *E. coli* (ElectroMAX DH10B Cells; Invitrogen, http://www.invitrogen.com) were transformed with this recombinant plasmid. Next, 4,560 genomic and 1,140 plasmid recombinant clones were arrayed in 96-well plates. Clone inserts were amplified by PCR with amino-modified SMART primers. Additionally, fragments from one enterococcal housekeeping gene, 13 virulence genes, and 20 genes involved with antibiotic resistance were PCR–amplified and included in the microarray (Table S7) [18,49,55,75–89]. PCR products were ethanol purified and resuspended in 1 × SSC (1 × SSC is 0.15 M NaCl plus 0.015 M sodium citrate). All genomic, plasmid, and additional gene PCR products were printed by using ESI three-axis DB-3 robot (Versarray ChipWriter Pro; Biorad, http://www.bio-rad.com) at a controlled humidity of 55% on CSS silylated slides (European Biotech Network, http://www.euro-bio-net.com). Slides were printed in two batches, after which they were blocked following the manufacturer's instructions. Genomic coverage of the library on the nucleotide level was calculated using Formula 1 [35]:

$$N = \ln(1 - P)/\ln[1 - (1 - (I/G))], \quad (1)$$

Table 3. Strains Used in Library

Epidemiology	Country	Number	Year	MLST		PAI[a]	vanR[a]	ampR[a]	Strain	Reference
				CC	ST					
E	USA	1–1	1995	17	17	1	1	1	E0155	[65]
E	NLD	3	1998–1999	17	16	1	1	0	E0470	[69]
Ci	GBR	6	1999	non 17	157	0	1	0	E0699	[72]
Ci	DEU	2	1998	17	17	0	0	1	E1284	[41]
Ci	NLD	1	1950	non 17	67	0	0	0	E1620	[22]
Ci	USA	1	1991	17	18	0	0	1	E1794	[70]
C	NLD	7	2000	non 17	32	0	1	0	E1071	[41]
C	NLD	3	1996	non 17	6	0	1	0	E0135	[20]
A Cat	NLD		1996	non 17	21	0	1	0	E0466	[71]

[a]1 indicates presence, 0 indicates absence.

A, animal; ampR, ampicillin resistance; C, community surveillance; Ci, clinical; DEU, Germany; E, epidemic; GBR, Great Britain; NLD, Netherlands; ST, sequence type; USA, United States of America; vanR, vancomycin resistance.

doi:10.1371/journal.ppat.0030007.t003

in which N = number of clones, P = probability of coverage, I = insert size and G = genome size.

This formula, however, is based on the assumption that that every single nucleotide should be present in the library. In the present approach, there is no need for a complete ORF to be present: border sequences with a minimal size of 100 nucleotides should be sufficient to obtain positive hybridization signals (Formula 2) [34]:

$$1 - (1 - (T + I - 2(RO/G)))^N, \quad (2)$$

in which T = transcript length and RO = required overlap.

In this study, both algorithms are used to estimate genomic coverage.

Labeling, hybridization, and data acquisition. Total DNA (0.5 μg) was labeled with fluorescent dyes by random priming with the Bioprime labeling system (Invitrogen). To normalize the two channels for label incorporation, DNA concentration differences, and variation in slide scanning, equal amounts of the library strains were mixed as the reference pool and labeled with Cy3 dUTP. Tester strains were labeled with Cy5 dUTP. Ten tester strains were hybridized in duplo for control of reproducibility. For each hybridization, Cy5 and Cy3 probes were combined with yeast tRNA, speed vacuum dried, resuspended in 40 μl Easy hyb buffer (Roche Diagnostics Netherlands B.V., http://www.roche-diagnostics.nl), and denatured for 2 min at 100 °C.

Silylated slides were prehybridized in prehybridization solution (1% BSA, 5 × SSC and 0.1% sodium dodecyl sulfate, filtered) at 42 °C during 45 min while rotating, washed twice with filtered milli Q water, dried with N_2 flow, and prewarmed at 42 °C. Easy hyb solution was pipetted on the microarray print of the slide, covered with a hybrislib, and placed in hybridization chambers (Corning Life Sciences B.V., http://www.corning.com/lifesciences). Hybridizations were performed overnight at 42 °C in a waterbath. Microarrays were then washed sequentially in (*i*) 1 × SSC/0.2% sodium dodecyl sulfate for 10 s at 37 °C, (*ii*) 0.5 × SSC for 10 s at 37 °C, and (*iii*) twice in 0.2 × SSC for 10 min at room temperature, and dried with N_2-flow. A Scanarray Express 680013 Microarray Analysis System (PerkinElmer Life and Analytical Sciences, http://las.perkinelmer.com) was used for scanning slides.

Microarray images were quantified with Imagene software version 4.2 (Biodiscovery, http://www.biodiscovery.com). Inferior spots (empty spots, exceeding SD of pixels, less than two times background in Cy3 channels), were excluded from normalization and data analysis.

Data processing, comparative phylogenomics, and identification of predictive genes. To correct for differences in labeling, hybridization conditions, slide quality, and scanning circumstances, each slide was normalized independently. At first, ratios of Cy5 minus background to Cy3 minus background were calculated and log2-transformed. Filtering was applied to exclude spots with flags; for estimating the correction factor in normalization, only spots were included with Cy3 values larger than two times background. Mean log2 ratios were calculated and applied to each independent ratio. Next, the data were transformed using GACK (http://falkow.stanford.edu/whatwedo/software/software.html) to assign a region of considerate absence and presence, corresponding with, respectively, minus 0.50 and plus 0.50, and an interval with indefinite presence/absence, or divergence, to interpret the data. For Bayesian modeling, maximum parsimony analysis, SDA, and character evolution modeling data was transformed to binary output.

Using a Nexus format matrix, the relationship of strains based on the presence and absence of hybridizing signal on spotted inserts was determined with Bayesian-based algorithms implemented through MR BAYES 3.0 software [90], as explained by Champion et al. [56]. With samples and saves from every 40th tree, 1,100,000 generations of four incrementally heated Markov chain Monte Carlos were performed on the DNA–DNA microarray data by using the default annealing temperature of 0.5, a burn-in of 100,000 Markov chain Monte Carlos generations, and an 8-category distribution. Ninety five percent majority rule consensus trees and clade credibility values were obtained by using TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html). In addition to the Bayesian-based approach, phylogeny was studied with hierarchical clustering and maximum parsimony analysis. [24]. Bootstrapped (1,000 iterations) complete linkage transversal hierarchical clustering with Euclidian distance was performed and visualized with TIGR MeV version 3.1 software (http://www.tm4.org/mev.html). One thousand times bootstrapped maximum parsimony analysis was performed with PAUP* 4.0 software (Sinauer Associates, http://paup.csit.fsu.edu).

In order to select a maximal variety of differential genes and gain insight into the degree of redundancy, hierarchical clustering (for details see above) was used to generate a dendrogram of genes based on their patterns of absence and presence across the strains. Detection of clusters of acquired genes is based on the assumption that co-inherited genes can be found co-located on the bacterial genome. Subsequently, a subset of inserts was considered highly specific for a clade according to the following two criteria: First, insert specificity for the clade was higher than 80%, estimated with the χ^2 test and followed by FDR correction [91]. Second, the insert clustered with a Euclidian distance of 1.1 or less for the genomic library and 0.8 or less for the plasmid library was selected for sequencing (see below). To gain insight into the core genome; 35 randomly chosen inserts that gave a positive hybridization signal in all strains were sequenced.

SDA based on the presence and absence of ORFs, with identity to genes belonging to contigs 658 and 656 in a selection of the most genetically diverse strains, was used to test for parallel changes in the gene order on these *E. faecium* DO contigs. The bootstrapping procedure for SDA was used as implemented in the SplitsTree program version 4.0 using Hamming correction (http://www.splitstree.org). Recombination, hybridization, gene conversion, and gene transfer all lead to histories that are not adequately modeled by a single tree. They effectively cause lineages to coalesce forward in time, resulting in trees that have reticulations or a network structure rather than the simple branching structure seen with most phylogenies. Split decomposition does not force tree topologies to be strictly bifurcating or multibranching but permits network relationships. A split decomposition graph will look less tree-like and more net-like as the influence of recombination becomes more important in the history of a set of taxa. Since splits graphs were sufficiently complex and the distances among strains sufficiently great, data sets had to be simplified by removing strains representing the longest branches to allow visualization of central networks and improve the fit parameter, which is similar to the method described in [92].

The character evolution maximum likelihood–based model of Mesquite 1.06 software (http://www.mesquiteproject.org) was used to identify clade-predictive genes on binary data in nexus format. Tracing shows the most likely hypothesis of ancestral states, and indicates how presence and absence or divergence of certain genes in an ancestral strain has led to the formation of a new clade.

Sequencing and analysis of differential inserts and validation of microarray results. Inserts selected for sequencing (see above) were PCR-amplified and sequenced single-stranded from one direction in combination with the BigDye Terminator reaction kit by using an ABI PRISM 3700 DNA analyzer (Applied Biosystems, http://www.appliedbiosystems.com). All sequences were blasted in GenBank. COGs for *E. faecium* genes and proteins were assigned according to the Oak Ridge National Laboratory Web site (http://maple.lsd.ornl.gov/cgi-bin/JGI_microbial/display_page.cgi?page=cog&org=efae&chr=08jun04). COGs for other species' genes and proteins were assigned according to GenBank.

Microarray hybridization results from ten spots, including six different ORFs (IS*16, esp, vanA,* transposase IS*111A*/IS*1328*/IS*1533*; transposase IS*110*/IS*116*/IS*902, glycosyl hydrolase family 88,* and *extracellular solute-binding protein*), were validated by Southern hybridization. For this purpose, chromosomal DNA preparations were digested with EcoRI, separated by agarose gel electrophoresis (0.8% agarose gels), transferred onto a Hybond N^+ nylon membrane (Nycomed Amersham plc, http://www.amersham.com), and subsequently hybridized to six ECl-labeled PCR products specific for the six ORFs according to the manufacturer's protocol (Amersham). Primers, if not already designed for the additional spots on the array (Table S7; see "Microarray fabrication" section of Materials and Methods), were listed in Table S8.

Determining the codon adaptation index and GC content. Codon usage patterns may vary considerably among genes. It is generally assumed that codons that are best recognized by the most abundant tRNA are those that are translated optimally (most accurate), and are often linked to genes expressed at high levels. The CAI measures codon bias and indicates the relative adaptiveness of the codon usage of a particular gene towards the codon usage of highly expressed genes. A gene that consists only of the most frequently used codons of a reference set of highly expressed genes has the maximal possible CAI value of 1.0 and is thought to be highly expressed. In general, highly expressed genes have high CAI values. The CAI adaptation index was calculated as described previously [36]. Briefly, a CAI calculator (http://www.evolvingcode.net/codon/cai/cai.php) was used to determine the relative synonymous codon usage (RSCU)—that is, the observed frequency of a particular codon divided by its expected frequency under the assumption of equal usage of the synonymous codons for an amino acid [36]—of analyzed genes. CAI is defined as

the geometric mean of the RSCU values corresponding to each of the codons used in that gene, divided by the maximum possible CAI for a gene of the same amino acid composition [36]. A subset of *E. faecium* DO core genes representing genes encoding elongation factors and ribosomal proteins was supposed to be highly expressed in *E. faecium* and was used as a reference for RSCU (Table S9). CAI values were then calculated using the CAI calculator (http://www.evolvingcode.net/codon/cai/cai.php). The DRAFT annotation of the *E. faecium* DO genes is from the Joint Genome Institute *E. faecium* Web site (http://genome.jgi-psf.org/draft_microbes/entfa/entfa.home.html). The mean CAI value with SD was calculated from the sequenced core genome genes (Table S1). Furthermore, the known putative *E. faecium* PAI genes were included. The observed codon frequencies in the *E. faecium* genome were compared with the expected codon frequencies calculated from the GC content at the first, second, and third codon positions under the assumption of the same amino acid composition. The significance of the differences was evaluated by *t*-test statistics.

PFGE and diversity index. PFGE analysis was performed as described previously [6]. Distance matrices of banding patterns between 339.5 kb and 97 kb were calculated with Bionumerics software (version 3.5; Applied Maths, http://www.applied-maths.com) by the Ward method for a subset of hospital clade strains ($n = 22$) and a subset of non–hospital clade strains ($n = 24$) with different sequence types and epidemiological origins. Distance matrices from previously identified MLST allelic profiles were calculated with the categorical coefficient (Bionumerics software). In order to compare the level of similarity among isolates belonging to the hospital clade with non–hospital clade isolates based on PFGE and MLST, a so-called diversity index was calculated. For this, the average of similarities between all possible pairs of hospital clade and non–hospital clade isolates based on MLST was calculated and divided by the average of similarities between all possible pairs of hospital clade and non–hospital clade isolates based on PFGE. When the level of genetic diversity (is 1–similarity) based on MLST and PFGE is identical, the diversity index equals 1. A diversity index greater than one indicates that the average pairwise similarity based on PFGE is higher than that based on MLST, and suggests enhanced genomic rearrangements.

Supporting Information

Dataset S1. Flag Filtered, Normalized Log-Transformed *E. faecium* Mixed Genome Array Data

Found at doi:10.1371/journal.ppat.0030007.sd001 (4.4 MB TXT).

Figure S1. CAI and GC Content of Core Genes on Contig 595 and Hospital Clade–Specific Genes on Contigs 656 and 658

Black lines represent the CAI of hospital clade–specific genes. Dark blue lines indicate hospital clade–specific IS elements with no corresponding CAI value. The mean CAI of the core genes is represented by the red straight lines; the red dotted lines indicate the 95% confidence interval.

Found at doi:10.1371/journal.ppat.0030007.sg001 (27 KB PDF).

Table S1. *E. faecium* Core Genome Sequenced Genes

Found at doi:10.1371/journal.ppat.0030007.st001 (24 KB XLS).

Table S2. CAI Values of *E. faecium* DO Sequenced Core Inserts

Found at doi:10.1371/journal.ppat.0030007.st002 (15 KB XLS).

Table S3. CAI Values of Putative PAI Genes

Found at doi:10.1371/journal.ppat.0030007.st003 (13 KB XLS).

Table S4. Ranking of Inserts Based on Predominance of Hospital Clade Ancestral State, Identified from Character Evolution Modeling

Found at doi:10.1371/journal.ppat.0030007.st004 (299 KB XLS).

Table S5. Distance Matrices of Banding Pattern Similarity of Hospital Clade and Non–Hospital Clade Strains Obtained from PFGE

Found at doi:10.1371/journal.ppat.0030007.st005 (20 KB XLS).

Table S6. Distance Matrices of Similarity of MLST Allelic Profiles of Hospital Clade and Non–Hospital Clade Strains

Found at doi:10.1371/journal.ppat.0030007.st006 (20 KB XLS).

Table S7. Primers Used for PCR-Amplified Antimicrobial Resistance, Virulence, and Housekeeping Genes

a, antimicrobial resistance; h, housekeeping genes; v, virulence.

Found at doi:10.1371/journal.ppat.0030007.st007 (29 KB XLS).

Table S8. Primers Used for Validation Array

Found at doi:10.1371/journal.ppat.0030007.st008 (13 KB XLS).

Table S9. *E. faecium* DO Elongation Factors and Ribosomal Proteins Used to Calculate Reference RSCU

Found at doi:10.1371/journal.ppat.0030007.st009 (15 KB XLS).

Accession Numbers

The GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html) accession numbers for the genes and gene products discussed in this paper are *E. faecium* isolate E300 pathogenicity island (AY322150), pEFNP1 (AB038522), and pKQ10 (U01917).

Acknowledgments

We thank W. Jansen for critical reading of the manuscript, and A. van Belkum and W. van Leeuwen for technical assistance.

Author contributions. HLL, RJLW, WJBvW, FHS, and MPMC conceived and designed the experiments, analyzed the data, and contributed reagents/materials/analysis tools. HLL and MPMC performed the experiments. HLL, RJLW, and MJMB wrote the paper. HLL, RJLW, WJBvW, and MJMB contributed to hypothesis generation and overall study design.

Funding. This work was supported by a ZonMW MD clinical research trainee grant 2903322 from The Netherlands Organization for Health Research and Development. Part of this manuscript was presented at the 106th American Society for Microbiology general meeting in Orlando and the 16th European Congress of Clinical Microbiology and Infectious Diseases in Nice (2006).

Competing interests. The authors have declared that no competing interests exist.

References

1. National Nosocomial Infections Surveillance System (2004) National Nosocomial Infections Surveillance (NNIS) System Report, data summary from January 1992 through June 2004, issued October 2004. Am J Infect Control 32: 470–485.
2. Murray BE (2000) Vancomycin-resistant enterococcal infections. N Engl J Med 342: 710–721.
3. Endtz HP, van den Braak N, van Belkum A, Kluytmans JA, Koeleman JG, et al. (1997) Fecal carriage of vancomycin-resistant enterococci in hospitalized patients and those living in the community in The Netherlands. J Clin Microbiol 35: 3026–3031.
4. Stobberingh E, van Den BA, London N, Driessen C, Top J, Willems R (1999) Enterococci with glycopeptide resistance in turkeys, turkey farmers, turkey slaughterers, and (sub)urban residents in the south of The Netherlands: Evidence for transmission of vancomycin resistance from animals to humans? Antimicrob Agents Chemother 43: 2215–2221.
5. van den Bogaard AE, Mertens P, London NH, Stobberingh EE (1997) High prevalence of colonization with vancomycin- and pristinamycin-resistant enterococci in healthy humans and pigs in The Netherlands: Is the addition of antibiotics to animal feeds to blame? J Antimicrob Chemother 40: 454–456.
6. van den Braak N, van Belkum A, van Keulen M, Vliegenthart J, Verbrugh HA, et al. (1998) Molecular characterization of vancomycin-resistant enterococci from hospitalized patients and poultry products in The Netherlands. J Clin Microbiol 36: 1927–1932.
7. van der Auwera P, Pensart N, Korten V, Murray BE, Leclercq R (1996) Influence of oral glycopeptides on the fecal flora of human volunteers: Selection of highly glycopeptide-resistant enterococci. J Infect Dis 173: 1129–1136.
8. van Belkum A, van Den BN, Thomassen R, Verbrugh H, Endtz H (1996) Vancomycin-resistant enterococci in cats and dogs. Lancet 348: 1038–1039.
9. Fridkin SK, Edwards JR, Courval JM, Hill H, Tenover FC, et al. (2001) The effect of vancomycin and third-generation cephalosporins on prevalence of vancomycin-resistant enterococci in 126 U.S. adult intensive care units. Ann Intern Med 135: 175–183.
10. Centers for Disease Control and Prevention (1995) Recommendations for preventing the spread of vancomycin resistance recommendations of the hospital infection control practices advisory committee (HICPAC). MMWR 44(RR12): 1–13. Available: http://www.cdc.gov/mmwr/preview/mmwrhtml/00039349.htm. Accessed 14 December 2006.
11. Vernon MO, Hayden MK, Trick WE, Hayes RA, Blom DW, et al. (2006) Chlorhexidine gluconate to cleanse patients in a medical intensive care

unit: The effectiveness of source control to reduce the bioburden of vancomycin-resistant enterococci. Arch Intern Med 166: 306–312.
12. Mascini EM, Troelstra A, Beitsma M, Blok HE, Jalink KP, et al. (2006) Genotyping and preemptive isolation to control an outbreak of vancomycin-resistant *Enterococcus faecium*. Clin Infect Dis 42: 739–746.
13. Chang S, Sievert DM, Hageman JC, Boulton ML, Tenover FC, et al. (2003) Infection with vancomycin-resistant *Staphylococcus aureus* containing the *vanA* resistance gene. N Engl J Med 348: 1342–1347.
14. Kacica M, McDonald LC (2004) Brief report: Vancomycin-resistant *Staphylococcus aureus*–New York, 2004. MMWR 53: 322–323.
15. Rudrik JT (2005) Michigan Department of Community Health communication. Available: http://www.michigan.gov/documents/VRSA__Feb05__HAN__118391__7.pdf. Accessed 26 December 2006.
16. Tenover FC, Weigel LM, Appelbaum PC, McDougal LK, Chaitram J, et al. (2004) Vancomycin-resistant *Staphylococcus aureus* isolate from a patient in Pennsylvania. Antimicrob Agents Chemother 48: 275–280.
17. Klare I, Konstabel C, Mueller-Bertling S, Werner G, Strommenger B, et al. (2005) Spread of ampicillin/vancomycin-resistant *Enterococcus faecium* of the epidemic-virulent clonal complex-17 carrying the genes *esp* and *hyl* in German hospitals. Eur J Clin Microbiol Infect Dis 24: 815–825.
18. Leavis H, Top J, Shankar N, Borgen K, Bonten M, et al. (2004) A novel putative enterococcal pathogenicity island linked to the *esp* virulence gene of *Enterococcus faecium* and associated with epidemicity. J Bacteriol 186: 672–682.
19. Rice LB, Carias L, Rudin S, Vael C, Goossens H, et al. (2003) A potential virulence gene, *hylEfm*, predominates in *Enterococcus faecium* of clinical origin. J Infect Dis 187: 508–512.
20. Willems RJ, Homan W, Top J, van Santen-Verheuvel M, Tribe D, et al. (2001) Variant *esp* gene as a marker of a distinct genetic lineage of vancomycin-resistant *Enterococcus faecium* spreading in hospitals. Lancet 357: 853–855.
21. Homan WL, Tribe D, Poznanski S, Li M, Hogg G, et al. (2002) Multilocus sequence typing scheme for *Enterococcus faecium*. J Clin Microbiol 40: 1963–1971.
22. Willems RJ, Top J, van Santen M, Robinson DA, Coque TM, et al. (2005) Global spread of vancomycin-resistant *Enterococcus faecium* from distinct nosocomial genetic complex. Emerg Infect Dis 11: 821–828.
23. Israel DA, Salama N, Krishna U, Rieger UM, Atherton JC, et al. (2001) *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. Proc Natl Acad Sci U S A 98: 14625–14630.
24. Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, et al. (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. Proc Natl Acad Sci U S A 97: 14668–14673.
25. Porwollik S, Wong RM, McClelland M (2002) Evolutionary genomics of *Salmonella:* gene acquisitions revealed by microarray analysis. Proc Natl Acad Sci U S A 99: 8956–8961.
26. Dobrindt U, Agerer F, Michaelis K, Janka A, Buchrieser C, et al. (2003) Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. J Bacteriol 185: 1831–1840.
27. Fukiya S, Mizoguchi H, Tobe T, Mori H (2004) Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray. J Bacteriol 186: 3911–3921.
28. Fitzgerald JR, Sturdevant DE, Mackie SM, Gill SR, Musser JM (2001) Evolutionary genomics of *Staphylococcus aureus:* Insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. Proc Natl Acad Sci U S A 98: 8821–8826.
29. Fitzgerald JR, Reid SD, Ruotsalainen E, Tripp TJ, Liu M, et al. (2003) Genome diversification in *Staphylococcus aureus:* Molecular evolution of a highly variable chromosomal region encoding the Staphylococcal exotoxin-like family of proteins. Infect Immun 71: 2827–2838.
30. Lindsay JA, Moore CE, Day NP, Peacock SJ, Witney AA, et al. (2006) Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes. J Bacteriol 188: 669–676.
31. Dziejman M, Serruto D, Tam VC, Sturtevant D, Diraphat P, et al. (2005) Genomic characterization of non-O1, non-O139 Vibrio cholerae reveals genes for a type III secretion system. Proc Natl Acad Sci U S A 102: 3465–3470.
32. Perrin A, Bonacorsi S, Carbonnelle E, Talibi D, Dessen P, et al. (2002) Comparative genomics identifies the genetic islands that distinguish *Neisseria meningitidis*, the agent of cerebrospinal meningitis, from other *Neisseria* species. Infect Immun 70: 7063–7072.
33. Yao Y, Sturdevant DE, Villaruz A, Xu L, Gao Q, et al. (2005) Factors characterizing *Staphylococcus epidermidis* invasiveness determined by comparative genomics. Infect Immun 73: 1856–1860.
34. Akopyants NS, Clifton SW, Martin J, Pape D, Wylie T, Li L, et al. (2001) A survey of the Leishmania major Friedlin strain V1 genome by shotgun sequencing: A resource for DNA microarrays and expression profiling. Mol Biochem Parasitol 113: 337–340.
35. Moore DD (1993) Overview of recombinant DNA libraries. In: Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, et al., editors. Current protocols in molecular biology. New York: John Wiley & Sons. pp. 5.1.1–5.1.3.
36. Sharp PM, Li WH (1987) The codon adaptation index–A measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15: 1281–1295.
37. Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72: 248–254.
38. Maki H, Murakami K (1997) Formation of potent hybrid promoters of the mutant *llm* gene by IS256 transposition in methicillin-resistant *Staphylococcus aureus*. J Bacteriol 179: 6944–6948.
39. McAshan SK, Vergin KL, Giovannoni SJ, Thaler DS (1999) Interspecies recombination between enterococci: Genetic and phenotypic diversity of vancomycin-resistant transconjugants. Microb Drug Resist 5: 101–112.
40. Dorrell N, Mangan JA, Laing KG, Hinds J, Linton D, et al. (2001) Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. Genome Res 11: 1706–1715.
41. Leavis HL, Willems RJ, Top J, Spalburg E, Mascini EM, et al. (2003) Epidemic and nonepidemic multidrug-resistant *Enterococcus faecium*. Emerg Infect Dis 9: 1108–1115.
42. Willems RJ, Top J, van Den Braak N, van Belkum A, Endtz H, et al. (2000) Host specificity of vancomycin-resistant *Enterococcus faecium*. J Infect Dis 182: 816–823.
43. Saunders NA, Underwood A, Kearns AM, Hallas G (2004) A virulence-associated gene microarray: A tool for investigation of the evolution and pathogenic potential of *Staphylococcus aureus*. Microbiology 150: 3763–3771.
44. Stabler RA, Marsden GL, Witney AA, Li Y, Bentley SD, Tang CM, Hinds J (2005) Identification of pathogen-specific genes through microarray analysis of pathogenic and commensal *Neisseria* species. Microbiology 151: 2907–2922.
45. Eisen JA, Benito MI, Walbot V (1994) Sequence similarity of putative transposases links the maize mutator autonomous element and a group of bacterial insertion sequences. Nucleic Acids Res 22: 2634–2636.
46. Quintiliani R Jr., Courvalin P (1996) Characterization of Tn1547, a composite transposon flanked by the IS16 and IS256-like elements, that confers vancomycin resistance in *Enterococcus faecalis* BM4281. Gene 172: 1–8.
47. Naas T, Fortineau N, Snanoudj R, Spicq C, Durrbach A, Nordmann P (2005) First nosocomial outbreak of vancomycin-resistant *Enterococcus faecium* expressing a VanD-like phenotype associated with a vanA genotype. J Clin Microbiol 43: 3642–3649.
48. Dahl KH, Lundblad EW, Rokenes TP, Olsvik O, Sundsfjord A (2000) Genetic linkage of the vanB2 gene cluster to Tn5382 in vancomycin-resistant enterococci and characterization of two novel insertion sequences. Microbiology 146 (Pt 6): 1469–1479.
49. Grady R, Hayes F (2003) Axe-Txe, a broad-spectrum proteic toxin-antitoxin system specified by a multidrug-resistant, clinical isolate of *Enterococcus faecium*. Mol Microbiol 47: 1419–1432.
50. Yang F, Yang J, Zhang X, Chen L, Jiang Y, et al. (2005) Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. Nucleic Acids Res 33: 6445–6458.
51. Diavatopoulos DA, Cummings CA, Schouls LM, Brinig MM, Relman DA, Mooi FR (2005) *Bordetella pertussis*, the causative agent of whooping cough, evolved from a distinct, human-associated lineage of *B. bronchiseptica*. PLoS Pathog 1: e45.
52. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, et al. (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet 35: 32–40.
53. Paulsen IT, Banerjei L, Myers GS, Nelson KE, Seshadri R, et al. (2003) Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. Science 299: 2071–2074.
54. Werner G, Hildebrandt B, Witte W (2003) Linkage of erm(B) and aadE-sat4-aphA-3 in multiple-resistant *Enterococcus faecium* isolates of different ecological origins. Microb Drug Resist 9 Suppl 1: S9–16.
55. Werner G, Hildebrandt B, Witte W (2001) Aminoglycoside-streptothricin resistance gene cluster aadE-sat4-aphA-3 disseminated among multiresistant isolates of *Enterococcus faecium*. Antimicrob Agents Chemother 45: 3267–3269.
56. Champion OL, Gaunt MW, Gundogdu O, Elmi A, Witney AA, Hinds J, Dorrell N, Wren BW (2005) Comparative phylogenomics of the food-borne pathogen *Campylobacter jejuni* reveals genetic markers predictive of infection source. Proc Natl Acad Sci U S A 102: 16043–16048.
57. Howard SL, Gaunt MW, Hinds J, Witney AA, Stabler R, Wren BW (2006) Application of comparative phylogenomics to study the evolution of *Yersinia enterocolitica* and to identify genetic differences relating to pathogenicity. J Bacteriol 188: 3645–3653.
58. Melles DC, Gorkink RF, Boelens HA, Snijders SV, Peeters JK, et al. (2004) Natural population dynamics and expansion of pathogenic clones of *Staphylococcus aureus*. J Clin Invest 114: 1732–1740.
59. Pupo GM, Lan R, Reeves PR (2000) Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc Natl Acad Sci U S A 97: 10567–10572.
60. Helgason E, Okstad OA, Caugant DA, Johansen HA, Fouet A, et al. (2000) *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*–One species on the basis of genetic evidence. Appl Environ Microbiol 66: 2627–2630.
61. Groisman EA, Ochman H (1996) Pathogenicity islands: Bacterial evolution in quantum leaps. Cell 87: 791–794.
62. Sokurenko EV, Hasty DL, Dykhuizen DE (1999) Pathoadaptive mutations:

Gene loss and variation in bacterial pathogens. Trends Microbiol 7: 191–195.
63. Rose M, Charlesworth B (1980) A test of evolutionay theories of senescence. Nature 287: 141–142.
64. Bonten MJ, Hayden MK, Nathan C, van Voorhis J, Matushek M, et al. (1996) Epidemiology of colonisation of patients and environment with vancomycin-resistant enterococci. Lancet 348: 1615–1619.
65. Bonten MJ, Hayden MK, Nathan C, Rice TW, Weinstein RA (1998) Stability of vancomycin-resistant enterococcal genotypes isolated from long-term-colonized patients. J Infect Dis 177: 378–382.
66. Dunne WM Jr, Wang W (1997) Clonal dissemination and colony morphotype variation of vancomycin-resistant *Enterococcus faecium* isolates in metropolitan Detroit, Michigan. J Clin Microbiol 35: 388–392.
67. Jordens JZ, Bates J, Griffiths DT (1994) Faecal carriage and nosocomial spread of vancomycin-resistant *Enterococcus faecium*. J Antimicrob Chemother 34: 515–528.
68. Mohn SC, Ericson Solid JU, Jureen R, Steen VM, Langeland N (2005) Novel insertion sequence between the *pbp5* and *psr* genes is associated with ampicillin resistance in *Enterococcus faecium* [dissertation].
69. Timmers GJ, van der Zwet WC, Simoons-Smit IM, Savelkoul PH, Meester HH, et al. (2002) Outbreak of vancomycin-resistant *Enterococcus faecium* in a haematology unit: Risk factor assessment and successful control of the epidemic. Br J Haematol 116: 826–833.
70. Arduino RC, Murray BE, Rakita RM (1994) Roles of antibodies and complement in phagocytic killing of enterococci. Infect Immun 62: 987–993.
71. Woodford N, Soltani M, Hardy KJ (2001) Frequency of *esp* in *Enterococcus faecium* isolates. Lancet 358: 584.
72. Henwood CJ, Livermore DM, Johnson AP, James D, Warner M, et al. (2000) Susceptibility of gram-positive cocci from 25 UK hospitals to antimicrobial agents including linezolid. The Linezolid Study Group. J Antimicrob Chemother 46: 931–940.
73. Willems RJ, Top J, van Den Braak N, van Belkum A, Mevius DJ, et al. (1999) Molecular diversity and evolutionary relationships of Tn1546-like elements in enterococci from humans and animals. Antimicrob Agents Chemother 43: 483–491.
74. Borucki MK, Kim SH, Call DR, Smole SC, Pagotto F (2004) Selective discrimination of *Listeria monocytogenes* epidemic strains by a mixed-genome DNA microarray compared to discrimination by pulsed-field gel electrophoresis, ribotyping, and multilocus sequence typing. J Clin Microbiol 42: 5270–5276.
75. Aymerich T, Holo H, Havarstein LS, Hugas M, Garriga M, et al. (1996) Biochemical and genetic characterization of enterocin A from *Enterococcus faecium,* a new antilisterial bacteriocin in the pediocin family of bacteriocins. Appl Environ Microbiol 62: 1676–1682.
76. Bozdogan B, Berrezouga L, Kuo MS, Yurek DA, Farley KA, et al. (1999) A new resistance gene, *linB,* conferring resistance to lincosamides by nucleotidylation in *Enterococcus faecium* HM1025. Antimicrob Agents Chemother 43: 925–929.
77. Dutka-Malen S, Evers S, Courvalin P (1995) Detection of glycopeptide resistance genotypes and identification to the species level of clinically relevant enterococci by PCR. J Clin Microbiol 33: 1434.
78. Eaton TJ, Gasson MJ (2001) Molecular screening of *Enterococcus* virulence determinants and potential for genetic exchange between food and medical isolates. Appl Environ Microbiol 67: 1628–1635.
79. Fines M, Perichon B, Reynolds P, Sahm DF, Courvalin P (1999) VanE, a new type of acquired glycopeptide resistance in *Enterococcus faecalis* BM4405. Antimicrob Agents Chemother 43: 2161–2164.
80. Gevers D, Danielsen M, Huys G, Swings J (2003) Molecular characterization of *tet*(M) genes in *Lactobacillus* isolates from different types of fermented dry sausage. Appl Environ Microbiol 69: 1270–1275.
81. Luna VA, Coates P, Eady EA, Cove JH, Nguyen TT, et al. (1999) A variety of gram-positive bacteria carry mobile *mef* genes. J Antimicrob Chemother 44: 19–25.
82. Min YH, Jeong JH, Choi YJ, Yun HJ, Lee K, et al. (2003) Heterogeneity of macrolide-lincosamide-streptogramin B resistance phenotypes in enterococci. Antimicrob Agents Chemother 47: 3415–3420.
83. Soltani M, Beighton D, Philpott-Howard J, Woodford N (2000) Mechanisms of resistance to quinupristin-dalfopristin among isolates of *Enterococcus faecium* from animals, raw meat, and hospital patients in Western Europe. Antimicrob Agents Chemother 44: 433–436.
84. Teng F, Kawalec M, Weinstock GM, Hryniewicz W, Murray BE (2003) An *Enterococcus faecium* secreted antigen, SagA, exhibits broad-spectrum binding to extracellular matrix proteins and appears essential for *E. faecium* growth. Infect Immun 71: 5033–5041.
85. Vakulenko SB, Donabedian SM, Voskresenskiy AM, Zervos MJ, Lerner SA, et al. (2003) Multiplex PCR for detection of aminoglycoside resistance genes in enterococci. Antimicrob Agents Chemother 47: 1423–1426.
86. Vankerckhoven V, Van Autgaerden T, Vael C, Lammens C, Chapelle S, et al. (2004) Development of a multiplex PCR for the detection of *asa1, gelE, cylA, esp,* and *hyl* genes in enterococci and survey for virulence determinants among European hospital isolates of *Enterococcus faecium*. J Clin Microbiol 42: 4473–4479.
87. Werner G, Hildebrandt B, Witte W (2001) The newly described *msrC* gene is not equally distributed among all isolates of *Enterococcus faecium*. Antimicrob Agents Chemother 45: 3672–3673.
88. Xu Y, Singh KV, Qin X, Murray BE, Weinstock GM (2000) Analysis of a gene cluster of *Enterococcus faecalis* involved in polysaccharide biosynthesis. Infect Immun 68: 815–823.
89. Nallapareddy SR, Weinstock GM, Murray BE (2003) Clinical isolates of *Enterococcus faecium* exhibit strain-specific collagen binding mediated by Acm, a new member of the MSCRAMM family. Mol Microbiol 47: 1733–1747.
90. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.
91. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Statist Soc Ser B 57: 289–300.
92. Holmes EC, Urwin R, Maiden MC (1999) The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. Mol Biol Evol 16: 741–749.