**Comparative Genomics of Enterococci: Variation in Enterococcus faecalis, Clade Structure in E. faecium, and Defining Characteristics of E. gallinarum and E. casseliflavus.**

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1128/mBio.00318-11">http://dx.doi.org/10.1128/mBio.00318-11</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>American Society for Microbiology</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Wed Nov 28 10:24:26 EST 2018</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/73611">http://hdl.handle.net/1721.1/73611</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher’s policy and may be subject to US copyright law. Please refer to the publisher’s site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
Comparative Genomics of Enterococci: Variation in Enterococcus faecalis, Clade Structure in E. faecium, and Defining Characteristics of E. gallinarum and E. casseliflavus

Comparative Genomics of Enterococci: Variation in *Enterococcus faecalis*, Clade Structure in *E. faecium*, and Defining Characteristics of *E. gallinarum* and *E. casseliflavus*


ABSTRACT The enterococci are Gram-positive lactic acid bacteria that inhabit the gastrointestinal tracts of diverse hosts. However, *Enterococcus faecium* and *E. faecalis* have emerged as leading causes of multidrug-resistant hospital-acquired infections. The mechanism by which a well-adapted commensal evolved into a hospital pathogen is poorly understood. In this study, we examined high-quality draft genome data for evidence of key events in the evolution of the leading causes of enterococcal infections, including *E. faecalis*, *E. faecium*, *E. casseliflavus*, and *E. gallinarum*. We characterized two clades within what is currently classified as *E. faecium* and identified traits characteristic of each, including variation in operons for cell wall carbohydrate and putative capsule biosynthesis. We examined the extent of recombination between the two *E. faecium* clades and identified two strains with mosaic genomes. We determined the underlying genetics for the defining characteristics of the motile enterococci *E. casseliflavus* and *E. gallinarum*. Further, we identified species-specific traits that could be used to advance the detection of medically relevant enterococci and their identification to the species level.

IMPORTANT The enterococci, in particular, vancomycin-resistant enterococci, have emerged as leading causes of multidrug-resistant hospital-acquired infections. In this study, we examined genome sequence data to define traits with the potential to influence host-microbe interactions and to identify sequences and biochemical functions that could form the basis for the rapid identification of enterococcal species or lineages of importance in clinical and environmental samples.

RESULTS AND DISCUSSION Phylogenetic analysis of enterococci. We recently announced the public release of genome sequence data for 28 enterococcal strains of diverse origin (8) (see Table S1 in the supplemental material). The 16 *E. faecalis* genomes sequenced represent the deepest nodes in the MLST phylogeny, providing the greatest diversity. The strains include those of clinical, animal, and insect origins and were isolated from 1926 to 2005 (9). These strains represent approximately 80 years of enterococcal evolution, spanning the periods prior to and during widespread antibiotic use. Additionally, the genomes of 6 *E. faecium*, 1 *E. gallinarum*, and 3 *E. casseliflavus* clinical isolates from 2001 to 2005 and 2 human fecal *E. faecium* strains were examined.

The enterococci are a diverse group of Gram-positive gastrointestinal (GI) tract colonizers with lifestyles ranging from intestinal symbiont to environmental persister to multidrug-resistant nosocomial pathogen (1, 2, 3). Enterococci are used in food production, in probiotic products, and for tracking fecal contamination. They are also of regulatory and industrial interest. In this study, we examined high-quality draft genome data for evidence of key events in the evolution of the leading causes of enterococcal infections, including *E. faecalis*, *E. faecium*, *E. casseliflavus*, and *E. gallinarum*. We characterized two clades within what is currently classified as *E. faecium* and identified traits characteristic of each, including variation in operons for cell wall carbohydrate and putative capsule biosynthesis. We examined the extent of recombination between the two *E. faecium* clades and identified two strains with mosaic genomes. We determined the underlying genetics for the defining characteristics of the motile enterococci *E. casseliflavus* and *E. gallinarum*. Further, we identified species-specific traits that could be used to advance the detection of medically relevant enterococci and their identification to the species level.
OrthoMCL (10) was used to identify ortholog groups in the 30 enterococcal genomes. Ortholog groups represented in all 31 genomes were considered core groups, which were further subdivided into single-copy (1 gene copy in each genome) and multi-copy (>1 gene copy exists in at least 1 genome). Genes not clustered were considered orphans. A phylogenetic tree generated from the concatenated sequences of 847 single-copy core genes is shown in Fig. 1. Relationships among the 18*E.*faecalis strains, despite their diverse origins, cannot be fully resolved by this analysis (based on lack of bootstrap support for branches within the *E.*faecalis branch; inset, Fig. 1). As expected, *E.*casseliflavus and *E.*gallinarum branch separately, supporting their designation as different species. Importantly, two clades were identified within the species *E.*faecium, as had been inferred by comparative genome hybridization, which suggested that hospital-associated isolates, including ST17 and ST18 isolates, may make up a distinct subspecies within *E.*faecium (11). The 3 vancomycin-resistant *E.*faecium strains in our collection are members of clade A, while the 2 human fecal isolates are members of clade B (Fig. 1). To quantify the relationships among these strains, we generated average nucleotide identity (ANI) plots (Fig. 2), which have been used to query and refine prokaryotic species definitions (12, 13).

*E.*faecium. The *E.*faecium ANI analysis refines phylogenetic relationships among clade A and clade B strains (Fig. 2). Within clade A, ST17 strain 410 and double-locus variants (DLVs) 933 (ST18) and 502 (ST203) are closely related (99.2 to 99.4% ANI) whereas strains 501 (ST52) and 408 (ST582, an ST17 DLV) have lower ANI values with those strains, and each other (96.9 to 98.2% ANI). Similar ANI values were observed among clade B strains (97.9 to 99.4%). However, pairwise comparisons of clade A and clade B strains ranged from 93.9 to 95.6% ANI, overlapping an ANI species line of 94 to 95%. ANI values of 94 to 95% correlate with experimentally derived 70% DNA-DNA hybridization values, a commonly accepted threshold for species designation (12, 13, 14). Clade A and clade B may be endogenous to the GI tracts of different hosts and now coexist among human flora as a result of

![Core gene tree.](image-url)

FIG 1  Core gene tree. Concatenated sequences of 847 genes core to 30 enterococci and the outgroup species *L.*lactis were aligned, and a phylogenetic tree was generated using RAxML with bootstrapping. The bootstrap value for all nodes outside the *E.*faecalis clade is 100. *E.*faecium clades A (blue) and B (red) are indicated.
antibiotic elimination of competitors, or clade A and clade B may be diverging from each other as a result of antibiotic use and ecological isolation (less likely because of the short time involved).

For the 8 E. faecium strains in our collection, the two clades are recapitulated using the 7 housekeeping genes selected for E. faecium MLST (see Fig. S1 in the supplemental material). Between clade A and clade B strains, the nucleotide identities of concatenated MLST sequences range from 96.2 to 96.9% (compared to a 93.9 to 95.6% ANI range). To determine whether a single marker is representative of either E. faecium clade, we examined the distribution of individual MLST alleles among the E. faecium STs assigned to clade A or clade B (see Fig. S1 in the supplemental material). A “minority allelic population” composed of 5 divergent STs was reported in seminal E. faecium MLST work (15). The 5 divergent STs (ST39, ST40, ST60, ST61, and ST62) identified by that study belong to clade B (see Fig. S1 in the supplemental material). The genomes of 7 additional E. faecium strains were recently sequenced (7), and we used MLST to assign them to clade A or B (see Fig. S1 in the supplemental material). The assignment of one of these strains, E. faecium E980, to clade B is consistent with previous analyses demonstrating the phylogenetic distance of this strain from the other 6 (clade A) strains in that sequencing collection (7). In the first-pass analysis, the allele adk-6, which differs from the ST17 allele adk-1 at 3 synonymous sites, was observed to occur almost exclusively in clade B strains (see Fig. S1 in the supplemental material). To further explore the distribution of adk alleles among E. faecium isolates, we extracted sequences of all 617 available STs in the E. faecium MLST database and determined the extents of identity to an ST17 (clade A) reference. In the MLST database, adk-1, adk-5, and adk-6 are the most abundant adk alleles, representing 87% of the E. faecium STs. Of the 85 STs possessing adk-6, 66 (78%) share 96 to 96.9% nucleotide identity with ST17, comparable to that observed for clade B-ST17 comparisons. Conversely, adk-1 and adk-5 occur primarily in STs with ≥99% identity to ST17. These data suggest that adk allele exchange is restricted, perhaps resulting from a barrier to DNA uptake such as clustered regularly interspaced short palindromic repeats (CRISPR)-cas defense and/or from the proximity of adk to the replication origin (Fig. 3).

E. faecium 408 is a DLV of ST17 that possesses adk-6 and ddl-13 (see Fig. S1 in the supplemental material). Because adk-6 occurs mostly among strains with lower relatedness to ST17, and ddl-13 is present in two clade B strains (see Fig. S1 in the supplemental material), we were curious about whether these alleles were acquired by recombination. Genome mosaicism is evident in E. faecium clade A strains 408 and 501 (Fig. 3). The occurrence of adk-6 and ddl-13 within a hybrid region in E. faecium 408 (Fig. 3; see data set S1 in the supplemental material) supports the acquisition of this region from a clade B strain. The putative genome defense system EfmtCRISPR1-cas (16), present in 2 of 3 clade B strains and in E. faecium 408 (see Table S1 in the supplemental material), occurs within this region, suggesting that CRISPR-cas was acquired by E. faecium 408 from a clade B strain via recombination. The hybrid region in E. faecium 501 includes pbp5 (Fig. 3; see data set S1 in the supplemental material), which can confer ampicillin resistance. Our results indicate that pbp5-S was acquired by E. faecium 501 from a clade B strain. The hybrid region in 501 is flanked by a putative phage integrase (EFRG_00906) that is conserved among all of the E. faecium strains in our collection (see data set S1 in the supplemental material). We recently reported an Hfr-like mechanism for the transfer of chromosomal genes between E. faecalis strains (17), and it seems likely that a similar mechanism functions in E. faecium.

To determine whether specific traits define the two E. faecium clades, we searched for clade-specific ortholog groups present in and exclusive to all of the members of each clade. We then used representative gene sequences from each to search for similar sequences in 7 additional E. faecium genomes (7) assigned to clade A or B (see Fig. S1 in the supplemental material). Of the clade A-specific genes (see data set S2 and Table S2 in the supplemental material), 8 are associated with a locus that has high sequence identity with and almost the same gene content as the yejM-NOPQRSTUV locus of Entercocilla coli, which is significantly enriched in enteric clades (18) and also occurs in Listeria (19). The organization of this locus is similar to that of a Lactobacillus actidophilus fructooligosaccharide (prebiotic) utilization locus (20). Of the genes unambiguously assigned to clade B (see data set S2 and Table S2 in the supplemental material), 5 encode putative transcriptional regulators with protein domain hits to Mga or Rgg, regulators of virulence, competence, and cell-cell signaling in streptococci (21, 22). Two of these putative regulators are divergently transcribed from genes that are also clade B specific, including a putative thioredoxin that could modulate the redox state of cellular targets in response to oxidative stress (23). A putative phospholipase C is also clade B specific. Finally, one clade B-specific gene (EFSG_01746) was useful in identifying a genomic insertion, composed of 17 genes, in E. faecium 733 (see Table S2 in the supplemental material). This region encodes a putative phosphotransferase system and a secreted hyaluronidase that could cleave the extracellular matrix of host cells. It is surprising that clade B (and not clade A, which contains all high-risk STs) strains encode a number of secreted factors that could interact with eukaryotic cell surfaces. This suggests that clade B strains may be more closely associated with host tissues in the GI tract than clade A strains are, which possibly contributes to their persistence in the
GI tract, whereas clade A strains may be more transient and associated with the GI lumen, which contributes to their dissemination.

*E. faecalis.* In contrast to *E. faecium*, little phylogenetic divergence was observed among *E. faecalis* strains (Fig. 2). Among 306 pairwise comparisons, ANI varies within a narrow range (97.8 to 99.5%). Instead, shared gene content among these strains varies (70.9 to 96.5%). For example, strain T11 shares 96.5% of its 2,511 genes with ST16 strain V583, while V583 shares only 72.8% of its 3,265 genes with T11; they possess 99.5% ANI in the genes that they share. The genome size of T11 is smaller than that of V583 (2.74 Mb versus 3.36 Mb) and is similar to that of the oral isolate OG1RF (24), likely representing the minimal *E. faecalis* genome. For all 18 *E. faecalis* strains, genome sizes vary between the extremes of T11 and V583 (see Table S1 in the supplemental material). We recently proposed that loss of CRISPR-cas in founders of modern *E. faecalis* high-risk MLST lineages facilitated the influx of acquired antibiotic resistance genes and other mobile traits into these lineages (16). Genome size distribution significantly differs between strains possessing or lacking CRISPR-cas (P = 0.026;
one-tailed Wilcoxon rank-sum test), with a greater average genome size in strains lacking CRISPR-cas (3.1 Mbp versus 2.9 Mbp). The distribution of domain motifs associated with mobile elements is significantly different in strains with genomes >3 Mb in size ($P < 0.05$), including the plasmid mobilization MobC domain (PF05713; $P = 0.001$), the antirestriction protein ArDA (PF07275; $P = 0.032$), the replication initiation factor domain (PF02486; $P = 0.001$), a plasmid addiction toxin domain (TIGR02385; $P = 0.032$), and a transposase domain (PF01526; $P = 0.021$). This supports a model where increased genome size is the result of mobile element accretion, consistent with the proposition that compromised genome defense facilitated the accretion of mobile elements (16), resulting in larger genomes.

We analyzed the 18 *E. faecalis* genomes for mosaicism (Fig. 4). Thirteen variable regions were previously defined for *E. faecalis* genomes using comparative genome hybridization to a V583-genome mosaicism plot. The outermost ring shows *E. faecalis* V583 chromosomal (scaffold 4) and plasmid scaffolds (scaffold 1, pTEF2; scaffold 2, pTEF3; scaffold 3, pTEF1), with each gene represented as a radial position along the ring. Each of the remaining 17 *E. faecalis* genomes is represented by the rings below V583. Genes are colored by phylogenetic distance from *E. faecalis* V583 (from dark to light green with increasing phylogenetic distance), as determined by individual gene trees built from ortholog groups. The strains shown, from the outermost to the innermost rings, are V583, T11, OG1RF, Merz96, T8, T2, D6, X98, T3, T1, Fly1, CH188, HIP11704, ATCC 4200, E1Sol, AB01/DG, D55, and JH1. The locations of *E. faecalis* variable regions are shown (9). A, integrated plasmid; B, prophage 1; C, *E. faecalis* pathogenicity island; D, prophage 2; E, prophage 3; F, putative island; G, prophage 4; H, prophage 5; I, putative island; J, vancomycin resistance (*vanB*) transposon; K, integrated plasmid; L, prophage 6; M, prophage 7.
TABLE 1 ANI and shared-gene analyses of \textit{E. casseliflavus} and \textit{E. gallinarum}

<table>
<thead>
<tr>
<th>Strain</th>
<th>EC10</th>
<th>EC20</th>
<th>EC30</th>
<th>EG2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC10</td>
<td>98, 85</td>
<td>100, 99</td>
<td>74, 72</td>
<td>65–66, 51–55</td>
</tr>
<tr>
<td>EC20</td>
<td>98, 88</td>
<td>98, 88</td>
<td>74, 72</td>
<td>65–66, 52–56</td>
</tr>
<tr>
<td>EC30</td>
<td>100, 99</td>
<td>98, 86</td>
<td>74, 72</td>
<td>65–66, 51–55</td>
</tr>
<tr>
<td>EG2</td>
<td>74, 78</td>
<td>74, 77</td>
<td>74, 78</td>
<td>65–67, 53–60</td>
</tr>
</tbody>
</table>

\textsuperscript{a} The data shown are for genome 1 (left) compared to genome 2 (top). Values were rounded to the closest whole number.

\textsuperscript{b} Ranges of values are shown.

based microarray (9). Regions of mosaicism were detected in strains Merz96, JH1, and T2 overlapping the \textit{E. faecalis} pathogenicity island (25) and two putative genomic islands containing \textit{Tn916}-like genes (5), respectively. These results are consistent with conjugative acquisition of these islands and surrounding sequence by Merz96, JH1, and T2 from strains closely related to \textit{V.583} (17). Collectively, much of the diversity of \textit{E. faecalis} can be attributed to the accretion of mobile genetic elements on a largely conserved genomic backbone, with those mobile elements facilitating recombinatorial exchange of chromosomally encoded traits.

The motile enterococci. Very little is known about the genomes of \textit{E. casseliflavus} and \textit{E. gallinarum}. Once thought to be associated primarily with vegetation (\textit{E. casseliflavus} (26)) and fowl (\textit{E. gallinarum} (27)) and only rarely found in humans, these species appear to be increasingly implicated in infections and hospital outbreaks (28, 29). Motility is a defining characteristic of most strains of \textit{E. casseliflavus} and \textit{E. gallinarum}, while \textit{E. casseliflavus} additionally produces a yellow pigment (30); however, there has been confusion because of phenotype variation (3). ANI analysis confirms that the \textit{E. casseliflavus} and \textit{E. gallinarum} strains in our collection, which possess \textasciitilde74\% ANI in shared genes, are members of two separate species (Table 1). Motile enterococci are reported to have \textasciitilde3 or 4 terminal or lateral flagella per cell (31). In the \textit{E. casseliflavus} and \textit{E. gallinarum} genomes, we identified conserved gene clusters encoding proteins predicted to synthesize, export, and power a flagellum, as well as a chemotactic response system (see data set S3 in the supplemental material). Most of the proteins predicted to be encoded by the representative \textit{E. casseliflavus} EC10 motility gene cluster have best BLASTP hits to \textit{Lactobacillus ruminis} proteins (see data set S3 in the supplemental material) (32).

Bacterial motility is often regulated by the second messenger cyclic di-GMP (c-di-GMP), as are attachment to surfaces and production of extracellular polysaccharides (33). We identified putative diguanylate cyclases possessing GGDEF domains (for c-di-GMP synthesis) and phosphodiesterases possessing EAL domains (for c-di-GMP turnover) in all 3 \textit{E. casseliflavus} strains (see data set S3 in the supplemental material) but not in \textit{E. gallinarum}. Two GGDEF proteins and one EAL protein are encoded 5’ to a predicted protein possessing glycosyltransferase, cellulose synthase, and PilZ domains (see data set S3 in the supplemental material). This protein shares high identity with a \textit{Clostridium difficile} protein thought to be regulated by c-di-GMP (CD2545 [34]) (see data set S3 in the supplemental material). PilZ domains bind c-di-GMP (33), and it is possible that this domain regulates cellulose synthesis or the production of another extracellular polymer in \textit{E. casseliflavus}.

\textit{E. casseliflavus} produces a cell-associated carotenoid pigment thought to facilitate its environmental persistence by protecting against photooxidation (35). \textit{Streptococcus aureus} also produces a carotenoid pigment and virulence factor, staphyloxanthin, that protects it from host-induced oxidative damage and antimicrobial peptides, and inhibitors show promise as novel therapeutics (36). Compared to \textit{S. aureus} CrtOPQMN, \textit{E. gallinarum} and \textit{E. casseliflavus} share CrtM and CrtN homologues that catalyze the first steps of staphyloxanthin biosynthesis (37). However, only \textit{E. casseliflavus} possesses CrtO, CrtP, and CrtQ (see data set S3 in the supplemental material). Most ligand interaction sites (36) are conserved in the \textit{E. casseliflavus} and \textit{E. gallinarum} CrtM proteins (see Fig. S2 in the supplemental material), suggesting that CrtM inhibitors could be usefully applied to these bacteria.

\textit{E. faecalis} and \textit{E. faecium} extracellular polysaccharides. Cell wall polymers produced by \textit{E. faecalis} include a lipoteichoic acid (LTA) with a poly(glycerol-phosphate) backbone (38, 39); a putative wall teichoic acid (WTA) composed of glycerol, glucose, and phosphate (40); and a rhamnopolysaccharide (the enterococcal polysaccharide antigen or Epa) composed of rhmnnose, N-acetylgalactosamine, N-acetylgalactosamine, glucose, and galactose (40, 41, 42). The \textit{E. faecalis} epa locus directs the synthesis of the Epa polymer (43), although the biochemical functions and essentiality of most Epa proteins are unknown. The production of an antiphagocytic capsule composed of galactose, glucose, and phosphate is strain variable in \textit{E. faecalis} and dependent on the presence of the \textit{cps} locus (9, 40). Other than the \textit{E. faecalis} epa and cps loci and \textit{E. faecalis} bgSA and bgSB, which are involved in LTA biosynthesis (44), the genetic bases of extracellular polymer biosynthesis in enterococci are largely uncharacterized, as is the genetic basis of variable phagocytosis resistance in the species \textit{E. faecium} (45). We therefore examined the distributions of epa, cps, and a predicted LTA biosynthesis pathway (46) and searched for new loci potentially important for decoration of the enterococcal cell surface.

As expected based on previous work (42), the entire \textit{epa} locus—comprising \textit{epaA} to \textit{epaI}—is core to the species \textit{E. faecalis}. An \textit{epa} locus varying in organization and content from that of \textit{E. faecalis} is also core to \textit{E. faecium}. In \textit{E. faecium}, the genes are ordered \textit{epaABCDDEFGH}-epaPQ-epaLM-[\textit{E. faecium}-specific gene]-epaOR. The intervening \textit{E. faecium}-specific gene encodes a protein with N-terminal similarity to \textit{E. faecalis} EpaN but was not identified as being orthologous to \textit{epaN} by OrthoMCL. Both the \textit{E. faecalis} EpaN and \textit{E. faecium}-specific proteins have a predicted S-adenosylmethionine binding site in the N terminus but are divergent in C-terminal sequence. The \textit{E. faecium} epa locus may direct the synthesis of a previously reported \textit{E. faecium} tetratetraglycan composed of galactose, rhamnose, N-acetylgalactosamine, glucose, and phosphate (47). The conservation of most of the \textit{epa} locus suggests that if Epa biosynthesis enzymes were targeted with novel antimicrobials, those antimicrobials could be effective against the enterococcal species of greatest concern to human health. Proteins predicted to be involved in \textit{E. faecalis} V583 LTA biosynthesis (46) were additionally identified as being core to \textit{E. faecalis} and \textit{E. faecium}, as well as the other enterococcal species in our collection (see data set S4 in the supplemental material).

Potentially important variation in \textit{E. faecalis} and \textit{E. faecium} epa
FIG 5  Putative *E. faecium* capsule loci. The core wzi-wzd-wze-wvb genes and downstream variable region are shown for 8 *E. faecium* strains. Conserved anchor genes flanking the core and variable regions are indicated. Variable region genes are colored by BLASTP and Pfam conserved-domain hits shown in data set S4 in the supplemental material. Multiple Pfam domains were collapsed into categories (for example, glycosyltransferases). Only the most abundant Pfam categories are shown. Orphan genes not grouped by OrthoMCL are indicated. Contig gaps in scaffolds are indicated by black bars; the size of the black bar is proportional to the number of N’s inserted during genome assembly. In *E. faecium* 502, the nucleotide sequence of *wze* is conserved but is interrupted by a contig gap, and a scaffold gap occurs between *wzh* and the EFVG_00414 flanking gene homologue (indicated by vertical slashes). The drawing is to scale, and a scale bar is shown.

operons occur between orthologs of *epaR* (EF2177 in V583) and EF2165 (see Fig. S3 in the supplemental material). Variation in this region was previously reported between *E. faecalis* strains V583 and OG1RF (42). The variable regions of the 26 *E. faecalis* and *E. faecium* strains, which consist of 37 ortholog groups and 11 orphans (excluding transposases), encode predicted glycosyltransferases and other proteins with likely roles in extracellular polysaccharide production (see data set S4 in the supplemental material). The 3 vancomycin-resistant *E. faecium* CC17 strains in our collection possess a unique *epa* locus configuration with putative sialic acid biosynthesis (*neuABCD*) genes within the variable region, and a divergently transcribed, predicted β-lactamase gene inserted in the core *epa* region between *epaO* and *epaR* (see Fig. S3 in the supplemental material). Sialic acid decoration by pathogenic bacteria is thought to be a form of molecular mimicry that interferes with detection by the host immune system (48). The *neuABCD* genes are not clade specific and are also present in clade B strain E980 and clade A strain U0317, suggesting that the *epa* region can be either lost or transferred between *E. faecium* clades. The potential for sialic acid decoration on high-risk, vancomycin-resistant strains has important implications for vaccine development. We additionally identified putative WTA biosynthesis genes (*tagFl/tarF* and *tagDl/tarD* [49]) in a subset of *E. faecalis* and *E. faecium* *epa* variable regions (see Fig. S3 and data set S4 in the supplemental material). The core *epaA* and *epaLM* genes encode proteins similar to *Bacillus subtilis* TagO and TagGH (43), which catalyze the initial step in WTA synthesis and the export of the assembled WTA polymer, respectively (49).

Phagocytosis resistance in *E. faecalis* is associated with capsule production (40), which is a variable trait of that species (9). We examined the distribution of the *E. faecalis* *cps* capsule locus and found that it occurs only in *E. faecalis*, with little variation (see data set S4 in the supplemental material). We identified a novel capsule-like region in *E. faecium* (Fig. 5; see data set S4 in the supplemental material) that includes a phosphoregulatory system conserved among all species except *E. faecalis* (see data set S4 in the supplemental material). The proteins encoded are similar to the YwqCDE proteins of *B. subtilis* and the CpsBCD proteins of *Streptococcus pneumoniae* (Table 2), protein-tyrosine kinase/phosphorylase regulatory systems that regulate UDP-glucose dehydrogenase activity (50) and capsule production (51), respectively. This system is located 5′ to a variable cohort of putative extracellular polymer biosynthesis genes in *E. faecium* (Fig. 5; see data set S4 in the supplemental material). This genetic configuration is similar to that of *S. pneumoniae* *cpsABCD*, which is core to the capsule biosynthesis loci of 90 pneumococcal serotypes (52). In *E. faecium*, these genes are oriented *cpsACDB* (Table 3; see data set S4 in the supplemental material). Because *cps* nomenclature is already used for the unrelated *E. faecalis* capsule biosynthesis locus (40), we refer here to the *E. faecium* *cpsACDB* genes by the alter-
nate *S. pneumoniae* gene names wzd, wze, wzh, and wzh, respectively (52).

We subjected the variable region between wzh and EFVG_00414 in each *E. faecium* genome to BLASTP and conserved-domain analyses (see data set S4 in the supplemental material). In *S. pneumoniae* capsule production, a sugar transferase (WchA) initiates capsule biosynthesis at the membrane by transferring an initial sugar to an undecaprenyl-phosphate carrier; additional repeat units are added by the Wzx flippase, and additional sugars are then transferred to the repeat unit by cosyltransferases and enzymes likely to generate activated sugar moieties for transfer (Fig. 5). Wzx and Wzy homologues were identified in each of the 8 *E. faecium* variable regions, as were many predicted glycosyltransferases and enzymes likely to generate activated sugar moieties for transfer (Fig. 5). Wzx and Wzy homologues were identified in some, but not all, *E. faecium* strains (Fig. 5). Phagocytosis resistance is variable among *E. faecium* isolates (45), and no mechanism has been reported for this clinically relevant phenotype. It is likely that the putative capsule locus and/or variable *epa* loci described here contribute. We additionally identified putative wzd, wze, wzh sequences in the enterococcal species *E. saccharolyticus* and *E. italicus* (data not shown), suggesting that, at least among the sequenced enterococci, *E. faecalis* is the exception, lacking this capsule biosynthesis pathway.

**Species-specific signatures.** We used a combination of data, including Biolog carbon substrate catabolism analysis of a subset of our strains (see Materials and Methods), ortholog groups, and

| TABLE 2 | BLAST and Pfam analyses of putative phosphoregulatory system present in *E. faecium*, *E. casseliflavus*, and *E. gallinarum* |
|-----------------|-----------------------------|-----------------------------|
| Representative locus | Pfam hit (E value) | BLASTP best hit* |
| **S. pneumoniae** | **TIGR4** | **B. subtilis** |
| EFPG_02020 | None | None |
| EFPG_02021 | LytR_cpsA_psr (2.7e-49) | SP_1942, transcriptional regulator (9e-75, 65%, 95%)a |
| EFPG_02022 | Wzz (1.1e-19) | Capsular polysaccharide biosynthesis protein |
| EFPG_02023 | ChiA (1.8e-14) | Capsular polysaccharide biosynthesis protein |
| EFPG_02024 | None | None |

* From *E. faecium* 933.
* Pfam hits with E values of \( \leq 10^{-5} \) are shown.
* Values in parentheses are E value, % similarity, and % query coverage.
* The second-best hit was capsular polysaccharide biosynthesis protein Cps4A (7e-31, 79%, 55%).

the Comparative Metabolism tool within a computationally generated database of predicted metabolic pathways (Enterocyc; http://enterocyc.broadinstitute.org), to identify species-specific biochemical traits and nucleotide sequences that could augment existing methodologies to classify enterococcal isolates. For Biolog analysis, we focused on carbon substrates having the strongest species-specific signatures (Table 3).

Inulin fermentation was reported to be a distinguishing characteristic of motile enterococci (31), and our Biolog analysis confirmed that inulin metabolism is restricted to *E. casseliflavus*. Additionally, genes for acetoin dehydrogenase (ECAG_00219 to ECAG_00222), which converts acetoin to acetaldehyde and acetyl coenzyme A, are unique to *E. casseliflavus*. Catabolism of \( \alpha \)-ketoketohal determine specific to *E. faecalis*, as are genes (*bkdDABC; EF1661 to EF1658) encoding a previously characterized branched-chain \( \alpha \)-ketohaldehydrogenase complex (53). The *eutBC* genes (EF1629 and EF1627, respectively) directing ethanolamine catabolism and the ornate dehydrogenase gene *fdhA* (EF1630) are also specific. Catabolism of the cyclic oligosaccharide \( \gamma \)-cyclodextrin, an additive in pharmaceuticals and other products (54), is enriched in *E. faecium*, while a gene for glutaminase (EF1659), which converts glutamate to glutamine, is unique to *E. faecium* specific. Probes targeting c-di-GMP signaling (see data set S3 in the supplemental material) and acetoin dehydrogenase genes (*E. casseliflavus*; the *eutBC*, *fdhA*, and *bdKDABC* genes (*E. faecalis*); and the glutaminase gene (*E. faecium*) could be used to discriminate these different enterococcal

| TABLE 3 | Biolog carbon catabolic substrates with the strongest species-specific signatures |
|-----------------|-----------------------------|-----------------------------|
| OD_{590} ratioa | E. casseliflavus | E. faecium |
| Chemical | EC10 | EC20 | EC30 | 408 | 410 | 933 | Com12 | Com15 | 733 | V583 | T8 | T1 | X98 | E1Sol | AR01/D.G. Fly1 | T3 |
| **\( \alpha \)-Ketoketohaldehydrogenase** | 4.6, 3.5 | 3.8, 3.4 | 3.7, 4.5 | 3.8, 3.7 | 3.5, 2.7 | 5.4, 5.6 | 3.6, 3.6 | 3.8, 3.8 | 2.4, 2.6 |
| **\( \gamma \)-Cyclodextrin** | 7.0, 7.2 | 7.4, 6.2 | 4.4, 3.4 | 7.5, 6.8 | 4.7, 4.3 | 6.7, 6.8 |
| Inulin | 5.2, 2.8 | 3.6, 4.0 | 4.2, 4.6 |

* Shown are ratios of the OD_{590} in a carbon-containing well to the OD_{590} of a no-carbon-added control well after 48 h incubation at 37°C. A ratio >2 was considered a positive result. Each strain was tested twice, and the data shown are for both trials. Ratios of <2 are not shown. All ratios for *E. gallinarum* were <2.
species. We did not detect _E. faecium_ clade-specific metabolism using Biolog analysis or EnterocCyC predictions; however, clade-specific gene sequences (see Table S2 in the supplemental material) could be used as molecular probes.

**Perspectives.** A comparative genomic approach was used to address gaps in our knowledge of Enterococcus, a bacterial genus of importance to human health. Our phylogenetic analysis of _E. faecium_ reveals and quantifies the distance that separates two distinct phylogenetic clades between which gene exchange has occurred. _E. faecium_ clade-specific genes (see data set S2 and Table S2 in the supplemental material) are suggestive of different niches for clade A and clade B _E. faecium_ in the GI tract. Additionally, conserved and variable pathways that appear to be important for cell wall polymer biosynthesis were identified. In contrast to _E. faecium_, a multicell structure was not observed in _E. faecalis_, for which the acquisition of mobile elements appears to be a major source of metabolic polymorphism. Antibiotic resistance and pathogenicity island acquisition of mobile elements appears to be a major source of second messenger only in some genes related to the formation and function of the c-di-GMP second messenger only in _E. casseliflavus_ and _E. gallinarum_. The importance of the occurrence of motility operons in both but of genes related to the formation and function of the c-di-GMP second messenger only in _E. casseliflavus_ and the impact of motility on metabolism represent interesting areas for future exploration.

**MATERIALS AND METHODS**

**Enterococcal strains and genome sequencing.** _E. faecalis_ strains were selected for genome sequencing to represent the diversity of a collection of 106 isolates previously characterized (9). The _E. faecalis_ V583 and OG1RF genome sequences were previously reported (5, 24). The _E. casseliflavus_, _E. gallinarum_, and _E. faecium_ strains were obtained from a repository of clinical isolates (Eurofins Medinet). _E. faecium_ Com12 and Com15 were isolated from fecal of healthy human volunteers under Schepens Eye Research Institute Institutional Review Board protocol 2006-02. Identification of Pathogenic Lineages of _E. faecalis_. _E. faecium_ STs were previously determined (16, 55), and _E. faecium_ MLST data were accessed at [http://efaecium.mlst.net](http://efaecium.mlst.net). The sequencing, assembly, annotation, and rapid public release of these genome sequences have been previously described (8).

**Standard analyses, OrthoMCL, and EnterocCyC.** Orthogene groups were identified using OrthoMCL (10), with an all-versus-all BLAST cutoff of 1e-5. _Lactococcus lactis_ subsp. cremoris SK11 plasmid (NC_008503 to NC_008507) and chromosomal (NC_008527) genes were included as the outgroup. Coding sequences were aligned using Muscle (56), and poorly conserved regions were trimmed using trimAl (57). All trimmed alignments were concatenated and used to estimate phylogeny using maximum likelihood and 1,000 bootstrap trials as implemented by RAxML (58) using the rapid-bootstrap option and the GTRMIX model. Conserved protein domains were predicted using HHpred (59) to search the Pfam (release 24; [http://pfam.janelia.org](http://pfam.janelia.org)) (60) and TIGRFam (release 10) (61) databases. The statistical significance of differences in genome size and conserved protein domain distribution was assessed using the one-tailed Wilcoxon rank sum test. Membrane helix predictions were generated with transmembrane protein topology with a hidden Markov model (14). Protein subcellular localization predictions were generated using PsortB (62). Sequence alignments and phylogenetic trees in the figures in the supplemental material were generated with ClustalW in MacVector. Enzyme Commission (EC) numbers for the proteins in EnterocCyC ([http://enterocyc.broadinstitute.org/](http://enterocyc.broadinstitute.org/)) were predicted using gene coding sequences (CDS) and BLASTX to search the KEGG database (release 56) (63) and assigning EC numbers based on the KEGG annotation. Only significant hits with an E value of <1e-10 and 70% overlap were considered. Pathways, operons, transporters, and pathway holes were predicted using the Pathway Tools software suite (64, 65). Unless otherwise noted, BLASTP and nucleotide megabLAST queries were executed against the NCBI nonredundant protein sequence, nucleotide collection, and whole-genome shotgun read databases using NCBI BLAST. Proteins encoded by the _E. casseliflavus_ EC10 motif locus were compared to a _B. subtilis_ 168 reference using BLASTP (see data set S3 in the supplemental material); the _B. subtilis_ 168 flagellum is a reference Gram-positive flagellum in the KEGG database ([http://www.genome.jp/kegg-bin/show_pathway?bsu02040](http://www.genome.jp/kegg-bin/show_pathway?bsu02040)).

**ANI and shared-genome analyses.** OrthoMCL orthologous groups were used to determine shared gene contents in pairwise genome comparisons. For a genome pair (genome 1 and genome 2), the total number of genes in genome 1 was determined and the number of genes in genome 1 shared with genome 2 (based on shared ortholog group membership) was determined. Percent shared gene content was calculated by dividing the number of genome 1 genes shared with genome 2 by the number of genes in genome 1. Nucleotide alignments of shared genes were used to determine the percentage of identical and different nucleotide residues in shared genes. For comparisons within species, at least 2,113 gene sequences were utilized. Percent ANI was calculated by dividing the number of identical nucleotide residues in shared genes by the total number of nucleotide residues.

**Recombination analysis.** See the Text S1 in the supplemental material for a description of the methods used for genome mosaicism analysis and plot generation.

**Biolig analysis.** A subset of strains (8/18 _E. faecalis_, 6/8 _E. faecium_, 3/3 _E. casseliflavus_, and 1/1 _E. gallinarum_) representing the diversity of the collection were analyzed in duplicate by Biolog Phenotype microarray in accordance with the manufacturer’s instructions. Optical density at 590 nm (OD590) was read using a synergy microplate reader (Bio-Tek). The 48-h OD590 reading of each well containing a carbon source was divided by the OD590 value obtained for the negative-control well. A ratio which gave a reproducible value of 2× the background was considered to be a positive result.

**ACKNOWLEDGMENTS**

This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH), Department of Health and Human Services, under grant AI072360, the Harvard-Wide Antibiotic Resistance Program (AI083214), and contract no. HHSN272200900018C. Additionally, K.L.P. was supported by NIH fellowships EY007145 and EY020734. We gratefully thank Rob Willems, Marvin Whiteley, Nathan Shankar, Willem van Schaik, and Mark Huycke for helpful comments during manuscript preparation.

**SUPPLEMENTAL MATERIAL**

Supplemental material for this article may be found at [http://mbio.asm.org](http://mbio.asm.org)/lookup/suppl/doi:10.1128/mBio.00318-12/-/DCSupplemental.

Text S1, DOCX file, 0.1 MB.

Data set S1, XLSX file, 0.1 MB.

Data set S2, XLSX file, 0.1 MB.

Data set S3, XLSX file, 0.1 MB.

Data set S4, XLSX file, 0.1 MB.

Figure S1, PDF file, 0.1 MB.

Figure S2, PDF file, 0.1 MB.

Figure S3, PDF file, 0.1 MB.

Table S1, DOC file, 0.1 MB.

Table S2, DOC file, 0.1 MB.

Table S3, DOC file, 0.1 MB.

Figure S4, PDF file, 0.1 MB.

Table S4, DOC file, 0.1 MB.
REFERENCES


38. Kadioglu A, Weiser JN, Paton JC, Andrew PW. 2008. The role of


