Identification and phenotypic characterization of a second collagen adhesin, Scm, and genome-based identification and analysis of 13 other predicted MSCRAMMs, including four distinct pilus loci, in Enterococcus faecium

Jouko Sillanpää,1,2† Sreedhar R. Nallapareddy,1,2† Vittal P. Prakash,1,2 Xiang Qin,3 Magnus Höök,4 George M. Weinstock3 and Barbara E. Murray1,2,5

1Division of Infectious Diseases, Department of Internal Medicine, University of Texas Medical School, Houston, TX, USA
2Center for the Study of Emerging and Re-emerging Pathogens, University of Texas Medical School, Houston, TX, USA
3Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA
4Center for Extracellular Matrix Biology, Institute of Biosciences and Technology, Texas A&M University Health Science Center, Houston, TX, USA
5Department of Microbiology and Molecular Genetics, University of Texas Medical School, Houston, TX, USA

Attention has recently been drawn to Enterococcus faecium because of an increasing number of nosocomial infections caused by this species and its resistance to multiple antibacterial agents. However, relatively little is known about the pathogenic determinants of this organism. We have previously identified a cell-wall-anchored collagen adhesin, Acm, produced by some isolates of E. faecium, and a secreted antigen, SagA, exhibiting broad-spectrum binding to extracellular matrix proteins. Here, we analysed the draft genome of strain TX0016 for potential microbial surface components recognizing adhesive matrix molecules (MSCRAMMs). Genome-based bioinformatics identified 22 predicted cell-wall-anchored E. faecium surface proteins (Fms), of which 15 (including Acm) had characteristics typical of MSCRAMMs, including predicted folding into a modular architecture with multiple immunoglobulin-like domains. Functional characterization of one [Fms10; redesignated second collagen adhesin of E. faecium (Scm)] revealed that recombinant Scm 65 (A- and B-domains) and Scm 36 (A-domain) bound to collagen type V efficiently in a concentration-dependent manner, bound considerably less to collagen type I and fibrinogen, and differed from Acm in their binding specificities to collagen types IV and V. Results from far-UV circular dichroism measurements of recombinant Scm 36 and of Acm 37 indicated that these proteins were rich in β-sheets, supporting our folding predictions. Whole-cell ELISA and FACS analyses unambiguously demonstrated surface expression of Scm in most E. faecium isolates. Strikingly, 11 of the 15 predicted MSCRAMMs clustered in four loci, each with a class C sortase gene; nine of these showed similarity to Enterococcus faecalis Ebp pilus subunits and also contained motifs essential for pilus assembly. Antibodies against one of the predicted major pilus proteins, Fms9 (redesignated EbpCfm), detected a ‘ladder’ pattern of high-molecular-mass protein bands in a Western blot analysis of cell surface extracts from E. faecium, suggesting that EbpCfm is polymerized into a pilus structure. Further analysis of the transcripts of the corresponding gene cluster indicated that fms1 (ebpAfm), fms5 (ebpBfm) and ebpCfm are co-transcribed, a result consistent with those for pilus-encoding gene clusters of other Gram-positive bacteria. All 15 genes occurred frequently in 30 clinically derived diverse E. faecium isolates tested. The common occurrence of MSCRAMM- and pilus-encoding genes and the presence of a second collagen-binding protein may have important implications for our understanding of this emerging pathogen.
INTRODUCTION

Enterococcus faecium, a member of the normal commensal flora, has recently emerged as a prominent nosocomial pathogen that causes serious infections, including infective endocarditis (Murray, 2000). Nosocomial infections due to E. faecium can be a life-threatening challenge to physicians because of this organism’s multi-drug resistance (Murray, 2000; Rice, 2001). Furthermore, in addition to the selective advantage conferred by antibiotics, strains that emerge in the hospital setting more often carry putative virulence genes such as esp, encoding enterococcal surface protein (Rice et al., 2003; Willems et al., 2001), hyl, encoding a putative hyaluronidase (Rice et al., 2003), and a functional copy of the acm gene, encoding a collagen adhesin (Nallapareddy et al., 2003, 2008a, b).

It is known that pathogenic bacteria have evolved a plethora of proteins to adhere to and invade host tissues and to resist host defences (Pizarro-Cerdà & Cossart, 2006). Among these is a family of surface proteins with well-established roles in host–pathogen adherence which have been termed microbial surface components recognizing adhesive matrix molecules (MSCRAMMs; Patti et al., 1994). MSCRAMMs share several characteristics, including (i) an N-terminal signal peptide, (ii) a non-repeated A-domain consisting of immunoglobulin (Ig)-like fold(s), (iii) a B-domain with a variable number of repeats among different strains, and (iv) a C-terminal cell wall anchor (CWA) domain. Some of these MSCRAMMs have been shown recently to be tethered to each other by a designated sortase to make up multimeric cell surface structures, named pili (Mora et al., 2005; Nallapareddy et al., 2006). Sortases, encoded by the srtA to srtD classes of genes (Dramsi et al., 2005), were originally described as membrane-bound transpeptidases that cleave the LPXTG-like motif in the CWA domain and covalently link CWA proteins to the peptidoglycan (Marraffini et al., 2006). While class A sortases appear to be ubiquitous and involved in cell surface anchoring of a large number of LPXTG-containing proteins (Marraffini et al., 2006), the class C (subfamily 3) sortase enzymes, which have a more limited substrate specificity, have recently been shown to be involved in pilus biogenesis, in addition to their role in surface anchoring (Kemp et al., 2007; Scott & Zahner, 2006; Telford et al., 2006).

Studies that have characterized the binding interactions of staphylococcal and enterococcal MSCRAMMs have identified that the ligand-binding A-domains consist of two to three subdomains (N1-N2/N3), each adopting an Ig-like fold (Liu et al., 2007; Nallapareddy et al., 2007; Ponnuraj et al., 2003; Zong et al., 2005). Based on the crystal structures of prototype MSCRAMMs, two slightly different models have been proposed to explain their binding mechanisms to linear peptides of fibrinogen and to triple-helical collagen. In the ‘dock, lock and latch’ binding model, a fibrinogen chain is inserted into a cleft between two Ig-folded subdomains and is then secured by a C-terminal N3 extension (latch) that is reoriented upon ligand binding and complements a β-sheet on the N2 subdomain (Ponnuraj et al., 2003). A variation of this two-subdomain binding model, ‘the collagen hug’, has been proposed for the collagen-binding MSCRAMMs (Liu et al., 2007; Zong et al., 2005).

Previous in silico analyses have identified a family of genes that encode MSCRAMM-like proteins in the genomes of several Gram-positive bacteria, including our reports of the ebp (endocarditis and biofilm-associated pilus of Enterococcus faecalis) operon, Ace (adhesin of collagen from E. faecalis) and Acm (adhesin of collagen from E. faecium) (Bowden et al., 2005; Nallapareddy et al., 2000, 2003, 2006; Roche et al., 2003; Sillanpää et al., 2004; Xu et al., 2004). Recently, Hendrickx et al. (2007) predicted 22 CWA proteins from the E. faecium TX0016 [formerly DO (Arduino et al., 1994)] genome, of which five were found to be enriched in isolates of the hospital-adapted clonal complex 17 (CC17). However, with the exception of the prototype MSCRAMM Acm (Nallapareddy et al., 2003), there has been no demonstration of other MSCRAMMs in E. faecium. Although our previous report identified an essential and secreted broad-spectrum adhesin, SagA, that exhibits binding to fibrinogen, collagen type I (CI), collagen type IV (CIV), fibronectin and laminin, this protein lacks a CWA domain and other MSCRAMM characteristics (Teng et al., 2003).

In the present study, we identified 14 genes (in addition to acm) that encode predicted MSCRAMMs in the genome of E. faecium TX0016. Recombinant forms of one of these proteins, designated Scm (for second collagen adhesin of E. faecium), were characterized to confirm the structural predictions and to identify its ligand. Cell-surface expression of Scm by selected E. faecium strains was quantified using FACS, and antibodies raised against a recombinant form of one of the major pilus proteins showed a high-molecular mass (HMM) ladder pattern characteristic of Gram-positive pili. Co-transcription of one of the pilus-encoding gene clusters was demonstrated by Northern

†These authors contributed equally to this work.

Abbreviations: CI, collagen type I; CIV, collagen type IV; CV, collagen type V; CD, circular dichroism; CWA, cell wall anchor; ECM, extracellular matrix; HMM, high molecular mass; Ig, immunoglobulin; MSCRAMM, microbial surface components recognizing adhesive matrix molecule.

Supplementary methods, describing bacterial isolates, identification and complements a second pilus proteins; and a supplementary table, listing oligonucleotides used in this study, are available with the online version of this paper.
hybridization and RT-PCR. In addition, we determined the distribution of the MSCRAMP-encoding genes among 30 diverse E. faecium clinical isolates.

**METHODS**

**Strains, plasmids and cultivation of bacteria.** Relevant characteristics of bacterial strains, growth conditions and plasmids used in this study are summarized in Table 1 and in the Supplementary Methods. All constructs were given TX numbers and plasmids from these constructs were assigned corresponding pTEX numbers (Table 1).

*Escherichia coli* and *E. faecium* cells were grown in Luria–Bertani (LB) broth/agar and Brain Heart Infusion (BHI) broth/agar (Difco), respectively. For *Escherichia coli* constructs, antibiotics were used at the following concentrations: 25 μg kanamycin mL⁻¹ and 100 μg ampicillin mL⁻¹.

**Identification and structural analysis of CWA proteins.** Bioinformatics methods used for genome-wide identification of CWA proteins are described in the Supplementary Methods. Domain architecture as well as fold-recognition analyses were carried out by comparing the protein sequences against domain databases, as described in Supplementary Methods.

**Construction of expression plasmids.** Genomic DNA from *E. faecium* strains was isolated as described earlier (Wilson, 1994). DNA regions encoding amino acids 27–624 (A-domain and B-repeats) and 27–333 (A-domain) of Scm, and amino acids 33–590 of Ebp_Cfm [Fms9 was renamed Ebp_Cfm based on its 74 % amino acid identity (84 % similarity) with EbpC of *E. faecalis* over the entire protein (mature protein without the signal peptide and the CWA domain)], were amplified using primers listed in Supplementary Table S1 and cloned into the expression vector pQE30 (Qiagen) to obtain constructs of rScm36, rScm65 and rEbp_Cfm62. Constructs were confirmed by sequencing.

**Purification of recombinant proteins.** Recombinant proteins with N-terminal His₆-tags were expressed and purified using nickel affinity chromatography followed by anion-exchange chromatography (Nallapareddy et al., 2007; Sillanpää et al., 2004). Protein concentrations were determined by absorption spectroscopy at 280 nm using calculated molar absorption coefficient values. Molecular masses were determined with MALDI-TOF MS for rScm36 and rScm65 based on their calculated molecular masses. Similarly, the cloned segment of Ebp_Cfm was designated rEbp_Cfm62. Constructs were confirmed by sequencing.

**Northern hybridization.** Using the RNAProtect bacteria reagent and RNeasy Mini kit (Qiagen), total RNA was isolated from TX0082 grown in BHI to mid-exponential phase (starting inoculum, OD₆₀₀=0.01) with mutanolysin as described previously (Nallapareddy et al., 2006). Equal amounts of concentrated mutanolysin extracts were separated using 4–15 % gradient SDS-PAGE gels (Bio-Rad) under reducing conditions, and transferred to PVDF membranes according to the manufacturer’s protocol. Membranes were probed with affinity-purified anti-Scm₆5 and anti-Ebp_Cfm₆2 antibodies (see above) followed by HRP-conjugated anti-goat IgG antibodies, and as a control, total IgG antibodies purified from preimmune goat sera were used. Signal was detected using SuperSignal West Pico chemiluminescent detection reagents (Thermo Scientific).

**RT-PCR.** Total RNA (isolated as above for Northern hybridization) was treated twice with 20 U RQ1 DNase (Promega) for 30 min at 37 °C. DNase was removed using the RNeasy Mini kit and purification protocol (Qiagen). Total RNA was then reverse-transcribed with specific primers (Supplementary Table S1) using the SuperScript One-Step RT-PCR with Platinum Taq kit (Invitrogen) according to the manufacturer’s instructions. DNA sequencing verified that the primer regions of TX0082 were 100 % identical to the corresponding sequences in TX0016. As an internal control, a 509 bp fragment of *gyrA* (encoding gyrase A) was amplified using the FmGyFF and FmGyFR primers (Supplementary Table S1). Reactions without reverse transcriptase were used as controls to verify lack of DNA contamination in the total RNA preparations.
Table 1. Bacterial strains and plasmids used in this study

<table>
<thead>
<tr>
<th>Strain or plasmid</th>
<th>Relevant characteristics*; origin; year of isolation</th>
<th>Reference or source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E. faecium strains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TX0016 (DO)</td>
<td>Endocarditis isolate, VanS, AmpR; Houston, TX; 1992</td>
<td>Arduino et al. (1994)</td>
</tr>
<tr>
<td>TX2535</td>
<td>Endocarditis isolate, VanS, AmpR; Houston, TX; 1995</td>
<td>Nallapareddy et al. (2003)</td>
</tr>
<tr>
<td>TX0068</td>
<td>Endocarditis isolate, VanS, AmpR; Worcester, MA; 1994</td>
<td>This study</td>
</tr>
<tr>
<td>TX0074</td>
<td>Endocarditis isolate, VanS, AmpR; Valhalla, NY; 1995</td>
<td>This study</td>
</tr>
<tr>
<td><strong>Escherichia coli strains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M15(pREP4)</td>
<td>Escherichia coli strain for expression of recombinant proteins</td>
<td>Qiagen</td>
</tr>
<tr>
<td>TX5432</td>
<td>M15 (pTEX5432), AmpR, KanR</td>
<td>This study</td>
</tr>
<tr>
<td>TX5628</td>
<td>M15 (pTEX5628), AmpR, KanR</td>
<td>This study</td>
</tr>
<tr>
<td><strong>Plasmids</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pQE30</td>
<td>Expression vector</td>
<td>Qiagen</td>
</tr>
<tr>
<td>pTEX5432</td>
<td>1794 bp fragment from TX0016 <em>scm</em> (encoding complete A- and B-domains) cloned into pQE30 expression vector</td>
<td>This study</td>
</tr>
<tr>
<td>pTEX5628</td>
<td>921 bp fragment from TX2535 <em>scm</em> (encoding complete A-domain) cloned into pQE30 expression vector†</td>
<td>This study</td>
</tr>
<tr>
<td>pTEX5630</td>
<td>1674 bp fragment from TX0016 ebpCfm (encoding mature EbpCfm without signal peptide or CWA domain) cloned into pQE30 expression vector</td>
<td>This study</td>
</tr>
</tbody>
</table>

* AmpR, ampicillin resistant; KanR, kanamycin resistant; VanR, vancomycin resistant.
† The DNA sequence of *scm* cloned into pTEX5628 from strain TX2535 is 100% identical with that of TX0016.

**RESULTS AND DISCUSSION**

**Identification of putative CWA proteins with Ig-like folds**

Our search of the nearly completed genome sequence of *E. faecium* endocarditis-derived strain TX0016 (G. M. Weinstock & B. E. Murray, unpublished data; http://www.hgsc.bcm.tmc.edu/) yielded a total of 22 ORFs encoding putative CWA proteins (designated Fms, for *E. faecium* surface proteins) with a tripartite pattern near the C terminus (an LPXTG motif or variant, followed by a membrane-spanning hydrophobic region and a positively charged tail). This number is in the range of predicted CWA proteins of related Gram-positive bacteria (Ponnuraj et al., 2003), which vary from eight (*Streptococcus mutans*) to 41 (*E. faecalis*). While the number of CWA proteins identified here is the same as the number reported in a recent study (Hendrickx et al., 2007) which identified CWA proteins from a partial TX0016 genome sequence (available at NCBI; accession no. NZ_AAAK00000000), the published report did not include one of the ORFs reported here (Fms6) and classified a pseudogene (Fms15) as two ORFs (Supplementary Results).

Among these 22 putative CWA proteins, 18 contained a predicted N-terminal signal peptide sequence required for Sec-dependent secretion. Further analysis of the 5' regions of the remaining four ORFs (Fms14, Fms15, Fms16 and Fms19) that lack a signal peptide revealed the presence of N-terminal signal peptides in the ORFs immediately upstream of Fms15, Fms16 and Fms19. Careful examination of the junction regions showed the presence of a premature stop codon due to a point mutation or frameshift in *fms15*, *fms16* and *fms19* (Supplementary Results). Sequencing of *fms16* and *fms19* regions from additional isolates identified intact *fms16* and *fms19* genes in two of 10 *E. faecium* clinical isolates tested (Supplementary Results). However, we did not find an *E. faecium* strain with an intact *fms15* gene (Supplementary Results).

Subsequently, using the fold recognition servers 3D-PSSM and PHYRE, we identified 15 of the predicted Fms proteins (Table 2) as containing one or more Ig-like folds enriched with β-sheets, and only a small quantity of α-helices; similar secondary structure compositions have been found in the ligand-binding A-regions of MSCRAMMs (Deivanayagam et al., 2002). One of these, Fms8, had been previously identified in our laboratory as Acm (Nallapareddy et al., 2003). The percentage of *E. faecium* LPXTG proteins containing Ig-like folds is greater than the percentages for other Gram-positive bacteria (22–45%) (Ponnuraj et al., 2003). The Ig-like fold-containing modules of the A-domain of a number of well-characterized *Staphylococcus aureus* MSCRAMMs (e.g. the fibronectin-, fibrinogen- and elastin-binding protein FnbpA, the fibrinogen-binding proteins ClfA and ClfB, and the collagen-binding protein Cna) and the prototype MSCRAMM of *E. faecium*, Acm, have been shown to play...
a direct role in interactions with their respective ECM protein ligands (Nallapareddy et al., 2007; Ponnuraj et al., 2003; Zong et al., 2005).

**Structural organization of the *E. faecium* putative MSRCAMMs**

Because of our interest in surface adhesins, we next performed additional analyses on the subset of predicted Ig-like fold-containing proteins and found that most were predicted to have a multi-domain architecture similar to the characterized MSRCAMMs of *S. aureus* and *E. faecalis* (Deivanayagam et al., 2002; Perkins et al., 2001; Sillanpää et al., 2004) (Supplementary Results). A schematic of the predicted primary sequence organization of these proteins is depicted in Fig. 1. The N-terminal signal peptide sequences, found in all but Fms14, were 24–56 aa in length. The majority had perfectly conserved canonical LPXTG motifs, while six had anchor motifs with a broader consensus (Table 2).

**Individually organized genes encoding putative MSRCAMMs**

Fms10, Fms15 and Fms18, like Acm, are encoded by individual genes and are not clustered with other MSRCAMM-encoding genes. Among these, Fms18 has 94% similarity (Table 2) to the *E. faecalis* MSRCAMM EF1896, which is predicted to be mostly Ig-folded and shows binding to a serum protein (J. Sillanpää, S. R. Nallapreddy, B. E. Murray and M. Hook, unpublished data). Scm (Fms10) of TX0016 has 13 B-repeats of 19 aa in length and a partial repeat of 10 aa, while pseudogene fms15 is predicted to encode eight B-repeats of 91–93 aa in length. The primary sequences of Scm and Fms15 each predict a single subdomain with an Ig-like fold in the A-domain, unlike the well-established MSRCAMMs, which have two to three subdomains. Hence, we hypothesize that if these proteins are adhesins, they may adopt a different binding mechanism, and we chose Scm for further characterization.

**Characteristics of recombinant Scm protein segments**

Total yields of the recombinant protein rScm65 (fully mature protein consisting of the A-domain and B-repeats; amino acids 27–624) were low. Although we estimated that this protein was over 95% pure immediately after eluting from the chromatographic columns, as shown in Supplementary Fig. S1, it was susceptible to partial degradation even after brief storage. rScm65 migrated as a larger band than expected from the calculated molecular mass of 64.5 kDa, likely due to the highly acidic nature (pI 3.9) of the protein. To obtain a stable construct with higher yields, we next expressed rScm36 encompassing the complete A-domain (amino acids 27–333; Supplementary Fig. S1). A MALDI-TOF MS analysis determined the molecular mass of rScm36 to be 35711 Da, which is in good agreement with the calculated mass of 35691 Da, and confirmed the absence of post-translational modifications and the homogeneity of the purified protein. As an additional confirmation of the identities of the purified protein segments and evidence of their intact N termini, a monoclonal anti-His6 antibody detected the N-terminal His6 fusion in both purified protein segments in a Western blot assay (data not shown).

**Recombinant Scm binds to collagen in a dose-dependent manner**

Since the adhesive functions of previously studied MSRCAMM proteins have mostly been ascribed to binding to protein components of the host ECM, we examined the potential ECM-adhesive functions of Scm by testing its binding to a panel of individual ECM proteins in an ELISA-type solid-phase ligand-binding assay. As seen in Fig. 2(a), both rScm65 (containing the full mature protein) and Scm36 (containing only the A-domain) exhibited significant binding to CV (35-fold and 82-fold, respectively, higher than the background level seen with immobilized BSA). Interestingly, stronger binding of these rScm proteins (especially rScm36) to fibrinogen and CI was also observed compared with their binding to CIV, laminin, and fibronectin (Fig. 2a).

Further analysis of the binding interactions of increasing concentrations of rScm36 with CI and CV, as well as fibrinogen, by dose–response assays showed concentration-dependent binding of rScm36 to CV that approached saturation, with apparent half-maximal binding (K_D) reached at 30.8 ± 1.8 μM (Fig. 2b). The apparent affinity of rScm36 for fibrinogen and CI was low and did not reach reliable saturation levels. In spite of partial degradation, Scm65 showed dose-dependent binding, with an affinity for CV similar to that for Scm36 (data not shown). Although K_D values from steady-state assays are estimates, our affinity determination for the Scm A-domain is in the same range as that shown previously for the minimal ligand-binding domain of Ace to CI (48 ± 7 μM) (Rich et al., 1999), but lower than that for the A-domain of Acm to CI (3.8 μM) (Nallapareddy et al., 2003). Interestingly, Acm did not show appreciable binding to CV (data not shown) and bound efficiently to CI, and to a lesser degree to CIV, in agreement with our previous report (Nallapareddy et al., 2003). Taken together, our results suggest that CV is a binding ligand for Scm and that this binding is mediated by the A-domain.

The three collagen types included in our assays have distinct structures and tissue distributions. CI is the most abundant form and the main component of the collagen fibres that are widely distributed in human tissues, while CIV forms structurally different cross-linked networks and is found nearly exclusively in basement membranes. CV, although quantitatively less abundant, has a critical role in formation of the fibrillar collagen matrix and connecting interstitial
Table 2. Summary of features of the predicted MSCRAMM proteins of *E. faecium* and the distribution of the corresponding genes in 30 *E. faecium* isolates associated with clinical infections

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Accession number</th>
<th>Protein length (aa)</th>
<th>Cell wall anchoring motif</th>
<th>Best match to other CWA proteins (similarity, %)</th>
<th>Number of isolates positive (%) (n=30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>acm (fms8)</td>
<td>ZP_00064683</td>
<td>721</td>
<td>LPKTS</td>
<td><em>Staph. aureus</em> Cna (62 %)*</td>
<td>30 (100 %)</td>
</tr>
<tr>
<td>scm (fms10)</td>
<td>ZP_00064835</td>
<td>660</td>
<td>LPKGT</td>
<td><em>Lactobacillus brevis</em> hypothetical protein (54 %)</td>
<td>30 (100 %)</td>
</tr>
<tr>
<td>fms15†</td>
<td>ZP_00062631</td>
<td>1412</td>
<td>LPKGT</td>
<td><em>Bacillus anthracis</em> collagen-binding protein BA0871 (49 %)</td>
<td>29 (97 %)</td>
</tr>
<tr>
<td>fms18</td>
<td>ZP_00063098</td>
<td>1075</td>
<td>FPQGT</td>
<td><em>E. faecalis</em> EF1896 (91 %)</td>
<td>24 (80 %)</td>
</tr>
<tr>
<td>Cluster 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ebpAfm (fms1)</td>
<td>ZP_0006289</td>
<td>1129</td>
<td>LPETG</td>
<td><em>E. faecalis</em> EbpA (75 %)</td>
<td>29 (97 %)</td>
</tr>
<tr>
<td>ebpBfm (fms5)</td>
<td>ZP_00062688</td>
<td>473</td>
<td>LPKTN</td>
<td><em>E. faecalis</em> EbpB (65 %)</td>
<td>29 (97 %)</td>
</tr>
<tr>
<td>ebpCfm (fms9)</td>
<td>ZP_00062687</td>
<td>625</td>
<td>LPSTG</td>
<td><em>E. faecalis</em> EbpC (84 %)</td>
<td>29 (97 %)</td>
</tr>
<tr>
<td>Cluster 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fms11</td>
<td>ZP_00064460</td>
<td>773</td>
<td>LPSTG</td>
<td><em>Lactobacillus johnsonii</em> hypothetical protein (46 %)</td>
<td>29 (97 %)</td>
</tr>
<tr>
<td>fms19‡</td>
<td>ZP_00064462</td>
<td>499</td>
<td>FPKTS</td>
<td><em>E. faecalis</em> EbpB (54 %)</td>
<td>29 (97 %)</td>
</tr>
<tr>
<td>fms16‡</td>
<td>ZP_00064464</td>
<td>470</td>
<td>LPSTG</td>
<td><em>E. faecalis</em> EbpC (46 %)</td>
<td>29 (97 %)</td>
</tr>
<tr>
<td>Cluster 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fms14§</td>
<td>ZP_00063043</td>
<td>1313</td>
<td>LPATG</td>
<td><em>E. faecalis</em> EbpA (41 %)</td>
<td>30 (100 %)</td>
</tr>
<tr>
<td>fms17</td>
<td>ZP_00063042</td>
<td>437</td>
<td>FPQTN</td>
<td><em>E. faecalis</em> EbpB (47 %)</td>
<td>30 (100 %)</td>
</tr>
<tr>
<td>fms13</td>
<td>ZP_00063041</td>
<td>565</td>
<td>LPATG</td>
<td><em>E. faecalis</em> EbpC (25 %)</td>
<td>30 (100 %)</td>
</tr>
<tr>
<td>Cluster 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fms21</td>
<td>ZP_00063529</td>
<td>658</td>
<td>VPMTG</td>
<td><em>E. faecalis</em> EbpC (41 %)</td>
<td>28 (93 %)</td>
</tr>
<tr>
<td>fms20</td>
<td>ZP_00063526</td>
<td>696</td>
<td>VPKTG</td>
<td><em>E. faecalis</em> EF2525 (41 %)</td>
<td>22 (73 %)</td>
</tr>
</tbody>
</table>

*Identified in our previous studies (Nallapareddy et al., 2003).
†In silico analysis predicts a one-nucleotide deletion in the B-repeat region.
§These genes had a nucleotide substitution or deletion in *E. faecium* strain TX0016 versus TX0068 and TX0074, resulting in pseudogenes.
§Fms14 lacks an identifiable signal peptide, but has the other characteristics of a CWA protein.

collagen fibrils with membranous collagen networks (Nicholls et al., 1996; Wenstrup et al., 2004). Considering the diverse tissue localizations and structural differences of the various collagen types, the possession of two collagen adhesins with different binding specificities to various collagen types could give *E. faecium* the ability to fine-tune its adherence phenotype to suit a given tissue. Since collagens, including CV, are also present in the intestinal submucosa (Liang et al., 2006), Scm could alternatively be involved in colonization and persistence in the intestinal tract, a major natural reservoir of both commensal and infection-associated strains of *E. faecium* in humans, or in facilitating translocation through damaged intestinal epithelium.

**Structural analysis of recombinant Scm protein**

To investigate the structural predictions made above, we analysed the full A-domain of Scm with far-UV CD spectroscopy. The spectra obtained showed a maximum at 197 nm, another at 191 nm, and a minimum at 217 nm (Fig. 3a). A similar overall pattern was observed with the collagen-binding A-domain of the *E. faecium* prototype MSCRAMM Acm, which we have characterized previously (Nallapareddy et al., 2003, 2007), and also with the minimal collagen-binding region of Ace from *E. faecalis*, which crystal structure studies have recently revealed folds into a similar DEv variant of the Ig-fold as that found previously in the ligand-binding A-domains of staphylococcal MSCRAMMs (Liu et al., 2007). Deconvolution of the collected data showed a secondary structure composition of 0.15 ± 0.03 α-helix and 0.34 ± 0.01 β-sheet for rScm36 (Fig. 3b). Although the helical content of this protein is slightly higher than in the two control proteins, these results generally resemble the overall secondary structure compositions of the ligand-binding regions of Acm and Ace, and are in good agreement with earlier CD analyses and crystal structure data of other MSCRAMM A-regions (Rich et al., 1999). While fold analyses and multiple alignments predicted a single Ig-folded subunit from amino acids 125–325 in the A-domain of Scm, our CD measurements indicate a high β-sheet composition for the entire Scm A-domain, thus suggesting that an Ig-folded or similar β-sheet-rich structure may extend over the whole A-domain (amino acids 27–333).

**Cell surface expression of Scm**

Using Scm-specific antibodies affinity-purified (with rScm36) from goat antisera, we assessed the surface expression of Scm by eight *E. faecium* clinical isolates (including the sequenced strain TX0016) using whole-cell
ELISA. Two *E. faecium* community-derived strains that lacked *scm* (identified below) were used as negative controls. Among eight *scm*+ isolates, four (TX0074, TX2535, TX2400 and TX2589) were strongly positive, two (TX0068 and TX2081) were positive, one (TX2416) was weakly positive and one (TX0016) was negative in this assay (see Supplementary Fig. S2), suggesting that Scm is produced on the cell surface of many *E. faecium* isolates during *in vitro* growth. Subsequent quantification of Scm surface expression in these isolates by FACS (Fig. 4) showed variable levels of surface expression, with the percentage of cells expressing Scm ranging from 2.3 to 96.9%. In strains TX0074, TX2535, TX2400 and TX2589, Scm was detected on the surface of 90–97% of cells and with relatively high fluorescence intensities, thus confirming our observations with whole-cell ELISA. Furthermore, these surface detection studies suggested that the majority of cells of *scm*+ isolates were actively producing Scm on the cell surface during *in vitro* growth. Consistent with our whole-cell ELISA and FACS results, a Western blot assay using the anti-Scm36 antibody identified a protein band corresponding to the expected size of mature Scm (~63 kDa) in mutanolysin cell wall extracts from TX0074 and TX2535, but not from TX0016 (data not shown).

**Eleven of 15 putative MSCRAMM-encoding genes are clustered at four different genomic loci, and nine are predicted to produce four distinct types of pili**

Unlike Acm, Scm (Fms10), Fms15 and Fms18, the remaining 11 putative MSCRAMM genes were found to be clustered in four genomic loci. Nine of the clustered genes showed significant similarity to pilus-associated proteins from other species (Table 2). Analysis of the nucleotide sequence of the locus spanning *fms1*, *fms5* and *ebpC* (em19) predicted the presence of four ORFs, all oriented in the same direction, including an *srtC1* gene encoding a class C (subfamily 3) sortase (Fig. 1b). We subsequently renamed *fms1*, *fms5* and *srtC1* as *EbpAfm*, *EbpBfm* and *Bpsfm*, respectively, based on their high sequence identity with the *E. faecalis* Ebp cluster proteins (Nallapareddy et al., 2006) (Table 2). A similar arrangement of predicted pilus-associated proteins with an adjacent class C sortase was found in the *fms14-17-13* and the *fms11-19-16* clusters. Besides our prediction of *ebpABCfm*, *fms14-17-13* and *fms11-19-16* co-transcription, due to short or overlapping intergenic regions, we were unable to identify transcriptional terminator-like sequences in these gene clusters; thus, it is likely that these genes are in three operons (see also below). The genes encoding Fms20 and Fms21 were found to be located close to each other in the genome, separated by two ORFs. One of these ORFs encodes a predicted class C sortase (*SrtC4*). In addition to *srtC4*, this locus also contains one class A sortase-encoding gene (*srtA*) immediately upstream of *fms21* (Fig. 1b). Another ORF located between SrtC4 and Fms20 shows 27% similarity to the EbpB pilus protein and has a signal peptide sequence but no CWA motif. Of note, the *Ebp cluster* (*ebpABCfm*) contains a sixth sortase gene (*srtC5*), predicted to encode a class C sortase (*SrtC5*). In addition to *srtC4*, this locus also contains one class A sortase-encoding gene (*srtA*) immediately upstream of *fms21* (Fig. 1b). Another ORF located between SrtC4 and Fms20 shows 27% similarity to the EbpB pilus protein and has a signal peptide sequence but no CWA motif. Of note, the *Ebp cluster* (*ebpABCfm*) contains a sixth sortase gene (*srtC5*), predicted to encode a class C sortase (*SrtC5*).
polymerization of the *Corynebacterium diphtheriae* pilus (Ton-That & Schneewind, 2003; Ton-That et al., 2004) were found to be 100% conserved. Two accessory proteins, namely EbpAfm and Fms14, have a von Willebrand factor type A-domain with a MIDAS motif which is frequently found in pili-associated accessory proteins. Recently, crystal structure analyses of the minor pilin protein GBS52 of *Streptococcus agalactiae* and the major pilin Spy0128 of *Streptococcus pneumoniae* have demonstrated that these proteins contain two Ig-like domains (Kang et al., 2007;
Krishnan et al., 2007), consistent with our prediction of Ig-like folds in *E. faecium* pilus proteins. These features, together with the presence of an independent class C sortase (class C sortases are used in the assembly of pili of related Gram-positive bacteria) in each of the four clusters, indicate that *E. faecium* may harbour genes for multiple pilus-like structures. Recent studies on the pili of *C. diphtheriae*, *E. faecalis*, group A and B streptococci, and pneumococci have demonstrated their role in bacterial adherence and biofilm formation, and their contribution to bacterial pathogenesis and modulation of the host immune system (Barocchi et al., 2006; Dramsi et al., 2006; Mandlik et al., 2007; Nallapareddy et al., 2006; Singh et al., 2007; Telford et al., 2006).

**EbpCfm is present as an HMM polymeric protein complex at the cell surface of *E. faecium***

To confirm the predicted location of EbpCfm on the surface of *E. faecium* as part of polymeric pili, we probed mutanolysin cell wall extracts of endocarditis-derived *E. faecium* strains TX0082 and TX0016 with affinity-purified anti-rEbpCfm62 antibodies. A ladder-like pattern of HMM bands with sizes greater than 200 kDa was detected from strain TX0082 (Fig. 5). In contrast, only a single band has been observed with Acm in mutanolysin extracts of TX0082 (Nallapareddy et al., 2003). The HMM banding pattern is consistent with observations of a large number of pilus proteins from other Gram-positive bacteria and suggests that EbpCfm is assembled similarly into multimeric pilus structures on the *E. faecium* cell surface. In contrast to TX0082, no signal was detected from the cell wall extract of TX0016, which is similar to our whole-cell ELISA results (data not shown) and indicates that the EbpCfm-containing pilus is not expressed in this strain under the growth conditions used.

**The ebpABCfm genes are transcribed as a single polycistronic mRNA***

As stated above, bioinformatic analyses indicated that each of the four pilus-encoding gene clusters is likely to be
organized as an operon. For the ebpA<sub>fm</sub> to bps<sub>fm</sub> cluster, a 12 bp inverted repeat that could form a stem–loop structure [free energy of −22.55 kcal mol<sup>−1</sup> (94.34 kJ mol<sup>−1</sup>)] located in the intergenic region between ebpC<sub>fm</sub> and bps<sub>fm</sub> is likely to function as a ρ-independent transcriptional terminator, thus predicting an ebpA<sub>fm</sub> to ebpC<sub>fm</sub> transcript and an independent bps<sub>fm</sub> transcript. To validate this, we carried out transcriptional analyses of the ebpA<sub>fm</sub> to bps<sub>fm</sub> locus using Northern hybridization. The ebpA<sub>fm</sub> ebpB<sub>fm</sub> and ebpC<sub>fm</sub> probes all hybridized to a single RNA band of ~7 kb (Fig. 6a) which corresponds to the expected size of a polycistronic ebpABC<sub>fm</sub> mRNA transcript, demonstrating that these three genes are cotranscribed. Probing with bps<sub>fm</sub> detected a very low intensity band (~1 kb), consistent with the expected size of a monocistronic bps<sub>fm</sub> mRNA transcript, suggesting that this transcript either is produced at low levels at this time point or has a relatively short half-life. RT-PCR using three independent sets of internal bps<sub>fm</sub> primer pairs amplified fragments of the expected sizes (lane 4 of Fig. 6b) (data not shown for two additional primer pairs), thus confirming the expression of bps<sub>fm</sub>. Another faint band of ~8 kb was also observed with the bps<sub>fm</sub>-probed Northern blots; while its identity is currently unclear, it might represent a related mRNA sequence such as one transcribed from the other predicted pilus operons.

RT-PCR experiments on the ebp<sub>fm</sub> operon with primers designed to amplify intergenic regions of co-transcribed genes confirmed the expected amplification between ebpA<sub>fm</sub> and ebpB<sub>fm</sub> as well as ebpB<sub>fm</sub> and ebpC<sub>fm</sub> but not between ebpC<sub>fm</sub> and bps<sub>fm</sub> (Fig. 6b). Taken together, our data demonstrate unambiguously that the <i>E. faecium</i> Ebpfm pilus-encoding gene cluster represents a three-gene operon, unlike the <i>E. faecalis</i> locus, which produces a four-gene ebp–bps transcript in addition to an independent bps transcript (Nallapareddy et al., 2006).
Genes encoding MSCRAMP and pili proteins are commonly found in *E. faecium* isolates of clinical origin

We next examined 30 diverse (source/year/geographical location) human clinical isolates by colony hybridization using PCR-generated DNA probes. *E. faecium* strain TX0016 and the *acm* gene served as controls for the hybridization studies. As shown in Table 2, the percentage of isolates showing hybridization to the individual gene probes varied from 73% (*fms20*) to 100%. The range in the number of probes hybridizing per isolate varied from 11 genes in one isolate (that lacked *fms15, fms18, fms20* and *fms21*) to all 15 genes in 16 isolates (53%). One isolate lacked the ebpABC*fn* (*fms1-5-9*) operon and another lacked the *fms11-19-16* operon. The *scm* (*fms10), *fms14, fms17* and *fms13* genes were present in all isolates tested, while *fms15* was present in all but one. Among the endocarditis isolates included in this analysis, the multilocus sequence type 18 (ST18) isolate TX0068, belonging to the global epidemic hospital-associated clonal complex CC17 (Leavis et al., 2003), had all 15 genes, as did the sequenced ST18 isolate TX0016. Another endocarditis isolate, TX2355 (ST17), which also belongs to CC17, lacked *fms18*. A fourth endocarditis isolate, TX0074 (non-CC17, ST337), had 14 genes and lacked *fms20*. Thus, our results indicate broad but variable distribution of MSCRAMP-encoding genes in clinical isolates, consistent with a recent study of 22 *E. faecium* CWA-protein-encoding genes in which four of the genes, *fms11-19-16* (orf903-905-907) and *fms18* (orf2430) were shown to be specifically enriched in CC17 isolates, while the remaining nine genes were considered to be widespread in both CC17 and non-CC17 nosocomial isolates (Hendrickx et al., 2007). Notably, our preliminary data from screening for *scm* in several hundred *E. faecium* isolates belonging to clinical and non-clinical groups indicate that *scm* is absent only rarely from clinical isolates but is frequently absent from stool isolates (S. R. Nallapareddy, J. Sillanpää and B. E. Murray, unpublished data).

In summary, this study identified genes for 14 new *E. faecium* CWA proteins with MSCRAMP-like features. We characterized the function of one of these proteins, Scm, and demonstrated that it is a second collagen-binding protein for *E. faecium* with specificity for CV. We further showed that the identified minimal ligand-binding region has a β-sheet-rich secondary structure composition characteristic of the A-domains of MSCRAMMs. Scm was shown to be expressed on the surface of many *E. faecium* clinical isolates during *in vitro* growth. Eleven of the 14 MSCRAMP-like protein-encoding genes were clustered in four loci, and nine are predicted to encode the components needed to form multiple, distinct, pilus-like structures. We characterized one of these clusters (*ebpABC*fn*+bpsfn*) and showed that the four genes of this cluster are expressed in TX0082 as two transcripts, *ebpABC*fn* and bpsfn*. Subsequent detection of an HMM ladder pattern with anti-EbpCfn in Western blots of cell surface extracts from TX0082 provides experimental evidence for the assembly of EbpCfn into polymeric structures, such as pili. Our hybridization results for all 14 genes indicate a broad distribution of MSCRAMP- and pilus-encoding genes among clinical isolates; a future study will assess their distribution in natural *E. faecium* populations from a variety of sources. The ability of *E. faecium* to produce two surface-anchored collagen adhesins with different collagen type specificities, and the common occurrence of the *scm* gene as well as the 13 other newly identified MSCRAMP- and pilus-encoding genes among clinical *E. faecium* isolates, may have important implications for colonization and infection by this opportunistic pathogen.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grant R01 AI067861 from the Division of Microbiology and Infectious Diseases, National Institute of Allergy and Infectious Diseases (NIAID) to B.E.M. Additional support for this work is funded by NIH grant R01 AI20624 from the NIAID to M.H.

REFERENCES


encoding genes, srtA and bps (srtC), to biofilm formation and a murine model of urinary tract infection. Infect Immun 75, 5399–5404.


Edited by: K. E. Weaver