A Type I Restriction-Modification System Associated with *Enterococcus faecium* Subspecies Separation

Wenwen Huo[a], Hannah M. Adams[a], Cristian Trejo[a], Rohit Badia[a] and Kelli L. Palmer[a]*

[a]Department of Biological Sciences, University of Texas at Dallas, Richardson, Texas 75080

*Contact information for corresponding author:

Kelli Palmer: kelli.palmer@utdallas.edu

ABSTRACT

The gastrointestinal colonizer *Enterococcus faecium* is a leading cause of hospital-acquired infections. Multidrug-resistant (MDR) *E. faecium* are particularly concerning for infection treatment. Previous comparative genomic studies revealed that subspecies referred to as Clade A and Clade B exist within *E. faecium*. MDR *E. faecium* belong to Clade A, while Clade B consists of drug-susceptible fecal commensal *E. faecium*. Isolates from Clade A are further grouped into two sub-clades, A1 and A2. In general, Clade A1 isolates are hospital epidemic isolates whereas Clade A2 isolates are isolates from animals and sporadic human infections. Such phylogenetic separation indicates that reduced gene exchange occurs between the clades. We hypothesize that endogenous barriers to gene exchange exist between *E. faecium* clades. Restriction-modification (R-M) systems are such barriers in other microbes. We utilized bioinformatics analysis coupled with second generation and third generation deep sequencing platforms to characterize the methylome of two representative *E. faecium* strains, one from Clade A1 and one from Clade B. We identified a Type I R-M system that is Clade A1-specific, is active for DNA methylation, and significantly reduces transformability of Clade A1 *E. faecium*. Based on our results, we conclude that R-M systems act as barriers to horizontal gene exchange in *E. faecium* and propose that R-M systems contribute to *E. faecium* subspecies separation.

IMPORTANCE

Enterococcus faecium is a leading cause of hospital-acquired infections around the world. Rising antibiotic resistance in certain *E. faecium* lineages leaves fewer treatment options. The overarching aim of the attached work was to determine whether restriction-modification (R-M) systems contribute to the structure of the *E. faecium* species, wherein hospital-epidemic and non-hospital-epidemic isolates have distinct evolutionary histories and highly resolved clade

34 structures. R-M provides bacteria with a type of innate immunity to horizontal gene transfer
35 (HGT). We identified a Type I R-M system that is enriched in the hospital-epidemic clade and
36 determined that it is active for DNA modification activity and significantly impacts HGT. Overall,
37 this work is important because it provides a mechanism for the observed clade structure of *E.*
38 *faecium* as well as a mechanism for facilitated gene exchange among hospital-epidemic *E.*
39 *faecium*.
40

41 **INTRODUCTION**

42 *Enterococcus faecium* is a Gram-positive opportunistic pathogen that normally resides in the
43 gastrointestinal tracts of humans and other animals (1, 2). *E. faecium* can cause life-threatening
44 infections such as endocarditis and is among the leading causes of catheter-associated
45 bloodstream and urinary tract infections in clinical settings (3).

46

47 Previous comparative genomic studies revealed that subspecies exist within *E. faecium* (4-7).
48 Different names have been used by different groups to describe these clades; in this study, we
49 use the Clade A/B nomenclature. Generally speaking, MDR *E. faecium* belong to Clade A, while
50 Clade B consists of drug-susceptible fecal commensal *E. faecium* (8). Clade A is further split
51 into two subclades, A1 and A2, with hospital-endemic strains generally clustering in Clade A1
52 and sporadic infection isolates and animal isolates generally clustering in Clade A2 (8). Specific
53 phenotypes and genomic features are enriched in Clade A1 isolates relative to Clade A2 and B
54 isolates (8). Specifically, Clade A1 isolates have significantly higher mutation rates, larger
55 overall genome sizes including a larger core genome, and possess more mobile elements. On
56 the other hand, Clade A2 possesses a larger pan-genome than Clade A1 and B, possibly
57 reflective of the broader host origins of these strains. Given that Clade A and Clade B strains
58 would be expected to co-mingle in certain environments (for example, in hospital and municipal
59 sewage), the phylogenetic separation among the *E. faecium* clades suggests that they are not
60 sharing genetic information freely because of endogenous barriers to genetic exchange.

61

62 Horizontal gene transfer (HGT) is the exchange of genetic material between cells rather than
63 the vertical inheritance of genetic material from a parental cell. Bacteria can encode genome
64 defense mechanisms that can act in opposition to HGT. Two examples of these mechanisms
65 are clustered regularly interspaced short palindromic repeats (CRISPR) and associated proteins
66 (CRISPR-Cas) systems and restriction-modification (R-M) systems. CRISPR-Cas is a dynamic

67 immune system that utilizes sequence complementarity between self (CRISPR RNAs) and
68 foreign nucleic acid to carry out its restrictive function, whereas R-M discriminates self from
69 foreign DNA by DNA methylation patterns. If the *E. faecium* clades encode different defense
70 mechanisms, they may not exchange genetic information freely, thereby facilitating and
71 maintaining phylogenetic separation. However, little is known about CRISPR-Cas and R-M in *E.*
72 *faecium*. Genomic analysis suggests that these systems could contribute to the observed clade
73 structure of *E. faecium*. For example, CRISPR-Cas systems have been identified exclusively in
74 Clade B *E. faecium* and in sporadic Clade A-Clade B recombinant strains (8). For R-M, a
75 predicted methyl-directed restriction endonuclease (REase) is enriched in Clade A2 and B *E.*
76 *faecium* genomes relative to Clade A1 genomes (8).
77
78 Here, we focused on R-M systems and their roles in regulating gene exchange in *E. faecium*
79 because little is known about R-M defense in this species. Moreover, there is precedent in the
80 literature for R-M systems contributing to bacterial clade structure, as has been observed in
81 *Burkholderia* (9) and *Neisseria* (10). Our overarching hypothesis is that the *E. faecium* clades
82 encode different R-M systems, thereby inhibiting genetic exchange between them. In general,
83 R-M systems are composed of cognate methyltransferase (MTase) and REase activities and
84 are classified into different types based on the specific number and types of enzymes in the
85 system, as well as characteristics such as methylation type and pattern, cofactor requirement,
86 and restriction activity (11). A MTase recognizes specific sequences in the bacterial genome
87 and transfers a methyl group to either an adenine or a cytosine, resulting in 6-methyladenine
88 (m6A), 4-methylcytosine (m4C), or 5-methylcytosine (m5C). A REase may recognize the same
89 sequence as a MTase and cleave that region if the sequence is unmethylated (or in some
90 cases, if methylated). With the activities of MTases and REases, bacteria can use R-M to
91 impede entry of non-self DNA.
92

93 In this study, we used single-molecule real-time (SMRT) sequencing and whole genome
94 bisulfite sequencing to characterize the methylomes of representative *E. faecium* strains from
95 Clade A1 (*E. faecium* 1,231,502; or Efm502) and Clade B (*E. faecium* 1,141,733; or Efm733).
96 Two unique m6A methylation patterns were identified, one in each strain. These patterns were
97 asymmetric and bipartite, which is characteristic of Type I R-M methylation motifs (12).
98 Bioinformatic analyses were performed to identify candidate genes responsible for the
99 methylation. A unique Type I R-M system is encoded by each strain. We have named these
100 systems Efa502I (for Efm502) and Efa733I (for Efm733). Expression of these candidate
101 systems in *E. faecalis* heterologous hosts followed by SMRT sequencing confirmed that they
102 are responsible for the methylation patterns observed in Efm502 and Efm733. A functional
103 analysis was performed in order to assess the abilities of these systems to reduce *E. faecium*
104 HGT by transformation. In a comparative analysis among 73 *E. faecium* genomes, we found
105 that Efa502I is significantly enriched among Clade A1 isolates, while the Type I R-M system of
106 Efm733 appears to be strain-specific. Overall, this study is a first step towards understanding
107 the role of R-M in regulating HGT in *E. faecium* and the potential for R-M as one mechanism for
108 the clade structure of *E. faecium*.
109

110 **METHODS**

111 **Bacterial strains and growth conditions.** The strains used in this study are shown in Table 1.
112 All enterococci were grown in Brain Heart Infusion (BHI) broth or agar at 37°C, unless otherwise
113 stated. *Escherichia coli* strains were grown in Luria Broth (LB) at 37°C and with shaking at 225
114 rpm unless otherwise stated. Antibiotic concentrations for enterococcal strains were as follows:
115 rifampin, 50 µg/mL; fusidic acid, 25 µg/mL; spectinomycin, 500 µg/mL; streptomycin, 500 µg/mL;
116 chloramphenicol, 15 µg/mL. Antibiotic concentrations for *E. coli* strains were as follows:
117 chloramphenicol, 15 µg/mL; ampicillin, 100 µg/mL. All REases were purchased from New
118 England Biolabs (NEB) and used per the manufacturer's instructions. PCR was performed using

119 Taq polymerase (NEB) or Phusion (Fisher). Sanger sequencing to validate all genetic
120 constructs was performed at the Massachusetts General Hospital DNA Core facility (Boston,
121 MA).
122
123 **Isolation of genomic DNA.** Enterococcal strains were cultured overnight in BHI broth prior to
124 genomic DNA (gDNA) extraction. The extraction was performed using a Qiagen Blood and
125 Tissue DNeasy Kit using a previously published protocol (13). To isolate *E. coli* gDNA, bacteria
126 were grown overnight in LB broth prior to extraction using either the Blood and Tissue DNeasy
127 kit (Qiagen) or the UltraClean Microbial DNA Isolation Kit (Qiagen) per the manufacturer's
128 instructions.
129
130 **SMRT sequencing and methylome detection.** SMRT sequencing was performed by the
131 Johns Hopkins Medical Institute Deep Sequencing and Microarray Core. After sequencing, the
132 reads were assembled into contigs and analyzed using the RS modification and motif detection
133 protocol in SMRT portal v1.3.3. An *in silico* control was used as a methylation baseline. SMRT
134 sequencing in *E. faecalis* expressing Efa502I or Efa733I was performed by the University of
135 Michigan sequencing core facility. Reads were mapped to the *E. faecalis* OG1RF reference
136 sequence (GenBank accession number NC_017316), and the methylation motifs were detected
137 using the RS modification and motif detection protocol in SMRT portal v.2.3.2.
138
139 **Bioinformatic analysis of R-M systems in eight *E. faecium* genomes.** The entire protein
140 complement for eight previously sequenced *E. faecium* isolates (14) was analyzed. To identify
141 potential MTases, the REBASE Gold Standard list (15) was used as a reference. This list is
142 comprised of biochemically verified MTases and REases. Each protein sequence from *E.*
143 *faecium* genomes was analyzed using BlastP against the REBASE Gold Standard list. The
144 protein sequences with significant (e-value $<1e^{-3}$) homology to REBASE Gold Standard proteins

145 were further filtered by protein size. If an *E. faecium* query protein length was less than half of
146 its subject's length, the match was removed from the prediction list. Due to the sequence
147 diversity of REases which complicates their bioinformatic identification (15), guilt-by-association
148 was used to identify full R-M systems as we have previously described (16). The proteins
149 encoded near candidate DNA MTases were analyzed using BLAST and Pfam for conserved
150 domains consistent with REase activities and/or sequence identity to confirmed REases. The
151 amino acid sequence of each R-M candidate was then pairwise compared among all the eight
152 strains to identify putative orthologs. If two protein sequences shared an amino acid identity
153 ≥90% with query coverage ≥90%, they were considered to be orthologous.
154
155 **Expression of R-M systems in *E. faecalis* heterologous hosts.** Genes encoding the
156 specificity and methylation subunits of Efa733I (EFSG_05028-EFSG_05027) were PCR-
157 amplified in their entirety, including the upstream region to retain the native promoter, using
158 primers 733_T1A_SM_F and 733_T1A_SM_R (see Table S1 for primer sequences). The PCR
159 product was digested with *Not*I and ligated into *Not*I-digested pWH03 (16) using T4 DNA Ligase
160 (NEB), generating pHA102. pWH03 is a pLT06 derivative for expression of genes from a
161 previously validated neutral genomic insertion site (EF2238-EF2239) for expression (GISE) (16,
162 17). pHA102 constructs were then introduced into *E. coli* DH5α via heat shock for propagation
163 and sequence confirmation. pHA102 was electroporated into *E. faecalis* OG1SSp using a
164 previously described method (18). An *E. faecalis* OG1SSp derivative with a chromosomal
165 integration of Efa733I, referred to as OG1SSp::*efa733I*, was generated by temperature shifts
166 and *p*-chlorophenylalanine counterselection, as previously described (19).
167
168 Genes encoding the specificity and methylation subunits of Efa502I (EFQG_01131-
169 EFQG_01132) were PCR-amplified using primers 502_T1A_SM_F and 502_T1A_SM_R. The
170 PCR product was then TA-cloned into the pGEM-T Easy Vector (Promega) and introduced into

171 DH5α via heat shock to generate pGEM-SMA1. pGEM-SMA1 was then digested with *Not*I, and
172 the digestion reaction was used as insert for ligation into *Not*I-digested pWH03. The ligation
173 reaction was then introduced into DH5α, and colonies were screened for chloramphenicol
174 resistance and ampicillin susceptibility to ensure the pGEM backbone was not ligated into
175 pWH03. Once the construct, referred to as pHA103, was confirmed via Sanger sequencing, it
176 was introduced into electrocompetent *E. faecalis* OG1RF. An *E. faecalis* OG1RF derivative with
177 a chromosomal integration of Efa502I, referred to as OG1RF::*efa502I*, was generated by
178 temperature shifts and *p*-chlorophenylalanine counterselection. All plasmids and strains for
179 heterologous expression were validated by PCR and Sanger sequencing.
180
181 **Generation of *E. faecium* R-M deletion mutants.** Regions up- and downstream of Efa502I
182 and Efa733I were PCR-amplified using primers listed in Table S1, ligated into pLT06, and
183 transformed into *E. coli* EC1000 (20), generating pWH16 and pWH17 (Table 1). Insert
184 sequences were confirmed using Sanger sequencing. *E. faecium* strains were made
185 electrocompetent using previously a published protocol (21). 2 µg of sequence-confirmed
186 plasmids were electroporated into electrocompetent Efm733 and Efm502. The generation of
187 deletion mutants was accomplished using temperature shifts and *p*-chlorophenylalanine
188 counterselection, as previously described (19). The successful deletion mutants were
189 sequence-confirmed by PCR and Sanger sequencing.
190
191 **Transformation efficiency test.** Efm733, Efm502, and their respective R-M deletion mutants
192 were made electrocompetent using a modified version of the previously published protocol (21).
193 Briefly, overnight cultures were diluted 10-fold in BHI and cultured to OD_{600nm} ~ 0.6. The bacteria
194 were then pelleted and treated with filter-sterilized lysozyme buffer (10 mM Tris-HCl pH 8.0, 10
195 mM EDTA pH 8.0, 50 mM NaCl) supplemented with 83 µL of 2.5 KU/mL mutanolysin stock for
196 30 min at 37°C. The cells were then pelleted and washed three times with ice-cold filter-

197 sterilized electroporation buffer (0.5 M sucrose and 10% glycerol). Finally, the cells were
198 pelleted and resuspended in electroporation buffer and aliquoted for storage at -80°C and future
199 use. 1 µg pAT28 (22) was electroporated into the electrocompetent *E. faecium* cells. The counts
200 of total viable cells and spectinomycin-resistant cells were determined by serial dilution and
201 plating. The transformation efficiency was expressed as percent of transformed (spectinomycin-
202 resistant) cells per total viable cells. Three independent experiments were performed and the
203 statistical significance was assessed using the unpaired one-tailed Student's t-test.
204
205 **Distribution analysis of putative R-M systems and orphan MTases.** The amino acid
206 sequences for select R-M system and orphan MTase candidates were queried against a
207 collection of 73 *E. faecium* isolates previously analyzed by Lebreton *et al* (8) using BLASTP.
208 Any proteins which shared >90% query coverage and amino acid identity were considered
209 orthologs. The Fisher's exact test was used to determine if an orphan MTase or R-M system
210 was significantly over- or under-represented in a particular clade.
211
212 **REase protection assays.** To identify m5C methylation, gDNA was treated with the
213 methylation-sensitive REases McrBC, FspEI, and MspJI (NEB). 500 ng gDNA was incubated
214 with each REase at 37°C for 3 h (McrBC) or 6 h (FspEI and MspJI) followed by analysis by
215 electrophoresis on a 1% agarose gel with ethidium bromide.
216
217 **Bisulfite sequencing.** Whole-genome bisulfite sequencing libraries were constructed using the
218 Illumina TruSeq LT PCR FREE kit and the Qiagen EpiTect Bisulfite kit. Native DNA was isolated
219 as described above. Whole-genome-amplified (WGA) control DNA was generated by
220 amplification of native gDNA using the Qiagen REPLI-g® kit, per the manufacturer's
221 instructions. For bisulfite sequencing, briefly, 2 µg each of native and WGA control DNA were
222 fragmented using NEB fragmentase. DNA fragments ranging from 200 bp to 700 bp were gel

223 extracted and end-repaired. After A-tailing of DNA fragments, Illumina TruSeq adapters were
224 ligated. Then, the bisulfite conversion was performed using the Qiagen EpiTect Bisulfite kit, per
225 the manufacturer's instruction. An 8-cycle PCR enrichment with Illumina primer mix was
226 performed, followed by size selection and gel purification. The libraries were sequenced using
227 Illumina MiSeq with 2x75 bp paired-end chemistry.
228
229 **Whole genome bisulfite sequencing analysis.** The sequencing reads were analyzed using
230 Bismark (23) with additional quality control and filtering as described previously (16). Briefly, the
231 Illumina reads were mapped to the *in silico* bisulfite-converted references (23). Then, we
232 quantified the conversion rate of each mapped read by calculating the percentage of converted
233 C (which will result in T) to the total number of C in the reference within the mapped region. The
234 mapped reads with ≤80% conversion rate were filtered out from analysis (16). Next, the
235 coverage depth and methylation ratio were calculated for each C site. The methylation ratio was
236 calculated by dividing the total number of C by the coverage depth at each C site. A fully
237 methylated C, thus protected from bisulfite conversion, will have a methylation ratio near 1. An
238 unmethylated C will have a methylation ratio near 0. To identify consensus methylation motifs, C
239 sites with ≥0.35 methylation ratio and ≥10 coverage depth, along with the sequences of 5 bp
240 upstream and 5 bp downstream, were extracted. The extracted sequences were subjected for
241 MEME motif search (24).
242
243 **Confirmation of m5C MTase activity**. Primers EFSG_00659_F and EFSG_00659_R (Table
244 S1) were used to amplify the entire Efm733 EFSG_00659 coding region and its upstream
245 predicted promoter. The PCR product was then cloned into the pGEM-T Easy Vector (Promega)
246 per the manufacturer's instructions and transformed into *E. coli* STBL4 (Fisher) to generate
247 pRB01. REase digestion assays with methylation-sensitive enzymes were performed on purified
248 *E. coli* and *E. faecium* gDNA as described above.

249

250 **Accession numbers**. DNA sequence data generated in this study have been deposited in the
251 Sequence Read Archive under accession numbers PRJNA397049 (for SMRT sequencing data)
252 and PRJNA488088 (for Illumina bisulfite sequencing data).

253

254 **RESULTS**

255 **Identification of Clade A1-specific putative Type I R-M system in *E. faecium***. We previously
256 reported that a Type II R-M system significantly reduces HGT via conjugation (18) and
257 transformation (16) in *E. faecalis*. Here, we hypothesize that the *E. faecium* clades encode
258 distinct R-M systems that reduce the exchange of genetic information between them. We
259 utilized an approach we previously developed for *E. faecalis* R-M analysis (18) to predict
260 potential R-M systems in eight previously sequenced *E. faecium* genomes. The 8 genomes
261 included 3 genomes from Clade A1, 3 genomes from Clade B, one genome from Clade A2, and
262 one recombinant Clade A1/B hybrid (5, 8). Because REases are difficult to identify with
263 bioinformatics, and MTase prediction is comparatively straightforward, as has been previously
264 reported by NEB (15), we first identified predicted DNA MTases in *E. faecium* genomes, and
265 then analyzed surrounding genes for predicted R-M-related activities. The complete list of
266 candidates for the eight strains is shown in Table 2.

267

268 Interestingly, we predicted at least one putative Type I R-M system for seven of the eight *E.*
269 *faecium* strains (Table 2). Type I R-M systems are multisubunit complexes comprised of a
270 specificity subunit (S), a methylation subunit (M), and a restriction subunit (R) (11, 25-27). The S
271 subunit is responsible for the specific DNA recognition motif and associates with the DNA to
272 bring the M and R subunits into contact. The system has two conformations: M_2S_1, which is
273 capable of methylating DNA based on the recognition sequence, and $R_2M_2S_1$, which is capable
274 of restricting DNA (27, 28). One predicted *E. faecium* Type I R-M system is comprised of highly

275 conserved (>90% amino acid sequence identity) M and R subunits in six of eight genomes
276 across both Clade A1 and Clade B (Table 2 and Dataset S1). The specificity subunit from this
277 system, however, is highly conserved in Clade A1 genomes but not in Clade B (Table 2 and Fig.
278 1a-b). S subunits possess two target recognition domains (TRDs) that determine the nucleotide
279 sequence the subunit binds to (29, 30). The variation in amino acid sequence between the S
280 subunits occurs within these TRDs (Fig 1a), suggesting that these S subunits recognize
281 different DNA sequences. Notably, the S subunits from Clade A1 strains are identical to each
282 other, indicating that they utilize the same recognition sequence.
283
284 To examine the distribution of this putative Clade A1-specific system in a larger collection of *E.*
285 *faecium* strains, we analyzed 73 *E. faecium* genomes of mostly draft status that were reported
286 previously (8). This list includes 15 clade B isolates, 21 clade A1 isolates, 35 clade A2 isolates,
287 and 2 hybrid isolates (Table S2). We selected Efm502 (Clade A1) as our representative Clade
288 A1 strain for this analysis and used Type I R-M sequences from this genome as references for
289 analysis against the broader collection of *E. faecium* strains. The M and R subunits of the
290 putative Clade A1-specific Type I system were detected in 51 and 52, respectively, of 73 *E.*
291 *faecium* genomes, including both Clade A and Clade B strains (Fig S1a). However, the
292 distribution of the S subunits varied (Fig S1a-b). The S subunit present in Efm502,
293 EFQG_01131, was significantly enriched within Clade A1 isolates (14/21; p-value <0.0001 using
294 Fisher exact test; Fig S1a) and absent from all other clades with the exception of strain
295 EnGen002, which is classified as a Clade A1/B hybrid strain. Interestingly, the S subunits
296 present in most other *E. faecium* strains are strain-specific by the strict thresholds applied here.
297 Given that the Efm502 S subunit is enriched in Clade A1, we hypothesize that many Clade A1
298 strains exchange genetic information freely with each other while exchange with other *E.*
299 *faecium* strains is restricted.
300

301 **SMRT sequencing for *E. faecium* methylome analysis.** We analyzed the Efm502 and
302 Efm733 genomes by SMRT sequencing. SMRT sequencing measures the kinetics of DNA
303 polymerase as it synthesizes DNA in order to identify bases that have been modified (31-33). It
304 has been extensively utilized for bacterial methylome analysis (34-42). With SMRT sequencing,
305 6-methyladenine (m6A) and 4-methylcytosine (m4C) can be easily detected with modest
306 sequence coverage (~25x per strand), while 5-methylcytosine (m5C) detections requires high
307 coverage (~250x per strand) (41, 43). Using SMRT sequencing, we identified two unique m6A
308 methylation motifs in Efm502 and Efm733. Efm502 possessed m6A methylation at the
309 underlined position of the motif 5'-RAYCNNNNNNTTRG-3' and Efm733 possessed m6A
310 methylation at the underlined position of the motif 5'-AGAWNNNNATTA-3' (Table 3). These
311 sequences are asymmetric and bipartite, which is characteristic of Type I R-M methylation (12).
312 Due to the coverage of our SMRT sequencing, m5C modification could not be accurately
313 detected. The two unique m6A methylation patterns indicate that DNA from one strain would be
314 recognized as foreign should it cross the strain barrier.
315
316 **Expression in heterologous hosts links methylation activity to genes in Efm502 and**
317 **Efm733.** According to our predictions (Table 2), there is only one complete Type I R-M system
318 encoded by each of Efm502 and Efm733. To determine if these systems are responsible for the
319 methylation patterns identified by SMRT sequencing, we expressed the respective S and M
320 subunits (EFQG_01131-01132 for Efm502 and EFSG_05028-05027 for Efm733) in the
321 heterologous host *E. faecalis* OG1RF or its spectinomycin/streptomycin-resistant relative
322 OG1SSp. Previous work in our lab had characterized the methylome of OG1RF using SMRT
323 and bisulfite sequencing (16). This allowed us to attribute any new methylation patterns
324 observed during SMRT sequencing to the *E. faecium* genes that were expressed in the OG1RF
325 background. SMRT sequencing of these strains detected the same methylation patterns
326 originally identified in Efm502 and Efm733 (Table 3). These data demonstrate that

327 EFQG_01131-01132 is responsible for the 5'-RAYCNNNNNNTTRG-3' methylation in Efm502
328 and that EFSG_05028-05027 is responsible for the 5'-AGAWNNNNATTA-3' methylation in
329 Efm733. Because we have confirmed the function of these genes, we have named them
330 Efa502I and Efa733I, which is consistent with the R-M system nomenclature convention
331 established by New England Biolabs (12).
332
333 **Efa502I and Efa733I reduce transformation efficiency in *E. faecium*.** To determine whether
334 the Type I R-M systems in Efm733 and Efm502 actively defend against exogenous DNA, we
335 constructed null strains (Efm733ΔRM and Efm502ΔRM; Table 1) and evaluated their
336 transformation efficiencies relative to their wild-type parent strains. Here, we utilized the broad
337 host range plasmid pAT28 (Table 1) (22). pAT28 sequence has motifs recognized by the Type I
338 R-M systems in both wild-type Efm733 (1 occurrence) and Efm502 (1 occurrence). The
339 transformation of pAT28 into Efm733 and Efm502 served as a baseline for the experiment. If
340 Efa733I and Efa502I are active, we expect to see higher pAT28 transformation efficiencies into
341 Efm733ΔRM and Efm502ΔRM, respectively. Indeed, we observed significantly higher
342 transformation efficiencies into Type I R-M system null strains (Fig 2; p-value<0.05 using one-
343 tailed Student's t-test). These data demonstrate that the Type I R-M systems in Efm733 and
344 Efm502 actively function as mechanisms of genome defense.
345
346 **m5C methylation occurs in Efm733.** As described previously, our SMRT sequencing had
347 insufficient coverage depth for m5C methylome characterization. Hence, we used REase
348 protection assays with commercially available methylation-sensitive REases to query the
349 presence of m5C methylation in our eight *E. faecium* strains. Table 4 summarizes the
350 recognition sequences and modifications of the enzymes used in this study. Only Efm733
351 showed evidence of cytosine modification, as it was digested by MspJI (Table 4; Fig S2).
352

353 To determine the exact cytosine methylation motif present in Efm733, gDNA was subjected to
354 whole-genome bisulfite sequencing. Whole genome amplified (WGA) DNA was used as
355 negative control since WGA removes all modifications. During bisulfite treatment, cytosine
356 bases are converted to thymine unless they are protected by either m4C or m5C methylation.
357 Additionally, our lab has previously published a method of distinguishing between m4C and
358 m5C methylation using thymine conversion ratios of sequencing reads after bisulfite treatment
359 (44). m5C methylation is sufficient to protect the cytosine residue completely from bisulfite
360 conversion, so that most sequencing reads at that position contain the original cytosine base.
361 However, m4C methylation provides only partial protection from bisulfite conversion, so a
362 thymine conversion rate of 0.5 at a particular position within the sequencing reads suggests the
363 presence of m4C methylation. Bisulfite conversion and subsequent sequencing revealed that
364 Efm733 possesses m5C modification at the motif 5'-R^{m}<u>C</u>CGGY-3' (Table 5, Fig 3, and Fig S3;
365 the methylation occurs at the underlined position), which overlaps the MspJI recognition site (5'-
366 CNNR-3') and hence supports the evidence of methylation obtained from the MspJI digestion
367 assays. Based on the cytosine conversion ratio of close to 1.0, m5C modification is supported,
368 which is consistent with why it was not detected by our SMRT sequencing.
369

370 **EFSG_00659 is responsible for m5C methylation in Efm733.** Based on the bioinformatics
371 analyses, we hypothesized that EFSG_00659 was responsible for the m5C methylation found in
372 Efm733 (Table 2). EFSG_00659 possessed no homologs in the other seven strains analyzed,
373 making it a good candidate for the unique methylation found in Efm733. We queried the
374 EFSG_00659 protein sequence against the REBASE gold standard list and identified that it has
375 high sequence similarity to M.AvaIX, M.VchO395I and M.VchAI (e-value ≤ $3e^{-125}$; recognition
376 sites are 5'-RCCGGY-3'). Interestingly, BLASTP identified no significant hits when
377 EFSG_00659 was queried against the larger collection of 73 *E. faecium* genomes, indicating its
378 unique presence in Efm733.

379

380 In order to link EFSG_00659 with the m5C methylation identified during bisulfite sequencing, we
381 expressed it in the heterologous host *E. coli* STBL4 and performed an REase protection assay.
382 The REase AgeI recognizes the motif 5'-A<u>C</u>CGGT-3', and its enzymatic activity is blocked if
383 m5C methylation is present at the underlined position. This motif overlaps the m5C methylation
384 motif in Efm733 identified by bisulfite sequencing. If the motif is methylated, DNA will be
385 protected against digestion. We cloned EFSG_00659 into the vector pGEM-T and transformed it
386 into *E. coli* STBL4, generating strain *E. coli* STBL4(pRB01). Genomic DNA from Efm733,
387 STBL4(pGEM-T), and STBL4(pRB01) was treated with AgeI per the manufacturer's instructions.
388 EcoRI was used as a positive control for digestion. Figure 4 shows representative results of the
389 digestions on a 1% agarose gel. As expected, Efm733 was digested by EcoRI and protected
390 against digestion from AgeI. STBL4(pGEM-T) was digested by both EcoRI and AgeI, indicating
391 that the original host and empty vector pGEM-T did not possess the appropriate m5C
392 methylation. STBL4(pRB01) was protected against digestion by AgeI, demonstrating that
393 EFSG_00659 is responsible for the 5'-R^{m}CCGGY-3' methylation found in Efm733.

394

395 **DISCUSSION**

396 In this study, we used a combination of genomic and genetic approaches to identify a functional
397 Type I R-M system that is enriched in Clade A1 *E. faecium* and that significantly alters
398 transformability of a model Clade A1 strain. We propose that this R-M system impacts HGT
399 rates among *E. faecium* mixed-clade communities, thereby helping to maintain the observed
400 phylogenetic structure of *E. faecium* and facilitating HGT specifically among Clade A1 strains.
401 Mixed communities of *E. faecium* clades are expected to occur in environments where healthy
402 and ill human hosts, human and animal hosts, and/or the feces of any of these hosts co-mingle
403 (i.e. in sewage). In future studies, we plan to assess the impact of R-M on conjugative plasmid

transfer, which is a major mode of HGT in enterococci and was not assessed in our current study.

An interesting observation from our study is the sequence diversity of Type I S subunits encoded within *E. faecium* Type R-M systems having nearly identical R and M subunits (Fig S1a-b). After further investigation into those alignments, we found that those S subunits sharing 50-70% overall amino acid sequence identities possess sequence diversity within one TRD domain, where the other TRD domain and the central conserved domain are conserved. This suggests that these systems share partial recognition sequences. Previous research has reported that the diversification of Type I R-M recognition sequences is driven by TRD exchanges, permutation of the dimerization domain, and circular permutation of TRDs (45). Our observation suggests that TRD recombination and reorganization events occur for *E. faecium* Type I R-M systems outside Clade A1. Future studies will use genomics to further explore the relationship between S subunit sequence diversity and its impact on *E. faecium* methylomes and inter-strain and inter-clade HGT.

Acknowledgments

This work was supported by Public Health Service grants K22AI099088 and R01AI116610 to K.L.P.

424 **References**

425 1. Noble CJ. 1978. Carriage of group D streptococci in the human bowel. J Clin Pathol 31:1182-6.

426 2. Chenoweth C, Schaberg D. 1990. The epidemiology of enterococci. Eur J Clin Microbiol Infect Dis
427 9:80-9.

428 3. Agudelo Higuita NI, Huycke MM. 2014. Enterococcal Disease, Epidemiology, and Implications for
429 Treatment. *In* Gilmore MS, Clewell DB, Ike Y, Shankar N (ed), Enterococci: From Commensals to
430 Leading Causes of Drug Resistant Infection, Boston.

431 4. Louis E, Galloway-Peña J, Roh JH, Latorre M, Qin X, Murray BE. 2012. Genomic and SNP Analyses
432 Demonstrate a Distant Separation of the Hospital and Community-Associated Clades of
433 Enterococcus faecium. PLoS ONE 7:e30187.

434 5. Palmer KL, Godfrey P, Griggs A, Kos VN, Zucker J, Desjardins C, Cerqueira G, Gevers D, Walker S,
435 Wortman J, Feldgarden M, Haas B, Birren B, Gilmore MS. 2012. Comparative genomics of
436 enterococci: variation in Enterococcus faecalis, clade structure in E. faecium, and defining
437 characteristics of E. gallinarum and E. casseliflavus. MBio 3:e00318-11.

438 6. Willems RJL, Top J, van Schaik W, Leavis H, Bonten M, Siren J, Hanage WP, Corander J. 2012.
439 Restricted Gene Flow among Hospital Subpopulations of Enterococcus faecium. mBio 3.

440 7. van Schaik W, Top J, Riley DR, Boekhorst J, Vrijenhoek JE, Schapendonk CM, Hendrickx AP,
441 Nijman IJ, Bonten MJ, Tettelin H, Willems RJ. 2010. Pyrosequencing-based comparative genome
442 analysis of the nosocomial pathogen Enterococcus faecium and identification of a large
443 transferable pathogenicity island. BMC Genomics 11:239.

444 8. Lebreton F, van Schaik W, McGuire AM, Godfrey P, Griggs A, Mazumdar V, Corander J, Cheng L,
445 Saif S, Young S, Zeng Q, Wortman J, Birren B, Willems RJ, Earl AM, Gilmore MS. 2013. Emergence
446 of epidemic multidrug-resistant Enterococcus faecium from animal and commensal strains.
447 MBio 4.

448 9. Nandi T, Holden MTG, Didelot X, Mehershahi K, Boddey JA, Beacham I, Peak I, Harting J,
449 Baybayan P, Guo Y, Wang S, How LC, Sim B, Essex-Lopresti A, Sarkar-Tyson M, Nelson M, Smither
450 S, Ong C, Aw LT, Hoon CH, Michell S, Studholme DJ, Titball R, Chen SL, Parkhill J, Tan P. 2015.
451 Burkholderia pseudomalleisequencing identifies genomic clades with distinct recombination,
452 accessory, and epigenetic profiles. Genome Research 25:129-141.

453 10. Budroni S, Siena E, Hotopp JCD, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR,
454 Daugherty SC, Angiuoli SV, Covacci A, Pizza M, Rappuoli R, Moxon ER, Tettelin H, Medini D. 2011.
455 Neisseria meningitidis is structured in clades associated with restriction modification systems
456 that modulate homologous recombination. Proceedings of the National Academy of Sciences
457 108:4494-4499.

458 11. Tock MR, Dryden DT. 2005. The biology of restriction and anti-restriction. Curr Opin Microbiol
459 8:466-72.

460 12. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev S,
461 Dryden DT, Dybvig K, Firman K, Gromova ES, Gumport RI, Halford SE, Hattman S, Heitman J,
462 Hornby DP, Janulaitis A, Jeltsch A, Josephsen J, Kiss A, Klaenhammer TR, Kobayashi I, Kong H,
463 Kruger DH, Lacks S, Marinus MG, Miyahara M, Morgan RD, Murray NE, Nagaraja V, Piekarowicz
464 A, Pingoud A, Raleigh E, Rao DN, Reich N, Repin VE, Selker EU, Shaw PC, Stein DC, Stoddard BL,
465 Szybalski W, Trautner TA, Van Etten JL, Vitor JM, Wilson GG, Xu SY. 2003. A nomenclature for
466 restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic
467 Acids Res 31:1805-12.

468 13. Adams HM, Li X, Mascio C, Chesnel L, Palmer KL. 2015. Mutations associated with reduced
469 surotomycin susceptibility in Clostridium difficile and Enterococcus species. Antimicrob Agents
470 Chemother 59:4139-47.

471 14. Palmer KL, Carniol K, Manson JM, Heiman D, Shea T, Young S, Zeng Q, Gevers D, Feldgarden M,
472 Birren B, Gilmore MS. 2010. High-quality draft genome sequences of 28 Enterococcus sp.
473 isolates. J Bacteriol 192:2469-70.

474 15. Roberts RJ, Vincze T, Posfai J, Macelis D. 2015. REBASE—a database for DNA restriction and
475 modification: enzymes, genes and genomes. Nucleic Acids Research 43:D298-D299.

476 16. Huo W, Adams HM, Zhang MQ, Palmer KL. 2015. Genome Modification in Enterococcus faecalis
477 OG1RF Assessed by Bisulfite Sequencing and Single-Molecule Real-Time Sequencing. J Bacteriol
478 197:1939-51.

479 17. DebRoy S, van der Hoeven R, Singh KV, Gao P, Harvey BR, Murray BE, Garsin DA. 2012.
480 Development of a genomic site for gene integration and expression in Enterococcus faecalis.
481 Journal of Microbiological Methods 90:1-8.

482 18. Price VJ, Huo W, Sharifi A, Palmer KL. 2016. CRISPR-Cas and Restriction-Modification Act
483 Additively against Conjugative Antibiotic Resistance Plasmid Transfer in Enterococcus faecalis.
484 mSphere 1.

485 19. Thurlow LR, Thomas VC, Hancock LE. 2009. Capsular polysaccharide production in Enterococcus
486 faecalis and contribution of CpsF to capsule serospecificity. J Bacteriol 191:6203-10.

487 20. Leenhouts K, Buist G, Bolhuis A, ten Berge A, Kiel J, Mierau I, Dabrowska M, Venema G, Kok J.
488 1996. A general system for generating unlabelled gene replacements in bacterial chromosomes.
489 Mol Gen Genet 253:217-24.

490 21. Bhardwaj P, Ziegler E, Palmer KL. 2016. Chlorhexidine Induces VanA-Type Vancomycin
491 Resistance Genes in Enterococci. Antimicrob Agents Chemother 60:2209-21.

492 22. Trieu-Cuot P, Carlier C, Poyart-Salmeron C, Courvalin P. 1990. A pair of mobilizable shuttle
493 vectors conferring resistance to spectinomycin for molecular cloning inEscherichia coliand in
494 Gram-positive bacteria. Nucleic Acids Research 18:4296-4296.

495 23. Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq
496 applications. Bioinformatics 27:1571-2.

497 24. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009.
498 MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202-8.

499 25. Luria SE, Human ML. 1952. A nonhereditary, host-induced variation of bacterial viruses. J
500 Bacteriol 64:557-69.

501 26. Bertani G, Weigle JJ. 1953. Host controlled variation in bacterial viruses. J Bacteriol 65:113-21.

502 27. Murray NE. 2000. Type I restriction systems: sophisticated molecular machines (a legacy of
503 Bertani and Weigle). Microbiol Mol Biol Rev 64:412-34.

504 28. Taylor I, Patel J, Firman K, Kneale G. 1992. Purification and biochemical characterisation of the
505 EcoR124 type I modification methylase. Nucleic Acids Res 20:179-86.

506 29. Gough JA, Murray NE. 1983. Sequence diversity among related genes for recognition of specific
507 targets in DNA molecules. J Mol Biol 166:1-19.

508 30. Gann AA, Campbell AJ, Collins JF, Coulson AF, Murray NE. 1987. Reassortment of DNA
509 recognition domains and the evolution of new specificities. Mol Microbiol 1:13-22.

510 31. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for
511 single-molecule nanopore DNA sequencing. Nature Nanotechnology 4:265-270.

512 32. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A,
513 Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M,
514 Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S,
515 Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R,
516 Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A,
517 Zaccarin D, et al. 2009. Real-time DNA sequencing from single polymerase molecules. Science
518 323:133-8.

519 33. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R,
520 Wegener J, Turner SW. 2010. Real-time DNA sequencing from single polymerase molecules.
521 Methods Enzymol 472:431-55.

522 34. Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, Fomenkov A, Roberts RJ,
523 Korlach J. 2012. Characterization of DNA methyltransferase specificities using single-molecule,
524 real-time DNA sequencing. Nucleic Acids Research 40:e29-e29.

525 35. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC,
526 Jabado OJ, Deikus G, Clark TA, Luong K, Murray IA, Davis BM, Keren-Paz A, Chess A, Roberts RJ,
527 Korlach J, Turner SW, Kumar V, Waldor MK, Schadt EE. 2012. Genome-wide mapping of
528 methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time
529 sequencing. Nature Biotechnology 30:1232-1239.

530 36. Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, Fomenkov A, Turner SW,
531 Korlach J, Roberts RJ. 2012. The methylomes of six bacteria. Nucleic Acids Res 40:11450-62.

532 37. Bendall ML, Luong K, Wetmore KM, Blow M, Korlach J, Deutschbauer A, Malmstrom RR. 2013.
533 Exploring the Roles of DNA Methylation in the Metal-Reducing Bacterium Shewanella oneidensis
534 MR-1. Journal of Bacteriology 195:4966-4974.

535 38. Davis BM, Chao MC, Waldor MK. 2013. Entering the era of bacterial epigenomics with single
536 molecule real time DNA sequencing. Current Opinion in Microbiology 16:192-198.

537 39. Kozdon JB, Melfi MD, Luong K, Clark TA, Boitano M, Wang S, Zhou B, Gonzalez D, Collier J, Turner
538 SW, Korlach J, Shapiro L, McAdams HH. 2013. Global methylation state at base-pair resolution of
539 the Caulobacter genome throughout the cell cycle. Proceedings of the National Academy of
540 Sciences 110:E4658-E4667.

541 40. Richardson PM, Lluch-Senar M, Luong K, Lloréns-Rico V, Delgado J, Fang G, Spittle K, Clark TA,
542 Schadt E, Turner SW, Korlach J, Serrano L. 2013. Comprehensive Methylome Characterization of
543 Mycoplasma genitalium and Mycoplasma pneumoniae at Single-Base Resolution. PLoS Genetics
544 9:e1003191.

545 41. Roberts RJ, Carneiro MO, Schatz MC. 2013. The advantages of SMRT sequencing. Genome
546 Biology 14.

547 42. Krebes J, Morgan RD, Bunk B, Spröer C, Luong K, Parusel R, Anton BP, König C, Josenhans C,
548 Overmann J, Roberts RJ, Korlach J, Suerbaum S. 2014. The complex methylome of the human
549 gastric pathogen Helicobacter pylori. Nucleic Acids Research 42:2415-2432.

550 43. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. 2010.
551 Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods
552 7:461-5.

553 44. Huo W. 2017. Enterococcus faecalis Genome Defense Systems and Their Impact on Conjugative
554 Antibiotic Resistance Plasmid Transfer. 10970358. The University of Texas at Dallas, Ann Arbor.

555 45. Loenen WAM, Dryden DTF, Raleigh EA, Wilson GG. 2013. Type I restriction enzymes and their
556 relatives. Nucleic Acids Research 42:20-44.

557 46. Gold OG, Jordan HV, van Houte J. 1975. The prevalence of enterococci in the human mouth and
558 their pathogenicity in animal models. Arch Oral Biol 20:473-7.

559

560

561 **Table 1. Strains used in this study.**

Strain	Description	Reference
Enterococcus faecium		
1,231,502	Clade A1 isolate; also referred to as Efm502	(14)
1,230,933	Clade A1 isolate	(14)
1,231,410	Clade A1 isolate	(14)
1,231,408	Hybrid Clade A1/B isolate	(14)
1,231,501	Clade A2 isolate	(14)
1,141,733	Clade B isolate; also referred to as Efm733	(14)
Com12	Clade B isolate	(14)
Com15	Clade B isolate	(14)
Efm733ΔRM	Efm733 with deletion of Efa733I (EFSG_05027-29)	This study
Efm502ΔRM	Efm502 with deletion of Efa502I (EFQG_01130-32)	This study
Enterococcus faecalis		
OG1RF	Rifampicin- and fusidic acid-resistant derivative of *E. faecalis* OG1	(46)
OG1SSp	Streptomycin- and spectinomycin-resistant derivative of *E. faecalis* OG1	(46)
OG1RF::*efa502I*	OG1RF with Efa502I inserted at GISE site	This study
OG1SSp::*efa733I*	OG1SSp with Efa733I inserted at GISE site	This study
Escherichia coli		
EC1000	Plasmid propagation host	(20)
DH5α	Plasmid propagation host	
STBL4	Plasmid propagation host	Fisher
STBL4(pGEM-T)	STBL4 with pGEM-T Easy	This study
STBL4(pRB01)	STBL4 with pGEM-T Easy vector containing EFSG_00659	This study
Plasmids		
pGEM-T Easy	Commercial plasmid for gene propagation	Promega
pLT06	Temperature-sensitive plasmid	(19)
pAT28	Shuttle vector for *E. faecalis*	(22)
pWH03	Used for gene insertion in the enterococci	(16)
pRB01	pGEM-T Easy Vector with EFSG_00659	This study
pHA102	pWH03 containing NotI-digested fragments of Efa733I M and S loci (EFSG_05028-9)	This study
pHA103	pWH03 containing NotI-digested fragments of Efa502I M and S loci (EFQG_01131-2)	This study
pWH16	pLT06 containing fragments from upstream and downstream of Efa502I	This study
pWH17	pLT06 containing fragments from upstream and downstream of Efa733I	This study

562 **Table 2. Distribution of predicted DNA MTases and R-M Systems*.**

Strain name		1,230,933	1,231,410	1,231,502	1,231,501	1,231,408	1,141,733	Com12	Com15
Clade		A1	A1	A1	A2	Hybrid A1/B	B	B	B
Type I Restriction Modification System	Specificity	EFPG_01270	EFTG_00783	EFQG_01131		EFUG_01527	EFSG_05028		EFWG_02518
	Modification	EFPG_01269	EFTG_00782	EFQG_01132		EFUG_01526	EFSG_05029		EFWG_02519
	Restriction	EFPG_01271	EFTG_00784	EFQG_01130		EFUG_01528	EFSG_05027		EFWG_02517
Type I Restriction Modification System	Specificity	EFPG_05435 and EFPG_05434	EFTG_02031						
	Modification	EFPG_05433	EFTG_02032						
	Restriction	EFPG_05436	EFTG_02030						
Type I Restriction Modification System	Specificity				EFRG_00106				
	Modification				EFRG_00107				
	Restriction				EFRG_00105				
Type II Methyltransferase		EFPG_01821	EFTG_02364; EFTG_02653	EFQG_01609, EFQG_02270			EFSG_00659		
Unspecified Methyltransferases			EFTG_02333		EFRG_02239	EFUG_01512		EFVG_00502	

563 *Loci that are in black are strain specific. Loci that are the same color are conserved in their protein sequences based on a >90% sequence
564 identity threshold. An empty cell indicates that the system was not detected.

565 **Table 3. SMRT Sequencing results.**

Strain	Motif (5' -> 3')	Modified Position	Type	% Motif Detected	# of Motifs Detected	# of Motifs in Genome	Mean Motif Coverage
Efm502	RAYCNNNNNNTTRG	2	m6A	98.8	905	916	53.7
	CYAANNNNNNGRTY	4	m6A	97.9	897	916	54.5
Efm733	AGAWNNNNATTA	1	m6A	78.5	278	354	22.3
	TAATNNNNWTCT	3	m6A	73.2	259	354	23.6
OG1RF:: *efa502I*	RAYCNNNNNNTTRG	2	m6A	99.9	795	796	162.8
	CYAANNNNNNGRTY	4	m6A	99.0	788	796	163.2
OG1SSp:: *efa733I*	AGAWNNNNATTA	1	m6A	99.0	323	326	180.4
	TAATNNNNWTCT	3	m6A	98.5	321	326	180.8

566

567 **Table 4. Methylation-sensitive REase digestion reaction results[1].**

REase	Recognition sequence	Recognition methylation	Strains Digested
McrBC	5'-$R^mC(N_{40}$-$N_{3000})R^mC$-3'	m5C, m4C	0/8
FspEI	5'-C^mC-3'	m5C	0/8
MspJI	5'-^{m}CNNR-3'	m5C	Efm733

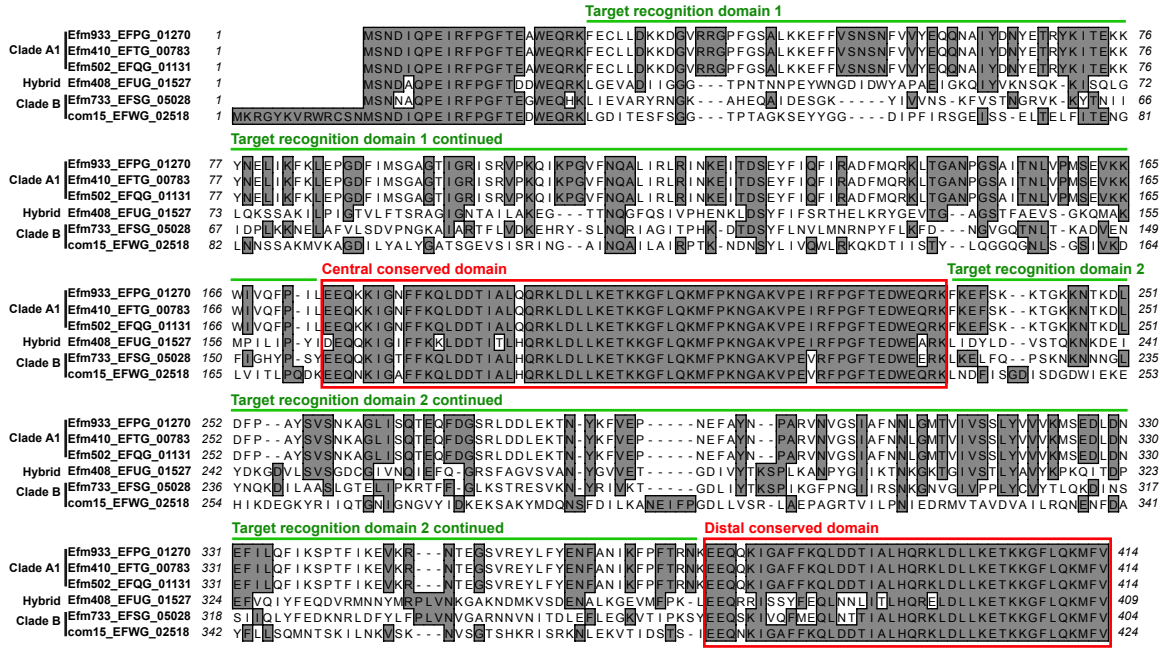
568 [1]Recognition sequences and methylation patterns were retrieved from NEB.

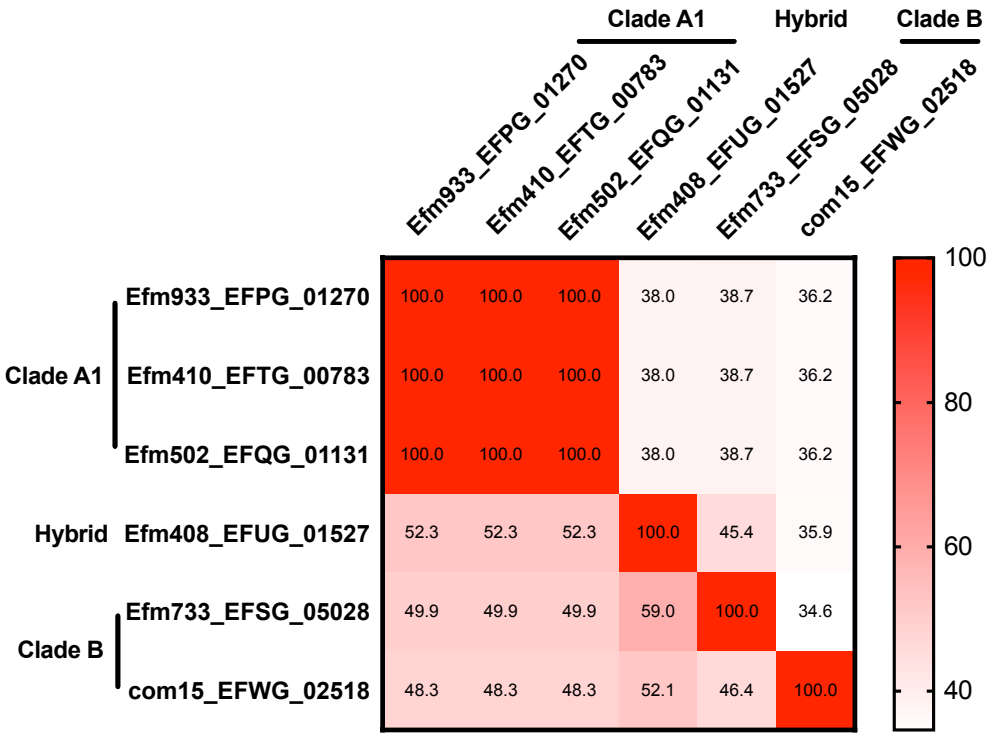
569 **Table 5. Bisulfite sequencing results.**

	methylation motif	methylation position	methylation type	# motifs detected[a]	# motifs in genome	# motif pass cvg filter	average methylation ratio[b]	mean motif coverage[c]
733	5'-RCCGGY-3'	2	m5C	748	780	750	96.5% (6.2%)	61.5X

570 [a]# motifs detected is defined as motifs with coverage depth more than 10X and methylation ratio >=0.35.
571 [b]Average methylation ratio is defined as the mean of the methylation ratio from all motifs in the genome.

a).





b).

Figure 1. Conservation and variability of S subunits. The protein sequences of predicted S subunits from 6 (out of 8) representative *E. faecium* genomes were pairwise

aligned using MacVector. The alignment is shown in (a). Central and distal conserved domain was identified based on sequence homology and labeled in red (30). The target recognition domains were interpreted based on conserved domains as labeled in green. The percent identity of each pair is shown in (b). White to red: low to high percent identities. Each number represents percent identity of one protein sequence (row name) to another (column name).

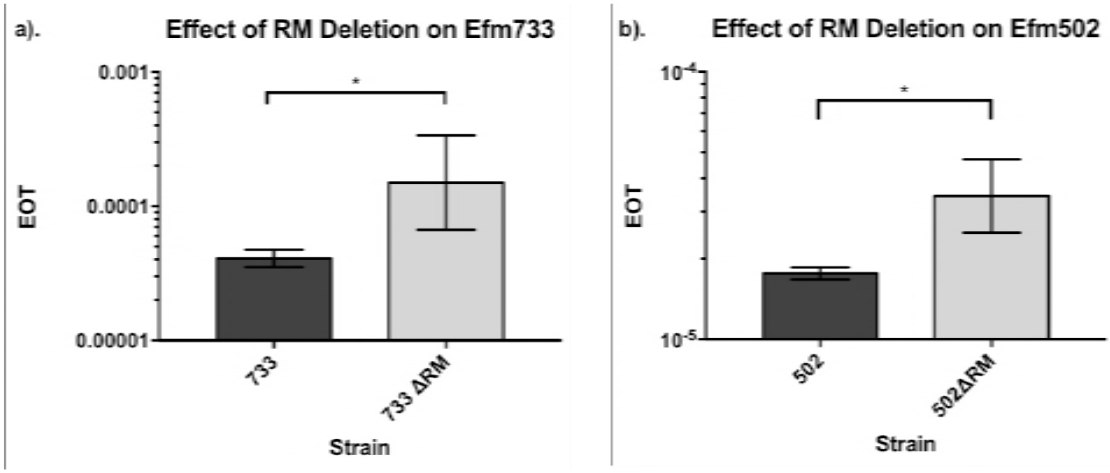


Figure 2. Type I R-M systems reduce transformability of Efm733 and Efm502. Three independent transformation experiments were performed. There is a statistical difference between the transformation efficiency of pAT28 into wild type and R-M null strains of Efm733 and Efm502. EOT: Efficiency of transfer. *: $p < 0.05$.

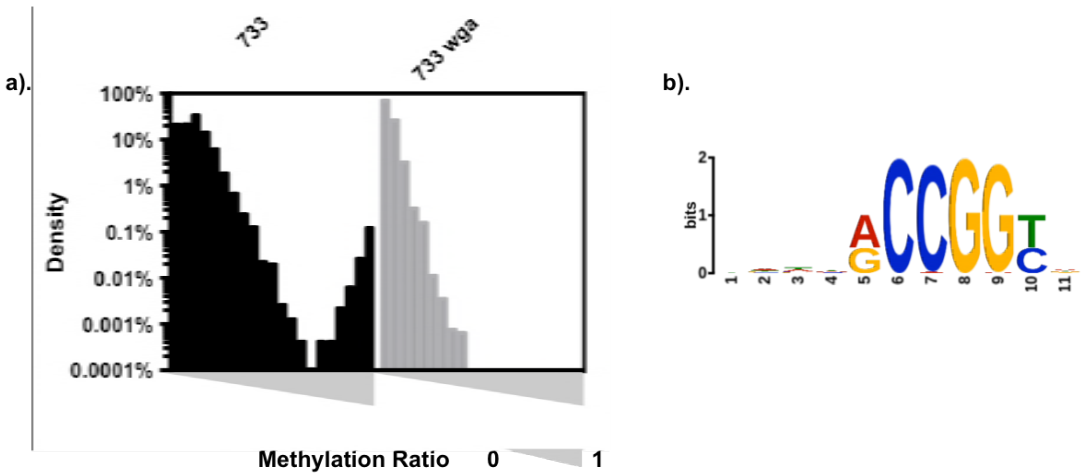


Figure 3. Bisulfite sequencing results for Efm733. a) Efm733 gDNA and its whole genome amplified (WGA) control were bisulfite-converted, and the methylation ratio for each cytosine site was calculated. Efm733 gDNA has cytosines with methylation ratios near 100%, indicating the presence of m5C methylation. b) All sites with ≥0.35 methylation ratio from native gDNA samples were extracted, together with 5 bp upstream and 5 bp downstream. The sequences are subjected for consensus motif analysis using MEME. The consensus motif identified by this analysis is shown.

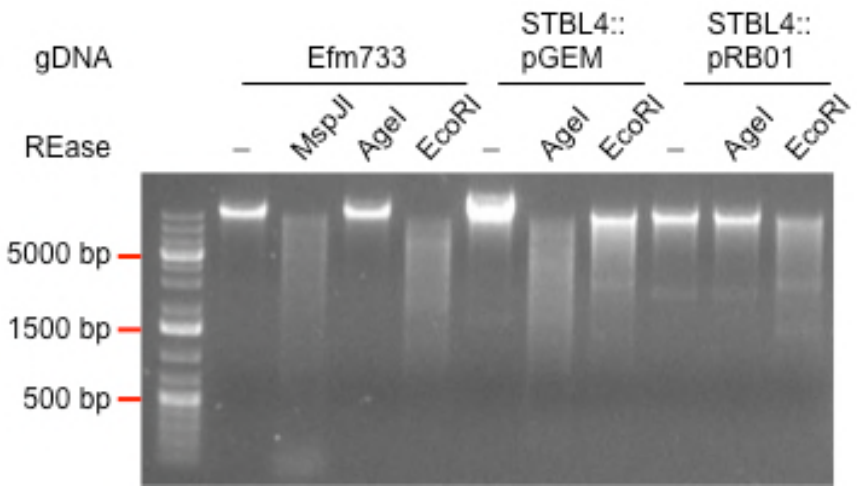


Figure 4. EFSG_00659 confers protection against AgeI digestion. AgeI digestion reactions were analyzed by agarose gel electrophoresis with ethidium bromide staining. Bacterial gDNA was used as substrate for REase reactions. Expression of the Efm733 gene EFSG_00659 in *E. coli* STBL4 protects *E. coli* gDNA from AgeI digestion. EcoRI is a positive control for digestion. -, no enzyme added.