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

Keywords: Restriction-Modification, *Enterococcus faecium*, genome defense

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

37 structures. R-M provides bacteria with a type of innate immunity to horizontal gene transfer
38 (HGT). We identified a Type I R-M system that is enriched in the hospital-epidemic clade and
39 determined that it is active for DNA modification activity and significantly impacts HGT. Overall,
40 this work is important because it provides a mechanism for the observed clade structure of *E.*
41 *faecium* as well as a mechanism for facilitated gene exchange among hospital-epidemic *E.*
42 *faecium*.
43

INTRODUCTION

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

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

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

70 immune system that utilizes sequence complementarity between self (CRISPR RNAs) and
71 foreign nucleic acid to carry out its restrictive function, whereas R-M discriminates self from
72 foreign DNA by DNA methylation patterns. If the *E. faecium* clades encode different defense
73 mechanisms, they may not exchange genetic information freely, thereby facilitating and
74 maintaining phylogenetic separation. However, little is known about CRISPR-Cas and R-M in *E.*
75 *faecium*. Genomic analysis suggests that these systems could contribute to the observed clade
76 structure of *E. faecium*. For example, CRISPR-Cas systems have been identified exclusively in
77 Clade B *E. faecium* and in sporadic Clade A-Clade B recombinant strains (8). For R-M, a
78 predicted methyl-directed restriction endonuclease (REase) is enriched in Clade A2 and B *E.*
79 *faecium* genomes relative to Clade A1 genomes (8).
80
81 Here, we focused on R-M systems and their roles in regulating gene exchange in *E. faecium*
82 because little is known about R-M defense in this species. Moreover, there is precedent in the
83 literature for R-M systems contributing to bacterial clade structure, as has been observed in
84 *Burkholderia* (9) and *Neisseria* (10). Our overarching hypothesis is that the *E. faecium* clades
85 encode different R-M systems, thereby inhibiting genetic exchange between them. In general,
86 R-M systems are composed of cognate methyltransferase (MTase) and REase activities and
87 are classified into different types based on the specific number and types of enzymes in the
88 system, as well as characteristics such as methylation type and pattern, cofactor requirement,
89 and restriction activity (11). A MTase recognizes specific sequences in the bacterial genome
90 and transfers a methyl group to either an adenine or a cytosine, resulting in 6-methyladenine
91 (m6A), 4-methylcytosine (m4C), or 5-methylcytosine (m5C). A REase may recognize the same
92 sequence as a MTase and cleave that region if the sequence is unmethylated (or in some
93 cases, if methylated). With the activities of MTases and REases, bacteria can use R-M to
94 impede entry of non-self DNA.
95

96 In this study, we used single-molecule real-time (SMRT) sequencing and whole genome
97 bisulfite sequencing to characterize the methylomes of representative *E. faecium* strains from
98 Clade A1 (*E. faecium* 1,231,502; or Efm502) and Clade B (*E. faecium* 1,141,733; or Efm733).
99 Two unique m6A methylation patterns were identified, one in each strain. These patterns were
100 asymmetric and bipartite, which is characteristic of Type I R-M methylation motifs (12).
101 Bioinformatic analyses were performed to identify candidate genes responsible for the
102 methylation. A unique Type I R-M system is encoded by each strain. We have named these
103 systems Efa502I (for Efm502) and Efa733I (for Efm733). Expression of these candidate
104 systems in *E. faecalis* heterologous hosts followed by SMRT sequencing confirmed that they
105 are responsible for the methylation patterns observed in Efm502 and Efm733. A functional
106 analysis was performed in order to assess the abilities of these systems to reduce *E. faecium*
107 HGT by transformation. In a comparative analysis among 73 *E. faecium* genomes, we found
108 that Efa502I is significantly enriched among Clade A1 isolates, while the Type I R-M system of
109 Efm733 appears to be strain-specific. Overall, this study is a first step towards understanding
110 the role of R-M in regulating HGT in *E. faecium* and the potential for R-M as one mechanism for
111 the clade structure of *E. faecium*.
112

113 **METHODS**

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

122 Taq polymerase (NEB) or Phusion (Fisher). Sanger sequencing to validate all genetic
123 constructs was performed at the Massachusetts General Hospital DNA Core facility (Boston,
124 MA).
125
126 **Isolation of genomic DNA.** Enterococcal strains were cultured overnight in BHI broth prior to
127 genomic DNA (gDNA) extraction. The extraction was performed using a Qiagen Blood and
128 Tissue DNeasy Kit using a previously published protocol (13). To isolate *E. coli* gDNA, bacteria
129 were grown overnight in LB broth prior to extraction using either the Blood and Tissue DNeasy
130 kit (Qiagen) or the UltraClean Microbial DNA Isolation Kit (Qiagen) per the manufacturer's
131 instructions. Whole genome amplification (WGA) control DNA was generated by amplification of
132 native gDNA using the REPLI-g kit (Qiagen) per the manufacturer's instructions.
133
134 **SMRT sequencing and methylome detection in *E. faecium*.** SMRT sequencing of *E. faecium*
135 gDNA and their WGA controls was performed by the Johns Hopkins Medical Institute Deep
136 Sequencing and Microarray Core. After sequencing, the reads were aligned to existing
137 references (for Efm502, NZ_GG688486-NZ_GG688546 and for Efm733, NZ_GG688461-
138 NZ_GG688485) and analyzed using the RS modification and motif detection protocol in SMRT
139 portal v1.3.3. WGA controls were used as methylation baselines.
140
141 **Bioinformatic analysis of R-M systems in eight *E. faecium* genomes.** The entire protein
142 complement for eight previously sequenced *E. faecium* isolates (14) was analyzed. To identify
143 potential MTases, the REBASE Gold Standard list (15) was used as a reference. This list is
144 comprised of biochemically verified MTases and REases. Each protein sequence from *E.*
145 *faecium* genomes was analyzed using BlastP against the REBASE Gold Standard list. The
146 protein sequences with significant (e-value $<1e^{-3}$) homology to REBASE Gold Standard proteins
147 were further filtered by protein size. If an *E. faecium* query protein length was less than half of

148 its subject's length, the match was removed from the prediction list. Due to the sequence
149 diversity of REases which complicates their bioinformatic identification (15), guilt-by-association
150 was used to identify full R-M systems as we have previously described (16). The proteins
151 encoded near candidate DNA MTases were analyzed using BLAST and Pfam for conserved
152 domains consistent with REase activities and/or sequence identity to confirmed REases. The
153 amino acid sequence of each R-M candidate was then pairwise compared among all the eight
154 strains to identify putative orthologs. If two protein sequences shared an amino acid identity
155 ≥90% with query coverage ≥90%, they were considered to be orthologous.
156
157 **Expression of R-M systems in *E. faecalis* heterologous hosts.** Genes encoding the
158 specificity and methylation subunits of Efa733I (EFSG_05028-EFSG_05027) were PCR-
159 amplified in their entirety, including the upstream region to retain the native promoter, using
160 primers 733_T1A_SM_F and 733_T1A_SM_R (see Table 2 for primer sequences). The PCR
161 product was digested with *Not*I and ligated into *Not*I-digested pWH03 (16) using T4 DNA Ligase
162 (NEB), generating pHA102. pWH03 is a pLT06 derivative for expression of genes from a
163 previously validated neutral genomic insertion site (EF2238-EF2239) for expression (GISE) (16,
164 17). pHA102 constructs were then introduced into *E. coli* DH5α via heat shock for propagation
165 and sequence confirmation. pHA102 was electroporated into *E. faecalis* OG1SSp using a
166 previously described method (18). An *E. faecalis* OG1SSp derivative with a chromosomal
167 integration of Efa733I, referred to as OG1SSp::*efa733I*, was generated by temperature shifts
168 and *p*-chlorophenylalanine counterselection, as previously described (19).
169
170 Genes encoding the specificity and methylation subunits of Efa502I (EFQG_01131-
171 EFQG_01132) were PCR-amplified using primers 502_T1A_SM_F and 502_T1A_SM_R. The
172 PCR product was then TA-cloned into the pGEM-T Easy Vector (Promega) and introduced into
173 DH5α via heat shock to generate pGEM-SMA1. pGEM-SMA1 was then digested with *Not*I, and

174 the digestion reaction was used as insert for ligation into *Not*I-digested pWH03. The ligation
175 reaction was then introduced into DH5α, and colonies were screened for chloramphenicol
176 resistance and ampicillin susceptibility to ensure the pGEM backbone was not ligated into
177 pWH03. Once the construct, referred to as pHA103, was confirmed via Sanger sequencing, it
178 was introduced into electrocompetent *E. faecalis* OG1RF. An *E. faecalis* OG1RF derivative with
179 a chromosomal integration of Efa502I, referred to as OG1RF::*efa502I*, was generated by
180 temperature shifts and *p*-chlorophenylalanine counterselection. All plasmids and strains for
181 heterologous expression were validated by PCR and Sanger sequencing.
182
183 SMRT sequencing in *E. faecalis* OG1 derivatives expressing Efa502I or Efa733I was performed
184 by the University of Michigan sequencing core facility. Reads were mapped to the *E. faecalis*
185 OG1RF reference sequence (GenBank accession number NC_017316), and the methylation
186 motifs were detected using the RS modification and motif detection protocol in SMRT portal
187 v.2.3.2. *In silico* controls were used as modification baselines.
188
189 **Generation of *E. faecium* R-M deletion mutants.** Regions up- and downstream of Efa502I
190 and Efa733I were PCR-amplified using primers listed in Table 2, ligated into pLT06, and
191 transformed into *E. coli* EC1000 (20), generating pWH16 and pWH17 (Table 1). Insert
192 sequences were confirmed using Sanger sequencing. *E. faecium* strains were made
193 electrocompetent using previously a published protocol (21). 2 µg of sequence-confirmed
194 plasmids were electroporated into electrocompetent Efm733 and Efm502. The generation of
195 deletion mutants was accomplished using temperature shifts and *p*-chlorophenylalanine
196 counterselection, as previously described (19). The successful deletion mutants were
197 sequence-confirmed by PCR and Sanger sequencing.
198

199 **Transformation efficiency test.** Efm733, Efm502, and their respective R-M deletion mutants
200 were made electrocompetent using a modified version of the previously published protocol (21).
201 Briefly, overnight cultures were diluted 10-fold in BHI and cultured to OD_{600nm} ~ 0.6. The bacteria
202 were then pelleted and treated with filter-sterilized lysozyme buffer (10 mM Tris-HCl pH 8.0, 10
203 mM EDTA pH 8.0, 50 mM NaCl) supplemented with 83 µL of 2.5 KU/mL mutanolysin stock for
204 30 min at 37°C. The cells were then pelleted and washed three times with ice-cold filter-
205 sterilized electroporation buffer (0.5 M sucrose and 10% glycerol). Finally, the cells were
206 pelleted and resuspended in electroporation buffer and aliquoted for storage at -80°C and future
207 use. 1 µg pAT28 (22) was electroporated into the electrocompetent *E. faecium* cells. The counts
208 of total viable cells and spectinomycin-resistant cells were determined by serial dilution and
209 plating. The transformation efficiency was expressed as percent of transformed (spectinomycin-
210 resistant) cells per total viable cells. Three independent experiments were performed and the
211 statistical significance was assessed using the unpaired one-tailed Student's t-test.
212
213 **Distribution analysis of putative R-M systems and orphan MTases.** The amino acid
214 sequences for select R-M system and orphan MTase candidates were queried against a
215 collection of 73 *E. faecium* isolates previously analyzed by Lebreton *et al* (8) using BLASTP.
216 Any proteins which shared >90% query coverage and amino acid identity were considered
217 orthologs. The Fisher's exact test was used to determine if an orphan MTase or R-M system
218 was significantly over- or under-represented in a particular clade.
219
220 **REase protection assays.** To identify m5C methylation, gDNA was treated with the
221 methylation-sensitive REases McrBC, FspEI, and MspJI (NEB). 500 ng gDNA was incubated
222 with each REase at 37°C for 3 h (McrBC) or 6 h (FspEI and MspJI) followed by analysis by
223 electrophoresis on a 1% agarose gel with ethidium bromide.
224

225 **Bisulfite sequencing.** Whole-genome bisulfite sequencing libraries were constructed using the
226 Illumina TruSeq LT PCR FREE kit and the Qiagen EpiTect Bisulfite kit. Native DNA was isolated
227 as described above. Whole-genome-amplified (WGA) control DNA was generated by
228 amplification of native gDNA using the Qiagen REPLI-g® kit, per the manufacturer's
229 instructions. For bisulfite sequencing, briefly, 2 µg each of native and WGA control DNA were
230 fragmented using NEB fragmentase. DNA fragments ranging from 200 bp to 700 bp were gel
231 extracted and end-repaired. After A-tailing of DNA fragments, Illumina TruSeq adapters were
232 ligated. Then, the bisulfite conversion was performed using the Qiagen EpiTect Bisulfite kit, per
233 the manufacturer's instruction. An 8-cycle PCR enrichment with Illumina primer mix was
234 performed, followed by size selection and gel purification. The libraries were sequenced using
235 Illumina MiSeq with 2x75 bp paired-end chemistry.
236
237 **Whole genome bisulfite sequencing analysis.** The sequencing reads were analyzed using
238 Bismark (23) with additional quality control and filtering as described previously (16). Briefly, the
239 Illumina reads were mapped to the *in silico* bisulfite-converted references (23). Then, we
240 quantified the conversion rate of each mapped read by calculating the percentage of converted
241 C (which will result in T) to the total number of C in the reference within the mapped region. The
242 mapped reads with ≤80% conversion rate were filtered out from analysis (16). Next, the
243 coverage depth and methylation ratio were calculated for each C site. The methylation ratio was
244 calculated by dividing the total number of C by the coverage depth at each C site. A fully
245 methylated C, thus protected from bisulfite conversion, will have a methylation ratio near 1. An
246 unmethylated C will have a methylation ratio near 0. To identify consensus methylation motifs, C
247 sites with ≥0.35 methylation ratio and ≥10 coverage depth, along with the sequences of 5 bp
248 upstream and 5 bp downstream, were extracted. The extracted sequences were subjected for
249 MEME motif search (24).
250

251 **Confirmation of m5C MTase activity**. Primers EFSG_00659_F and EFSG_00659_R (Table 2)
252 were used to amplify the entire Efm733 EFSG_00659 coding region and its upstream predicted
253 promoter. The PCR product was then cloned into the pGEM-T Easy Vector (Promega) per the
254 manufacturer's instructions and transformed into *E. coli* STBL4 (Fisher) to generate pRB01.
255 REase digestion assays with methylation-sensitive enzymes were performed on purified *E. coli*
256 and *E. faecium* gDNA as described above.
257
258 **Accession numbers**. DNA sequence data generated in this study have been deposited in the
259 Sequence Read Archive under accession numbers PRJNA397049 (for SMRT sequencing data)
260 and PRJNA488088 (for Illumina bisulfite sequencing data).
261
262 **RESULTS**
263 **Identification of Clade A1-specific putative Type I R-M system in *E. faecium***. We previously
264 reported that a Type II R-M system significantly reduces HGT via conjugation (18) and
265 transformation (16) in *E. faecalis*. Here, we hypothesize that the *E. faecium* clades encode
266 distinct R-M systems that reduce the exchange of genetic information between them. We
267 utilized an approach we previously developed for *E. faecalis* R-M analysis (18) to predict
268 potential R-M systems in eight previously sequenced *E. faecium* genomes. The 8 genomes
269 included 3 genomes from Clade A1, 3 genomes from Clade B, one genome from Clade A2, and
270 one recombinant Clade A1/B hybrid (5, 8). Because REases are difficult to identify with
271 bioinformatics, and MTase prediction is comparatively straightforward, as has been previously
272 reported by NEB (15), we first identified predicted DNA MTases in *E. faecium* genomes, and
273 then analyzed surrounding genes for predicted R-M-related activities. The complete list of
274 candidates for the eight strains is shown in Table 3.
275

276 Interestingly, we predicted at least one putative Type I R-M system for seven of the eight *E.*
277 *faecium* strains (Table 3). Type I R-M systems are multisubunit complexes comprised of a
278 specificity subunit (S), a methylation subunit (M), and a restriction subunit (R) (11, 25-27). The S
279 subunit is responsible for the specific DNA recognition motif and associates with the DNA to
280 bring the M and R subunits into contact. The system has two conformations: M_2S_1, which is
281 capable of methylating DNA based on the recognition sequence, and $R_2M_2S_1$, which is capable
282 of restricting DNA (27, 28). One predicted *E. faecium* Type I R-M system is comprised of highly
283 conserved (>90% amino acid sequence identity) M and R subunits in six of eight genomes
284 across both Clade A1 and Clade B (Table 3 and Dataset S1). The specificity subunit from this
285 system, however, is highly conserved in Clade A1 genomes but not in Clade B (Table 3; Fig. 1
286 and Fig S1). S subunits possess two target recognition domains (TRDs) that determine the
287 nucleotide sequence the subunit binds to (29, 30). The variation in amino acid sequence
288 between the S subunits occurs within these TRDs (Fig S1), suggesting that these S subunits
289 recognize different DNA sequences. Notably, the S subunits from Clade A1 strains are identical
290 to each other, indicating that they utilize the same recognition sequence.
291
292 To examine the distribution of this putative Clade A1-specific system in a larger collection of *E.*
293 *faecium* strains, we analyzed 73 *E. faecium* genomes of mostly draft status that were reported
294 previously (8). This list includes 15 clade B isolates, 21 clade A1 isolates, 35 clade A2 isolates,
295 and 2 hybrid isolates (Table S1). We selected Efm502 (Clade A1) as our representative Clade
296 A1 strain for this analysis and used Type I R-M sequences from this genome as references for
297 analysis against the broader collection of *E. faecium* strains. The M and R subunits of the
298 putative Clade A1-specific Type I system were detected in 51 and 52, respectively, of 73 *E.*
299 *faecium* genomes, including both Clade A and Clade B strains (Fig S2a). However, the
300 distribution of the S subunits varied (Fig S2a-b). The S subunit present in Efm502,
301 EFQG_01131, was significantly enriched within Clade A1 isolates (14/21; p-value <0.0001 using

302 Fisher exact test; Fig S2a) and absent from all other clades with the exception of strain
303 EnGen002, which is classified as a Clade A1/B hybrid strain. Interestingly, the S subunits
304 present in most other *E. faecium* strains are strain-specific by the strict thresholds applied here.
305 Given that the Efm502 S subunit is enriched in Clade A1, we hypothesize that many Clade A1
306 strains exchange genetic information freely with each other while exchange with other *E.*
307 *faecium* strains is restricted.
308
309 **SMRT sequencing for *E. faecium* methylome analysis.** We analyzed the Efm502 and
310 Efm733 genomes by SMRT sequencing. SMRT sequencing measures the kinetics of DNA
311 polymerase as it synthesizes DNA in order to identify bases that have been modified (31-33). It
312 has been extensively utilized for bacterial methylome analysis (34-42). With SMRT sequencing,
313 6-methyladenine (m6A) and 4-methylcytosine (m4C) can be easily detected with modest
314 sequence coverage (~25x per strand), while 5-methylcytosine (m5C) detections requires high
315 coverage (~250x per strand) (41, 43). Using SMRT sequencing, we identified two unique m6A
316 methylation motifs in Efm502 and Efm733. Efm502 possessed m6A methylation at the
317 underlined position of the motif 5'-RAYCNNNNNNTTRG-3' (and its complementary strand 5'-
318 CYAANNNNNNGRTY-3') and Efm733 possessed m6A methylation at the underlined position of
319 the motif 5'-AGAWNNNNATTA-3' (and its complementary strand 5'-TAATNNNNWTCT-3')
320 (Table 4; Dataset S2). These sequences are asymmetric and bipartite, which is characteristic of
321 Type I R-M methylation (12). Due to the coverage of our SMRT sequencing, m5C modification
322 could not be accurately detected. The two unique m6A methylation patterns indicate that DNA
323 from one strain would be recognized as foreign should it cross the strain barrier.
324
325 **Expression in heterologous hosts links methylation activity to genes in Efm502 and**
326 **Efm733.** According to our predictions (Table 3), there is only one complete Type I R-M system
327 encoded by each of Efm502 and Efm733. To determine if these systems are responsible for the

328 methylation patterns identified by SMRT sequencing, we expressed the respective S and M
329 subunits (EFQG_01131-01132 for Efm502 and EFSG_05028-05027 for Efm733) in the
330 heterologous host *E. faecalis* OG1RF or its spectinomycin/streptomycin-resistant relative
331 OG1SSp. Previous work in our lab had characterized the methylome of OG1RF using SMRT
332 and bisulfite sequencing (16). This allowed us to attribute any new methylation patterns
333 observed during SMRT sequencing to the *E. faecium* genes that were expressed in the OG1RF
334 background. SMRT sequencing of these strains detected the same methylation patterns
335 originally identified in Efm502 and Efm733 (Table 4; Dataset S2). These data demonstrate that
336 EFQG_01131-01132 is responsible for the 5'-RAYCNNNNNNTTRG-3' methylation in Efm502
337 and that EFSG_05028-05027 is responsible for the 5'-AGAWNNNNATTA-3' methylation in
338 Efm733. Because we have confirmed the function of these genes, we have named them
339 Efa502I and Efa733I, which is consistent with the R-M system nomenclature convention
340 established by New England Biolabs (12).
341
342 **Efa502I and Efa733I reduce transformation efficiency in *E. faecium*.** To determine whether
343 the Type I R-M systems in Efm733 and Efm502 actively defend against exogenous DNA, we
344 constructed null strains (Efm733ΔRM and Efm502ΔRM; Table 1) and evaluated their
345 transformation efficiencies relative to their wild-type parent strains. Here, we utilized the broad
346 host range plasmid pAT28 (Table 1) (22). pAT28 sequence has motifs recognized by the Type I
347 R-M systems in both wild-type Efm733 (1 occurrence) and Efm502 (1 occurrence). The
348 transformation of pAT28 into Efm733 and Efm502 served as a baseline for the experiment. If
349 Efa733I and Efa502I are active, we expect to see higher pAT28 transformation efficiencies into
350 Efm733ΔRM and Efm502ΔRM, respectively. Indeed, we observed significantly higher
351 transformation efficiencies into Type I R-M system null strains (Fig 2; p-value<0.05 using one-
352 tailed Student's t-test). These data demonstrate that the Type I R-M systems in Efm733 and
353 Efm502 actively function as mechanisms of genome defense.

354

355 **m5C methylation occurs in Efm733.** As described previously, our SMRT sequencing had
356 insufficient coverage depth for m5C methylome characterization. Hence, we used REase
357 protection assays with commercially available methylation-sensitive REases to query the
358 presence of m5C methylation in our eight *E. faecium* strains. Table 5 summarizes the
359 recognition sequences and modifications of the enzymes used in this study. Only Efm733
360 showed evidence of cytosine modification, as it was digested by MspJI (Table 5; Fig S3).

361

362 To determine the exact cytosine methylation motif present in Efm733, gDNA was subjected to
363 whole-genome bisulfite sequencing. Whole genome amplified (WGA) DNA was used as
364 negative control since WGA removes all modifications. During bisulfite treatment, cytosine
365 bases are converted to thymine unless they are protected by either m4C or m5C methylation.
366 Additionally, our lab has previously published a method of distinguishing between m4C and
367 m5C methylation using thymine conversion ratios of sequencing reads after bisulfite treatment
368 (44). m5C methylation is sufficient to protect the cytosine residue completely from bisulfite
369 conversion, so that most sequencing reads at that position contain the original cytosine base.
370 However, m4C methylation provides only partial protection from bisulfite conversion, so a
371 thymine conversion rate of 0.5 at a particular position within the sequencing reads suggests the
372 presence of m4C methylation. Bisulfite conversion and subsequent sequencing revealed that
373 Efm733 possesses m5C modification at the motif 5'-R^{m}<u>C</u>CGGY-3' (Table 6, Fig 3, and Fig S4;
374 the methylation occurs at the underlined position), which overlaps the MspJI recognition site (5'-
375 CNNR-3') and hence supports the evidence of methylation obtained from the MspJI digestion
376 assays. Based on the cytosine conversion ratio of close to 1.0, m5C modification is supported,
377 which is consistent with why it was not detected by our SMRT sequencing.

378

379 **EFSG_00659 is responsible for m5C methylation in Efm733.** Based on the bioinformatics
380 analyses, we hypothesized that EFSG_00659 was responsible for the m5C methylation found in
381 Efm733 (Table 3). EFSG_00659 possessed no homologs in the other seven strains analyzed,
382 making it a good candidate for the unique methylation found in Efm733. We queried the
383 EFSG_00659 protein sequence against the REBASE gold standard list and identified that it has
384 high sequence similarity to M.AvaIX, M.VchO395I and M.VchAI (e-value ≤ $3e^{-125}$; recognition
385 sites are 5'-RCCGGY-3'). Interestingly, BLASTP identified no significant hits when
386 EFSG_00659 was queried against the larger collection of 73 *E. faecium* genomes, indicating its
387 unique presence in Efm733.
388
389 In order to link EFSG_00659 with the m5C methylation identified during bisulfite sequencing, we
390 expressed it in the heterologous host *E. coli* STBL4 and performed an REase protection assay.
391 The REase AgeI recognizes the motif 5'-A<u>C</u>CGGT-3', and its enzymatic activity is blocked if
392 m5C methylation is present at the underlined position. This motif overlaps the m5C methylation
393 motif in Efm733 identified by bisulfite sequencing. If the motif is methylated, DNA will be
394 protected against digestion. We cloned EFSG_00659 into the vector pGEM-T and transformed it
395 into *E. coli* STBL4, generating strain *E. coli* STBL4(pRB01). Genomic DNA from Efm733,
396 STBL4(pGEM-T), and STBL4(pRB01) was treated with AgeI per the manufacturer's instructions.
397 EcoRI was used as a positive control for digestion. Figure 4 shows representative results of the
398 digestions on a 1% agarose gel. As expected, Efm733 was digested by EcoRI and protected
399 against digestion from AgeI. STBL4(pGEM-T) was digested by both EcoRI and AgeI, indicating
400 that the original host and empty vector pGEM-T did not possess the appropriate m5C
401 methylation. STBL4(pRB01) was protected against digestion by AgeI, demonstrating that
402 EFSG_00659 is responsible for the 5'-R^{m}CCGGY-3' methylation found in Efm733.
403
404 **DISCUSSION**

405 In this study, we used a combination of genomic and genetic approaches to identify a functional
406 Type I R-M system that is enriched in Clade A1 *E. faecium* and that significantly alters
407 transformability of a model Clade A1 strain. We propose that this R-M system impacts HGT
408 rates among *E. faecium* mixed-clade communities, thereby helping to maintain the observed
409 phylogenetic structure of *E. faecium* and facilitating HGT specifically among Clade A1 strains.
410 Mixed communities of *E. faecium* clades are expected to occur in environments where healthy
411 and ill human hosts, human and animal hosts, and/or the feces of any of these hosts co-mingle
412 (i.e. in sewage). In future studies, we plan to assess the impact of R-M on conjugative plasmid
413 transfer, which is a major mode of HGT in enterococci and was not assessed in our current
414 study.
415
416 An interesting observation from our study is the sequence diversity of Type I S subunits
417 encoded within *E. faecium* Type R-M systems having nearly identical R and M subunits (Fig
418 S2a-b). After further investigation into those alignments, we found that those S subunits sharing
419 50-70% overall amino acid sequence identities possess sequence diversity within one TRD
420 domain, where the other TRD domain and the central conserved domain are conserved. This
421 suggests that these systems share partial recognition sequences. Previous research has
422 reported that the diversification of Type I R-M recognition sequences is driven by TRD
423 exchanges, permutation of the dimerization domain, and circular permutation of TRDs (45). Our
424 observation suggests that TRD recombination and reorganization events occur for *E. faecium*
425 Type I R-M systems outside Clade A1. Future studies will use genomics to further explore the
426 relationship between S subunit sequence diversity and its impact on *E. faecium* methylomes
427 and inter-strain and inter-clade HGT.
428
429 **Acknowledgments**

430 This work was supported by Public Health Service grants K22AI099088 and R01AI116610 to
431 K.L.P.
432

433 **References**

434 1. Noble CJ. 1978. Carriage of group D streptococci in the human bowel. J Clin Pathol
435 31:1182-6.

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

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

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

444 5. Palmer KL, Godfrey P, Griggs A, Kos VN, Zucker J, Desjardins C, Cerqueira G, Gevers
445 D, Walker S, Wortman J, Feldgarden M, Haas B, Birren B, Gilmore MS. 2012.
446 Comparative genomics of enterococci: variation in *Enterococcus faecalis*, clade structure
447 in *E. faecium*, and defining characteristics of *E. gallinarum* and *E. casseliflavus*. MBio
448 3:e00318-11.

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

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

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

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

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

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

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

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

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

488 15. Roberts RJ, Vincze T, Posfai J, Macelis D. 2015. REBASE—a database for DNA
489 restriction and modification: enzymes, genes and genomes. Nucleic Acids Research
490 43:D298-D299.

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

494 17. DebRoy S, van der Hoeven R, Singh KV, Gao P, Harvey BR, Murray BE, Garsin DA.
495 2012. Development of a genomic site for gene integration and expression in
496 *Enterococcus faecalis*. Journal of Microbiological Methods 90:1-8.

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

500 19. Thurlow LR, Thomas VC, Hancock LE. 2009. Capsular polysaccharide production in
501 *Enterococcus faecalis* and contribution of CpsF to capsule serospecificity. J Bacteriol
502 191:6203-10.

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

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

508 22. Trieu-Cuot P, Carlier C, Poyart-Salmeron C, Courvalin P. 1990. A pair of mobilizable
509 shuttle vectors conferring resistance to spectinomycin for molecular cloning in
510 *Escherichia coli* and in Gram-positive bacteria. Nucleic Acids Research 18:4296-4296.

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

513 24. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble
514 WS. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res
515 37:W202-8.

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

518 26. Bertani G, Weigle JJ. 1953. Host controlled variation in bacterial viruses. J Bacteriol
519 65:113-21.

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

522 28. Taylor I, Patel J, Firman K, Kneale G. 1992. Purification and biochemical
523 characterisation of the EcoR124 type I modification methylase. Nucleic Acids Res
524 20:179-86.

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

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

529 31. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base
530 identification for single-molecule nanopore DNA sequencing. Nature Nanotechnology
531 4:265-270.

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

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

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

545 35. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B,
546 Mahajan MC, Jabado OJ, Deikus G, Clark TA, Luong K, Murray IA, Davis BM, Keren-
547 Paz A, Chess A, Roberts RJ, Korlach J, Turner SW, Kumar V, Waldor MK, Schadt EE.
548 2012. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia*
549 *coli* using single-molecule real-time sequencing. Nature Biotechnology 30:1232-1239.

550 36. Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, Fomenkov A, Turner
551 SW, Korlach J, Roberts RJ. 2012. The methylomes of six bacteria. Nucleic Acids Res
552 40:11450-62.

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

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

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

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

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

568 42. Krebes J, Morgan RD, Bunk B, Spröer C, Luong K, Parusel R, Anton BP, König C,
569 Josenhans C, Overmann J, Roberts RJ, Korlach J, Suerbaum S. 2014. The complex
570 methylome of the human gastric pathogen *Helicobacter pylori*. Nucleic Acids Research
571 42:2415-2432.

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

575 44. Huo W. 2017. PhD thesis. The University of Texas at Dallas, Richardson, TX.
576 *Enterococcus faecalis* Genome Defense Systems and Their Impact on Conjugative
577 Antibiotic Resistance Plasmid Transfer. ProQuest Dissertations Publishing: 10970358.
578 Available through link: http://hdl.handle.net/10735.1/5776.

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

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

583

584

585 **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

586

587

588 **Table 2. Primers used in this study.**

Primer name	Primer sequence[a]
502SMR_A1F_XbaI	CTAGTCTAGACCCTTGTTCGATATAGACCC
502SMR_A1R_BamHI	CGCGGATCCCCTAATATGAAAGCAATTATCAAC
502SMR_A2F_BamHI	CGCGGATCCGAGATTATCTGCCGCACTAAAT
502SMR_A2R_XmaI	TCCCCCCGGGGTAGAAACATACACAGCTATAC
733SMR_A1F_XbaI	CTAGTCTAGAGGACGTAAACCTTCGACA
733SMR_A1R_BamHI	CGCGGATCCATGTTAGATGATGAGTTGATTCC
733SMR_A2F_BamHI	CGCGGATCCGCTTACTGTGTCATATTCCTC
733SMR_A2R_XmaI	TCCCCCCGGGCTCTACGAACTTGATTGAATCC
733_T1A_SM_F	TACGATGCGGCCGCGGGAAGAACCATTTACACAA
733_T1A_SM_R	ACATTGGCGGCCGCCGAGCATCATTTATTAACATGTTT
502_T1A_SM_F	TTGCATGCGGCCGCCGGATTTGGAGCCCGACT
502_T1A_SM_R	TACGATGCGGCCGCCCCATAGGTAAAGTGCCAC
EFSG_00659_F	CGTTACCTGCAGGCAGCGAAAACTAGTTGGG
EFSG_00659_R	GAATCAGGATCCGCCTCTAAAGAATAAAGATCC

589 [a]Underlined sequences are restriction enzyme recognition sites.

590 **Table 3. 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	

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

593 **Table 4. SMRT Sequencing results.**

Strain	Motif[a]	Type	% Motif Detected	# of Motifs Detected	# of Motifs in Genome	Mean Motif Coverage
Efm502	5'-R^{m}AYCNNNNNNTTRG-3'	m6A	98.8	905	916	53.7
	3'-YTRGNNNNNN^{m}AAYC-5'	m6A	97.9	897	916	54.5
Efm733	5'-^{m}AGAWNNNNATTA-3'	m6A	78.5	278	354	22.3
	3'-TCTWNNNNT^{m}AAT-5'	m6A	73.2	259	354	23.6
OG1RF:: *efa502I*	5'-R^{m}AYCNNNNNNTTRG-3'	m6A	99.9	795	796	162.8
	3'-YTRGNNNNNN^{m}AAYC-5'	m6A	99.0	788	796	163.2
OG1SSp:: *efa733I*	5'-^{m}AGAWNNNNATTA-3'	m6A	99.0	323	326	180.4
	3'-TCTWNNNNT^{m}AAT-5'	m6A	98.5	321	326	180.8

594 [a]The underlined base indicates modified base.

595 **Table 5. 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

596 [1]Recognition sequences and methylation patterns were retrieved from NEB.

597 **Table 6. 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

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

600 **FIGURE LEGENDS**

601 **Figure 1. Conservation and variability of S subunits.** The protein sequences of predicted S
602 subunits from 6 (out of 8) representative *E. faecium* genomes were pairwise aligned using
603 MacVector. The percent identity of each pair is shown. White to red: low to high percent
604 identities. Each number represents percent identity of one protein sequence (row name) to
605 another (column name).

606

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

611

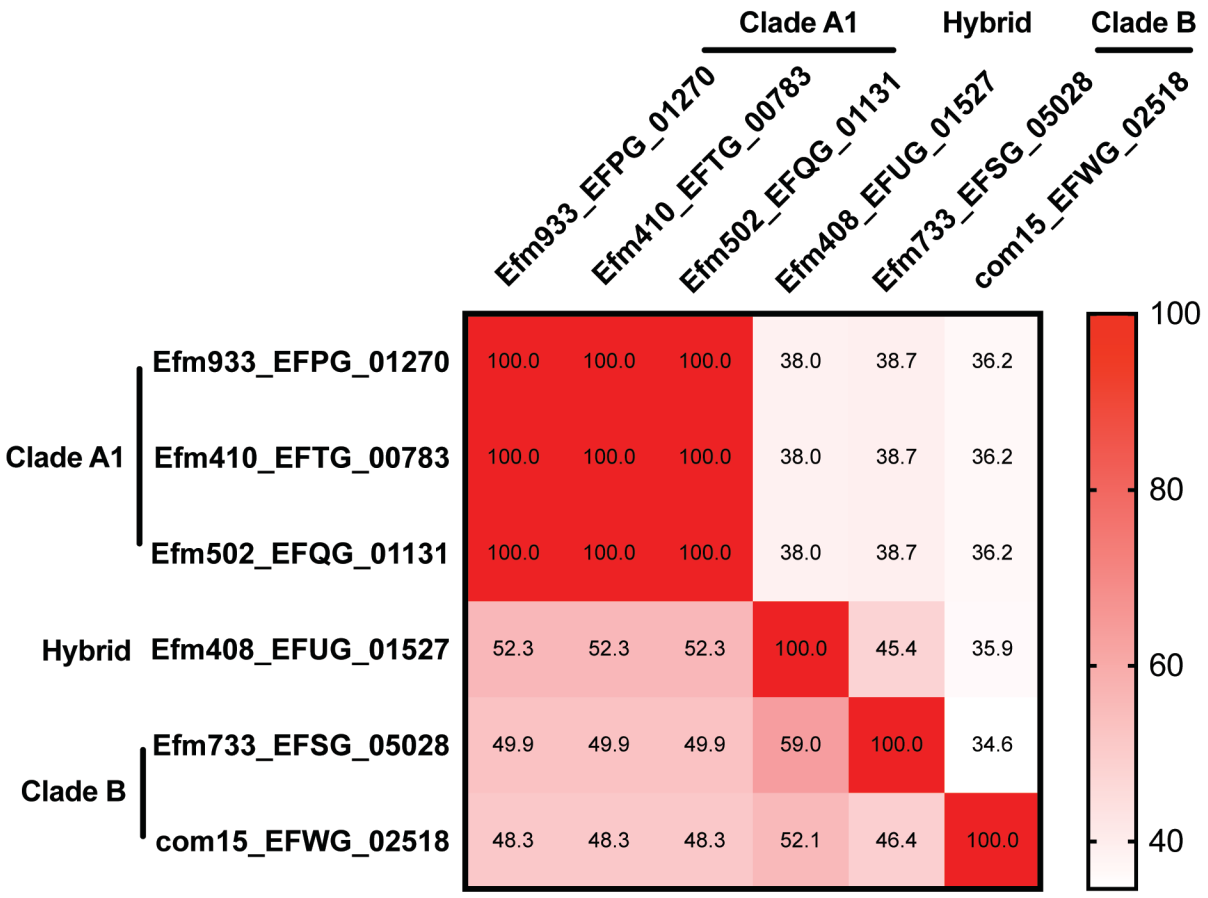
612 **Figure 3. Bisulfite sequencing results for Efm733.** Efm733 gDNA and its whole genome
613 amplified (WGA) control DNA were bisulfite-converted and deep sequenced. The methylation
614 ratio for each cytosine site was calculated. A). The methylation ratio was plotted against the
615 density of cytosine sites with that methylation ratio. The presence of cytosine sites with
616 methylation ratio near 100% in Efm733 native gDNA but not WGA control indicates the
617 presence of m5C methylation. All cytosine sites with ≥0.35 methylation ratio as indicated with
618 red boxes from native gDNA samples were extracted, together with 5 bp upstream and 5 bp
619 downstream sequences. The sequences are subjected to consensus motif analysis using
620 MEME. The consensus motif identified by this analysis is shown in (b) and the center position
621 (position 6) indicates where the modification was detected.

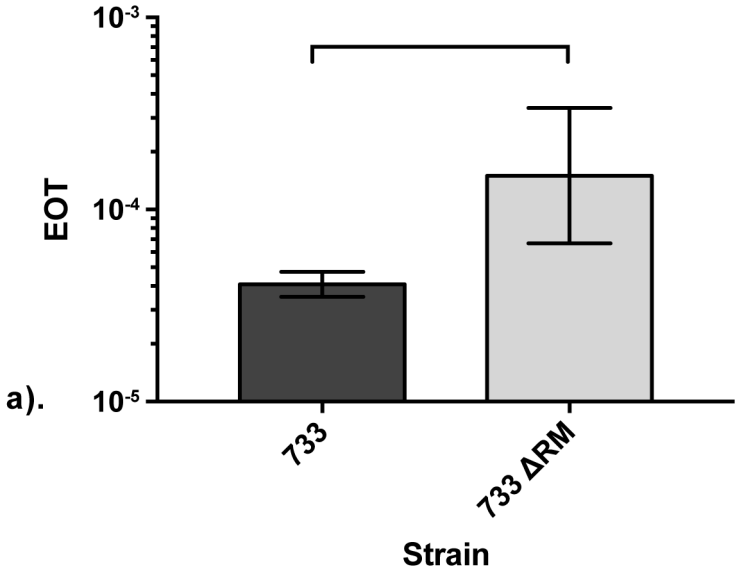
622

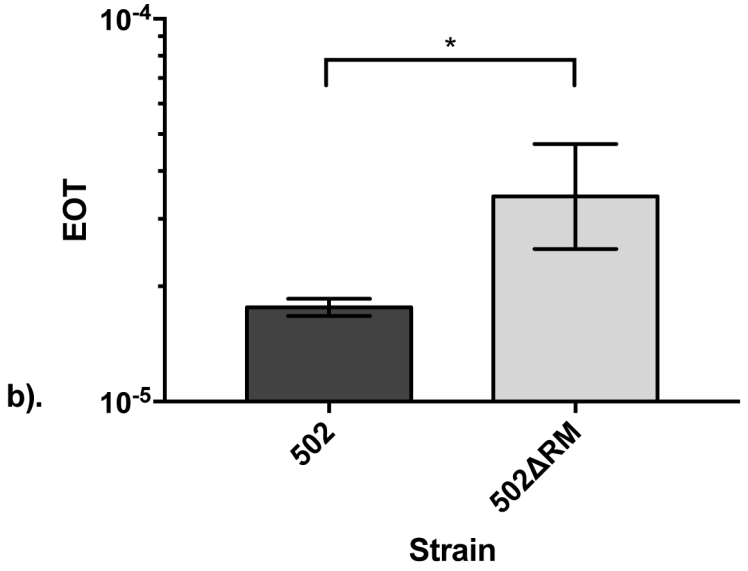
623 **Figure 4. EFSG_00659 confers protection against AgeI digestion.** AgeI digestion reactions
624 were analyzed by agarose gel electrophoresis with ethidium bromide staining. Bacterial gDNA
625 was used as substrate for REase reactions. Expression of the Efm733 gene EFSG_00659 in *E.*

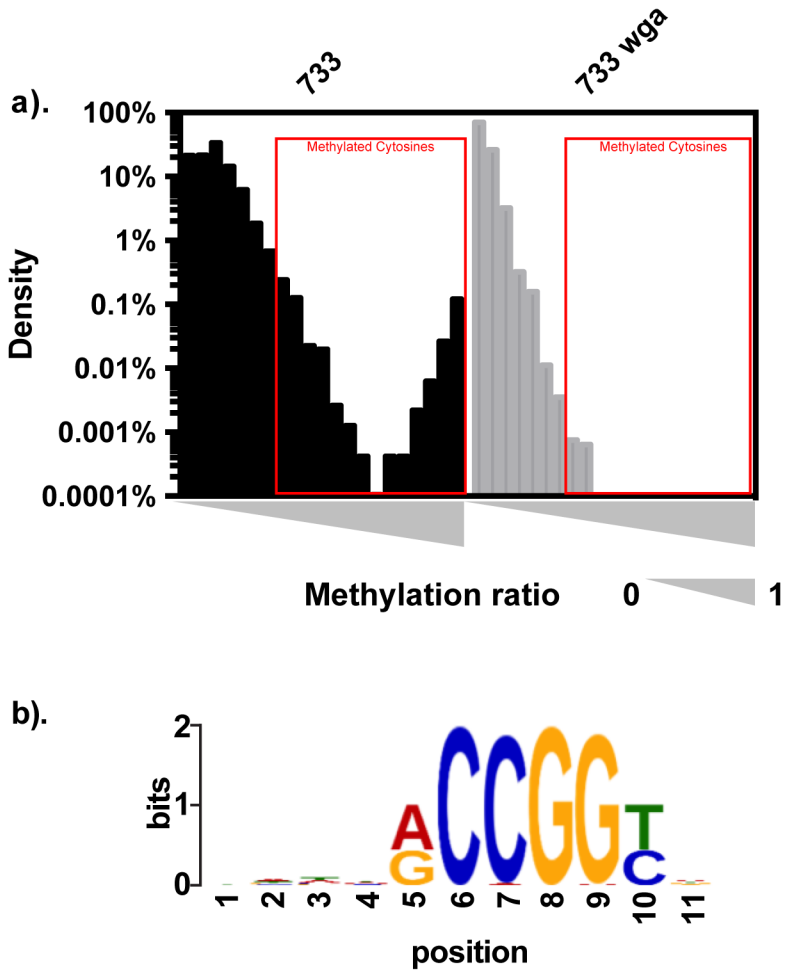
626 *coli* STBL4 protects *E. coli* gDNA from AgeI digestion. EcoRI is a positive control for digestion. -,

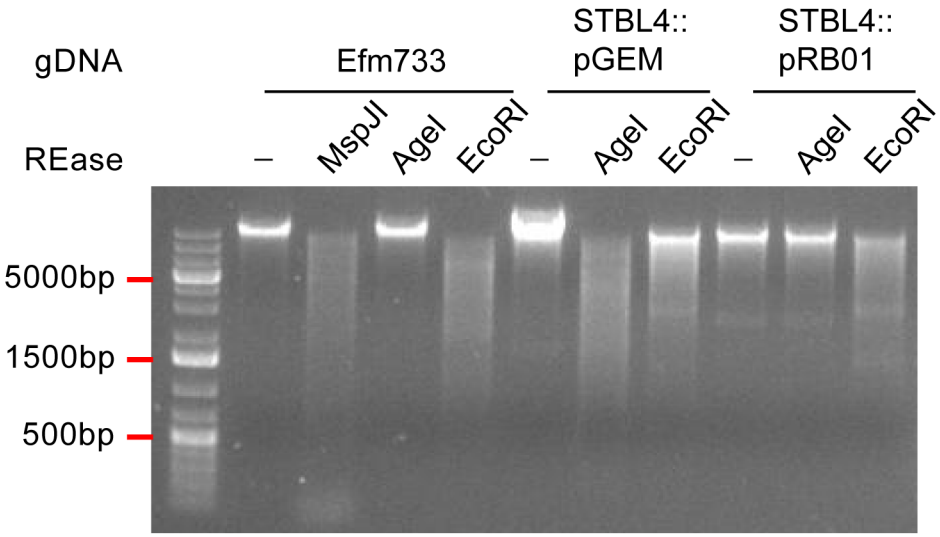
627 no enzyme added.


Clade A1
Hybrid
Clade B
Efm933_EFPG_01270
Efm410_EFTG_00783
Efm502_EFQG_01131
Efm408_EFUG_01527
Efm733_EFSG_05028
com15_EFWG_02518
Clade A1
Hybrid
Clade B
100.0 100.0 100.0 38.0 38.7 36.2
100.0 100.0 100.0 38.0 38.7 36.2
100.0 100.0 100.0 38.0 38.7 36.2
52.3 52.3 52.3 100.0 45.4 35.9
49.9 49.9 49.9 59.0 100.0 34.6
48.3 48.3 48.3 52.1 46.4 100.0
100
80
60
40


a).
EOT
10^{-3}
10^{-4}
10^{-5}
733
733 ΔRM
Strain


b).
EOT
10^{-4}
10^{-5}
*
502
502ΔRM
Strain


a).
733
733 wga
Density
100%
10%
1%
0.1%
0.01%
0.001%
0.0001%
Methylated Cytosines
Methylated Cytosines
Methylation ratio
0
1
b).
bits
2
1
0
1
2
3
4
5
6
7
8
9
10
11
position


gDNA
Efm733
STBL4::
pGEM
STBL4::
pRB01
REase
–
MspJI
AgeI
EcoRI
–
AgeI
EcoRI
–
AgeI
EcoRI
5000bp
1500bp
500bp