# Homologous recombination between tandem paralogues drives evolution of a subset of Type VII secretion system immunity genes in firmicute bacteria

Stephen R. Garrett[1]*, Giuseppina Mariano[1], Jo Dicks[2] and Tracy Palmer[1]*

[1] Microbes in Health and Disease Theme, Newcastle University Biosciences Institute, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK.

[2] Culture Collections, UK Health Security Agency, London, UK.

*To whom correspondence should be addressed.

e-mail: s.garrett2@newcastle.ac.uk, tracy.palmer@newcastle.ac.uk

Tel +44 191 208 3219

**Key Words:** T7SS, homologous recombination, immunity gene families, *Staphylococcus aureus, Streptococcus*

**Repositories:** The sequence data for *Staphylococcus aureus strain* RN6390 can be found on NCBI under BioProject: PRJNA789916

## ABSTRACT

The Type VII secretion system (T7SS) is found in many Gram-positive firmicutes and secretes protein toxins that mediate bacterial antagonism. Two T7SS toxins have been identified in *Staphylococcus aureus*, EsaD a nuclease toxin that is counteracted by the EsaG immunity protein, and TspA, which has membrane depolarising activity and is neutralised by TsaI. Both toxins are polymorphic, and strings of non-identical *esaG* and *tsaI* immunity genes are encoded in all *S. aureus* strains. To investigate the evolution of *esaG* repertoires, we analysed the sequences of the tandem *esaG* genes and their encoded proteins. We identified three blocks of high sequence similarity shared by all *esaG* genes and identified evidence of extensive recombination events between *esaG* paralogues facilitated through these conserved sequence blocks. Recombination between these blocks accounts for loss and expansion of *esaG* genes in *S. aureus* genomes and we identified evidence of such events during evolution of strains in clonal complex 8. TipC, an immunity protein for the TelC lipid II phosphatase toxin secreted by the streptococcal T7SS, is also encoded by multiple gene paralogues. Two blocks of high sequence similarity locate to the 5' and 3' end of *tipC* genes, and we found strong evidence for recombination between *tipC* paralogues encoded by *Streptococcus mitis* BCC08. By contrast, we found only a single homology block across *tsaI* genes, and little evidence for intergenic recombination within this gene family. We conclude that homologous recombination is one of the drivers for the evolution of T7SS immunity gene clusters.


## DATA SUMMARY

All sequence data for strains used in this study are available on NCBI under BioProject PRJNA789916. Sequences from the NCTC3000 project are available on NCBI under BioProject PRJEB6403. Supplementary data 2 and all custom scripts are available on Github: https://github.com/GM110Z/Garret-et-al.-recombination-paper.

48 **IMPACT STATEMENT**

49 The type VII secretion system (T7SS) in firmicutes secretes polymorphic toxins that target
50 other bacteria. To protect from the action of these toxins, bacteria carry multiple paralogous
51 copies of immunity protein-encoding genes that are sequence-related but non-identical. To
52 date, little is known about how T7 immunity gene families evolve. In this study we analysed a
53 cluster of EsaG-encoding genes in *Staphylococcus aureus* which are found at the *ess/T7*
54 secretion locus and provide immunity against the T7 secreted nuclease toxin, EsaD. We
55 identified three homology blocks covering *esaG* genes and their downstream intergenic
56 regions, which are separated by two variable regions. We have shown that recombination can
57 occur between these homology blocks, leading to loss or expansion of *esaG* genes at this
58 locus. Using a historical dataset of closely related *S. aureus* strains from clonal complex 8, we
59 identified several independent recombination events leading to changes in the *esaG*
60 repertoire. We further showed that similar events are observed for an immunity protein
61 encoded by Group B *Streptococcus* spp. suggesting that recombination plays a broader role
62 in the evolution of T7SS immunity-encoding genes. We speculate that gain and loss of T7
63 immunity genes is weighed in response to environmental pressure and metabolic burden.

64 **INTRODUCTION**

65 The type VII protein secretion system (T7SS) is found in many Gram-positive bacteria.
66 Following its discovery in pathogenic Mycobacteria, it has since been described in a range of
67 other actinobacteria and in firmicutes (1-5). Cryo-electron microscopy studies have shown that
68 the ESX-5 T7SS from Mycobacteria exists as a 2.3 MDa membrane complex, with a central
69 ATPase, EccC, forming a hexameric pore (6, 7). The firmicutes T7SS is only distantly related
70 to the actinobacterial T7SSa system and has been termed T7SSb. The hexameric ATPase is
71 the only common component found across all T7SSs and is designated EssC in the T7SSb
72 systems (8, 9).

73

74 While the T7SSa is heavily linked with Mycobacterial virulence, there is growing evidence that
75 the T7SSb plays an important role in bacterial antagonism (10, 11). In *Streptococcus*
76 *intermedius*, three T7SSb-secreted antibacterial effectors have been identified, including TelB,
77 an NADase, and TelC a lipid II phosphatase (12). The *Enterococcus faecalis* T7SSb also
78 mediates contact-dependent inhibition of some firmicute bacteria, and a bioinformatic analysis
79 in *Listeria monocytogenes* has identified over 40 potential antibacterial substrates of the T7SS
80 (13, 14).

81

82 A role for the T7SSb in interbacterial competition was first characterised in the opportunistic
83 pathogen *Staphylococcus aureus* (15). The T7SS-encoding locus in this organism is highly
84 variable. Sequence divergence initiates towards the 3' end of *essC*, with *essC* sequences
85 falling into one of four variants, termed *ess1* – *ess4* (16). Downstream of each *essC* subtype
86 is a cluster of variant-specific genes. In *essC1* variant strains one of these genes, *esaD*,
87 encodes a secreted nuclease toxin with antibacterial activity. Protection from the toxic activity
88 of EsaD is mediated by EsaG, which is encoded immediately adjacent to *esaD* in *essC1*
89 strains. EsaG inactivates EsaD by forming a tight complex with the EsaD nuclease domain
90 (15). While all *essC1* strains encode *esaD*, the toxin nuclease domain is polymorphic, and
91 these strains also encode additional copies of *esaG* genes in a highly variable immunity gene

92 island located at the 3' end of the *ess*/*T7SS* locus (10, 15, 16). These additional *esaG* copies
93 are genetically diverse but share some core regions of similarity within the encoded amino
94 acid sequences (17, 18). Strings of up to ten non-identical *esaG* genes are also found in a
95 similar genomic location in *essC2*, *essC3* and *essC4* strains (10 and our Supplementary
96 dataset 1).
97
98 TspA is a second antibacterial toxin secreted by the *S. aureus* T7SS. TspA has a C-terminal
99 membrane-depolarising domain, and immunity from intoxication is provided by the membrane-
100 bound TsaI protein (19). The *tspA-tsaI* locus is encoded distantly from the T7 gene cluster and
101 is found across all four *essC* variant strains. Similar to EsaD, the TspA toxin domain is
102 polymorphic, and all strains encode clusters of TsaI variants (between two and sixteen copies)
103 directly downstream of *tspA* (19).
104
105 It is common for bacteria to encode repertoires of immunity proteins for protection against
106 polymorphic effector proteins. For example, bacteria in the gut accumulate immunity genes in
107 genomic islands that provide protection against type VI secretion system effectors (20). This
108 includes orphan immunity genes in strains that do not encode the cognate effector protein, as
109 we observe for *esaG* genes in *S. aureus essC2*, *essC3* and *essC4* strains that do not contain
110 *esaD*. While it is common for immunity islands to carry many predicted immunity genes, it is
111 less common to see so many homologues of the same immunity gene clustered together. At
112 present, little is known about the origin of the T7SS immunity repertoires.
113
114 In this study we used gene phylogeny and cluster analyses to identify the processes that drive
115 the accumulation of T7SS immunity genes. Our findings indicate that intergenic recombination
116 is a major factor in the expansion and deletion of *esaG* genes, and that this is at play during
117 the evolution of *S. aureus* strains. We also noted that a similar process drives the evolution of
118 *tipC* genes, which encode Streptococcal TelC immunity proteins (12). By contrast, there is

119 limited evidence of intergenic recombination within *tsaI* clusters, and it does not appear to be

120 the main mechanism of evolution for this gene family.

121 **METHODS**

122

123 **Strains and genome sequencing**

124 Strain RN6390 was obtained from Professor Jan Maarten van Dijl (University of Groningen,
125 NL). Enhanced genome sequencing was carried out by MicrobesNG (Birmingham, UK).
126 Sequence data for RN6390 can be found under BioProject: PRJNA789916.

127

128 **Gene and protein alignment**

129 Nucleotide sequences were obtained from NCBI and aligned with MAFFT v7.489 (21). Amino
130 acid sequences were obtained from NCBI and aligned using MUSCLE v3.8.1551 (22). To
131 construct similarity plots for both nucleotide and amino acid sequences, Plotcon
132 (https://www.bioinformatics.nl/cgi-bin/emboss/plotcon) was executed using aligned
133 sequences, with a window size of 5. For *esaG4,* the full pseudogene was used for the
134 nucleotide alignment. For the alignment of the EsaG4 amino acid sequence, the two predicted
135 open reading frames (ORFs) were used. For Plotcon analysis on large input sequences,
136 alignments were manually curated to remove partial sequences and pseudogenes.

137

138 **Recombination Prediction**

139 Aligned nucleotide sequences were opened in the RDP4 software (23) and Run All selected.
140 TrimAl v1.2 (24) was used to remove the unaligned 5' region of *tsaI* genes before RDP4
141 analysis was carried out.

142

143 **Gene phylogeny**

144 Maximum likelihood trees for nucleotide sequences were built with IQTREE v 2.1.4 (25), with
145 1000 ultrafast bootstraps. Trees were visualised and annotated using iTOL (26).

146

147 **Comparison of genetic loci and gene cluster analysis**

148 T7SS-encoding loci from each *essC* variant were subjected to pairwise comparisons with the
149 RN6390 *T7SS* locus using BLAST (27). The report produced by BLAST was used as a
150 comparison file in genoPlotR package (28) within RStudio (29) to plot regions of similarity
151 between the two loci. FlaGs (30) was used to assess the variation in number of *tsaI* repeats
152 across *S. aureus* strains. Examples were selected to represent diversity at this locus. A bash
153 script was written to extract *ess/T7* loci from 133 *Staphylococcus aureus* strains whole-
154 genome sequenced within the NCTC3000 project (PRJEB6403). Nucleotide profile Hidden
155 Markov Models for genes *essC*, *esaD*, *esaG* (generated using all variants found within
156 NCTC8325) and *focA* were used to search with HMMER (31) for the presence and copy
157 number of key genes and to extract locus co-ordinates. In particular, the presence of *esaD*
158 was used to identify *essC1* variant strains and the number of *esaG* copies was recorded.
159 Custom bash scripts were used to extract *esaG* and *DUF5079* genes from the *ess/T7* loci of
160 NCTC CC8 strains and alignments carried out as described. Easyfig was used to analyse the
161 *ess/T7* clusters of NCTC CC8 strains by pairwise comparison and subsequently indicate
162 regions of high sequence similarity between each. These were ordered based on when they
163 were accessioned within the NCTC culture collection.
164
165

166 **RESULTS**

167

168 **RN6390 encodes twelve *esaG* copies.**

169 Much of our previous work on the *S. aureus* T7SS has used strain RN6390, which is a
170 derivative of the parental strain NCTC8325. Two genome sequences are available for
171 NCTC8325, NCTC8325-Oklahoma (GenBank accession number CP000253) and
172 NCTC8325-Sanger (GenBank accession number LS483365). The sequences are closely
173 related but differ primarily by the number of *esaG* genes at the T7 locus, with six in the
174 Oklahoma sequence and 12 in the Sanger sequence. To determine the number of *esaG*
175 copies present in RN6390, we undertook whole genome sequencing (deposited under
176 GenBank accession number CP90001.1), finding 12 *esaG* genes, of identical sequence to
177 those in the NCTC8325-Sanger sequence. We noted that the DNA sequence is highly
178 repetitive in this region, which may have confounded sequence assembly for the Oklahoma
179 accession. Nonetheless our sequence analysis shows that our model strain, RN6390,
180 harbours 12 copies of *esaG*.

181

182 **Intergenic recombination is a driving force in the evolution of *esaG* genes.**

183 As the sequence at the *esaG* locus of RN6390 is repetitive, we examined this region of the
184 chromosome in more detail. Of the 12 homologues of *esaG*, encoded in this region, the first
185 copy is found directly downstream of *esaD* and encodes the cognate immunity protein for the
186 nuclease toxin (Fig 1a; 15). We have named this *esaG1*. This is followed by three genes
187 encoding hypothetical proteins (SAOUHSC_00270 _00271 and _00272), before a stretch of
188 four further *esaG* homologues (*esaG2* - *esaG5*). This is followed by a further homologue of
189 *SAOUHSC_00270* (annotated as *SAOUHSC_00270b* on Fig 1a, sharing 96.49% identity with
190 *SAOUHSC_00270*) and then an additional seven *esaG* homologues (*esaG6* – *esaG12*; Fig
191 1a). Note that while *esaG4* is annotated as a pseudogene, due to a short additional stretch of
192 nucleotides close to the centre of the gene introducing a premature stop and subsequent start
193 codon, it actually encodes two smaller ORFs (Fig 1a).

194

195 To assess variation between the 12 homologues of EsaG, the amino acid sequences were
196 aligned (Fig 1b), and regions of high sequence similarity between the proteins were analysed
197 using Plotcon (Fig 2a). Numerous regions of sequence similarity were observed, including two
198 major blocks of high sequence homology at amino acids 13-31 and 84-119. To assess whether
199 these were general features of *S. aureus* EsaG proteins, we downloaded approximately 4,000
200 *S. aureus* EsaG sequences from RefSeq and collectively analysed them using Plotcon. Fig
201 S1a indicates that a similar profile of sequence similarity is seen across all EsaG proteins.
202

203 We next examined the intergenic regions between the RN6390 *esaG* homologues. Strikingly,
204 we noted that most of them were of a very similar length, other than when they directly
205 preceded a non-*esaG* gene (for example *SAOUHSC_00270b*). They also share a high degree
206 of sequence similarity (Fig S2a). When we undertook Plotcon analysis on the *esaG* genes,
207 including the 3' intergenic regions (Fig 2a, Fig S2b), we noted the same two major homology
208 blocks that we had seen from the amino acid analysis, but in addition a third block
209 encompassing the end of the gene and the downstream intergenic region (Fig 2a, Fig S2b,
210 Fig S2c).
211

212 Given the substantial levels of sequence similarity between the *esaG* genes, we used the
213 recombination prediction software, RDP4, to determine whether there had been recombination
214 events between the genes. The RDP4 output, shown in Table S1 and summarised Fig 2b,
215 predicts with high significance that there have been extensive recombination events within
216 most (but not all) of these genes. As shown in Fig 2b, recombination appears to occur at the
217 three points within the genes that correspond to the regions of high nucleotide sequence
218 homology. To support these findings, we constructed a maximum-likelihood tree for each of
219 the variable regions of *esaG1-12* (Fig 2ci-iii). The *esaG* genes do not cluster in the same way
220 in each variable region, confirming that these genes have not evolved solely by divergent
221 evolution. For example, *esaG7* clusters with *esaG3* in variable region 1, *esaG8* in variable

222 region 2 and *esaG10* in variable region 3 (Fig 2ciii). This suggests extensive recombination
223 has occurred between these genes to generate this repertoire.
224
225 Based on the RDP4 results, we built a schematic representation showing the homologous
226 regions of the *esaG* genes that are likely involved in the recombination events (Fig 2d).
227 Specific regions of the genes seem to share high sequence similarity with others, for example,
228 the mid-section of many of the homologues share high similarity to equivalent sections of
229 *esaG1* (coloured teal). Conversely, genes such as *esaG3* and *esaG6* appear to be much more
230 diverse.
231
232 **Evidence of *esaG* recombination in the USA300 epidemic strains of *S. aureus.***
233 To determine whether recombination could be detected between *esaG* genes across *S.*
234 *aureus* strains, we first examined genome sequences from a recent study analysing the
235 community spread and evolution of a USA300 variant during a New York outbreak (32).
236 USA300 is a methicillin-resistant *S. aureus essC1* strain, and a dominant cause of community-
237 acquired *S. aureus* infection in the USA (33). Comparing the whole-genome sequences of the
238 epidemic lineage showed that these strains carry only seven *esaG* genes in comparison to
239 the ten copies in the closely related USA300 FPR3757. Using strain BKV_2 as a
240 representative of the outbreak lineage, we aligned the region spanning from *esaE* to
241 *SAUSA300_0303* with USA300 FPR3757. The alignment showed almost complete sequence
242 identity, other than in a region spanning from the middle of *SAUSA300_0295* to the middle of
243 *SAUSA300_0299* (Fig 3a) which was absent from BKV_2. When the nucleotide sequences of
244 *SAUSA300_0295* and *SAUSA300_0299* were aligned with *esaG4* of BKV_2, *esaG4* was seen
245 to be mosaic, with most of the gene being identical to *SAUSA300_0295* but the 3' end showing
246 100% identity to *SAUSA300_0299* (Fig 3b). This is consistent with recombination between
247 homology block 2 of *SAUSA300_0295* and *SAUSA300_0299*, with loss of the intervening
248 DNA.
249

250 **Analysis of *esaG* recombination within a historical culture collection.**

251 To obtain further evidence of *esaG* recombination events, we accessed genome sequences
252 from the National Collection of Type Cultures (NCTC) which has recently generated long read
253 assemblies from 133 *S. aureus* strains sampled over the past 100 years. From this resource
254 we initially constructed a dataset of all the genome sequences (including NCTC8325 which is
255 part of this collection). The coordinates of the *ess/T7* locus were identified in each strain using
256 a custom bash script and the locus extracted using the coordinates for DNA immediately 5' of
257 *essC* and 3' of *focA* (which defines the 3' boundary of the T7 immunity island (10);
258 Supplementary Data 1). Next the clonal complex of each strain was annotated based on a
259 core genome SNP tree constructed previously (34), because for the purpose of our analysis it
260 was essential to analyse very closely related strains to ensure that any differences could not
261 be due to divergent evolution. From the annotated tree we selected to analyse the *ess/T7*
262 locus from the 31 CC8 strains as there are a significant number of strains in this clonal complex
263 (including NCTC8325) and all strains are extremely closely related to one another (34). Using
264 Easyfig to assess the genetic differences at this locus (35), we observed that the most
265 common genetic composition was to encode 12 *esaG* paralogues, as seen for NCTC8325 and
266 RN6390 (Fig S3). However, there were clear differences in the *esaG* repertoire among some
267 of these closely related strains, including two unique expansion events, five unique deletion
268 events and two instances of recombination occurring in other regions of the locus (Fig S3).
269 These were confirmed by long read sequence mapping (Supplementary Data 2).

270

271 Two unique expansion events were observed resulting in an additional *esaG7*-like gene in
272 strain NCTC10724 and a duplication of *esaG11* in NCTC10702. Further analysis of the events
273 in NCTC10724 indicate that recombination has occurred at homology block 1 of *esaG7* with
274 homology block 1 of *esaG8*, most likely during DNA replication (Fig 4a). The duplication of
275 *esaG11* in NCTC10702 is most likely due to recombination at homology block 3 between the
276 3' of *esaG11* and the 3' of *esaG10* (Fig 4b). These observations confirm that expansion of

277 *esaG* repertoires can occur due to recombination between the homology blocks we have
278 identified.
279
280 Five unique deletion events were also identified among the strains. Deletions due to
281 recombination at homology block 3 were responsible for three of these (Fig 4c-e).
282 Recombination between homology block 1 of *esaG9* and *esaG12* resulted in a deletion event
283 in NCTC13140, generating a novel *esaG* mosaic (Fig 4f). In NCTC13196, a deletion event has
284 occurred to produce a novel *esaG* mosaic comprising the 5' part of *esaG7* and 3' end of
285 *esaG12*. Notably in this instance, recombination in these genes has not occurred within the
286 homology blocks, but instead, in variable region 2 of the gene (Fig 4g). This is possible
287 because *esaG7* and *esaG12* share high sequence similarity in this region (Fig S2c, Fig 2d).
288
289 In RN6390 and other CC8 strains carrying the full repertoire of 12 *esaG* homologs, two
290 *DU5079* genes are present, one encoded immediately downstream of *esaG1* and the other
291 downstream of *esaG5* – these genes are annotated *DUF5079-1 (SAOUHSC_00270)* and
292 *DUF5079-2 (SAOUHSC_00270b)*, respectively (Fig 5). In ten of the CC8 strains analysed, we
293 noted that only a single *DUF5079* gene was encoded, and that these strains carried only
294 *esaG1* along with *esaG6* – *esaG12*. Analysis reveals that rather than a recombination event
295 between homology block 3 of *esaG1* and *esaG5*, recombination was instead highly likely to
296 have occurred between the *DU5079* regions. In strain NCTC10702, recombination occurred
297 between *DUF5079-1* and *DUF5079-2* to produce a recombinant *DUF5079* gene (Fig 5a),
298 whereas in strain NCTC12232 only the *DUF5079-2* gene was found, suggesting
299 recombination had occurred between the intergenic regions upstream of these genes (Fig 5b).
300 Aligning the region spanning *esaG1* and *DUF5079-1* with *esaG5* and *DUF5079-2* from
301 RN6390, we noted that the intergenic region and *DUF5079* gene share much higher sequence
302 similarity than most of the *esaG* genes (Fig 5c). To explore this further, these genomic regions
303 were extracted from the CC8 collection and an alignment visualised using Plotcon, confirming
304 that across these strains, high sequence similarity is observed within the intergenic regions

305 and *DUF5079* genes (Fig 5d). This indicates that recombination in other regions of the *ess/T7*
306 locus also plays a role in dictating the repertoire of T7SS immunity genes carried at this locus.
307
308 **_esaG_ diversity across _S. aureus essC2_, _essC3_ and _essC4_ variants.**
309 Although *S. aureus essC2*, *essC3* and *essC4* variants do not encode EsaD, these strains all
310 accumulate *esaG* genes at the 3' end of their *ess/T7* loci. To ascertain whether homologous
311 recombination may similarly play a role in the evolution of these *esaG* families we analysed
312 the *esaG* genes encoded in strains ST398, MRSA252 and HO 5096 0412, as representatives
313 of *essC2*, *essC3* and *essC4* variants, respectively (Fig S4). We noted that none of the *esaG*
314 genes found in these strains appear to be direct copies of the 12 *esaG* genes we identified in
315 the CC8 *essC1* strains (Fig S4 a-d). Using RDP4 analysis to detect recombination events,
316 again we could detect clear signatures of recombination in almost all of the genes (Fig S4 e-
317 g; Table S2-4), with the exception of *SAPIG_0314-0315* pseudogene. We conclude that
318 recombination events are common among *esaG* genes in all *S. aureus* strains analysed.
319
320 **No clear evidence for recombination within _tsaI_ genes.**
321 TspA is a T7SS-secreted antibacterial toxin that is highly conserved across all *essC* variant
322 strains (19). In all strains it is encoded away from the T7SS gene cluster, at a genomic location
323 bounded by *SAOUHSC_00583* and *iolS* (*SAOUHSC_00603*). The toxic activity of TspA is
324 neutralised by TsaI, a membrane protein of the DUF443 family (19). Multiple copies of *tsaI*
325 genes are encoded downstream of *tspA* (Fig S5a,b), and in RN6390 there are 11 copies, *tsaI1*
326 – *tsaI11* (Fig S5a). A small pseudogene, encoded by *SAOUHSC_00600* shares high
327 sequence similarity to part of the toxin region of TspA and is also found at this locus.
328
329 To determine whether recombination events also contributed to the evolution of the TsaI
330 repertoires, we aligned the amino acid sequences of RN6390 TsaI proteins (Fig S5c). We
331 noted much greater sequence variability between these proteins than the EsaG homologues,
332 particularly in the C-terminal region of the protein. Similar variability was observed in a

333 representative alignment of around 3000 TsaI sequences (Fig S1b). At the nucleotide
334 sequence level, much greater variability was also observed in the *tsaI* intergenic regions, in
335 both length and DNA sequence compared with *esaG* (Fig S5d). Using Plotcon we identified
336 only a single region of high similarity among TsaI sequences, covering approximately the first
337 75 amino acids, which is also mirrored at the DNA level (Fig S6a). As only one block of high
338 sequence similarity is detected and there is a high degree of sequence variability in the *tsaI*
339 intergenic regions, recombination within individual genes is unlikely. To analyse this, we used
340 RDP4 to predict recombination events within the 11 *tsaI* genes (Fig S6b). Far fewer potential
341 recombination events were predicted than for *esaG* genes, and with much lower probability
342 (Table S5). We conclude that evolutionary processes other than tandem homologous
343 recombination give rise to the variable *tsaI* gene clusters in *S. aureus* genomes.
344
345 **Intergenic recombination between *tipC* immunity genes in *Streptococcus*.**
346 Three T7SS-secreted antibacterial toxins have been identified in *Streptococcus intermedius*.
347 TelA and TelB are both cytoplasmic-acting toxins neutralised by the TipA and TipB immunity
348 proteins, respectively (12). Our genome analysis indicates that strains generally encode only
349 a single *tipA* gene, while *tipB* is found in up to five copies. The third *S. intermedius* toxin is
350 TelC, a lipid II phosphatase. Protection from TelC toxicity is provided by TipC, a membrane-
351 bound immunity protein that faces the extracellular space (12, 36). Klein *et al*. reported that
352 the number of *tipC* genes encoded at the *telC* locus was highly variable between strains (36).
353 We used gene neighbourhood analysis across Streptococcal genomes to compare the
354 number of *tipC* genes present at the *telC* gene cluster, identifying 15 of them in a strain of
355 *Streptococcus mitis* BCC08 (Fig 6a).
356
357 Alignment of the *S. mitis* BCC08 TipC sequences and their encoding DNA (Fig 6b, Fig S7)
358 showed two regions of high sequence conservation close to the start and end, with a central
359 region of much higher sequence variability (Fig 7a). Using RDP4 to screen for recombination,
360 at least five recombination events were predicted between these genes (Fig 7b, Table S6), in

361 each case almost certainly through the homology blocks we identified. To analyse this further
362 we constructed a maximum likelihood tree to compare *tipC* gene sequences. Genes
363 *D8786_RS05910* and *D8786_RS0585* cluster closely in this tree, which corresponds to the
364 recombination event predicted between these two genes (Fig 7b, Table S6). Likewise,
365 *D8786_RS05920*, which is predicted to be the major parent to *D8786_RS05865* (Table S6)
366 also clusters phylogenetically with this gene. We conclude that similar to *esaG*, intergenic
367 recombination drives the evolution of *tipC* immunity gene repertoires.

368 **DISCUSSION**

369 In this study we have undertaken sequence analysis of *S. aureus* EsaG proteins and their
370 encoding DNA, including their 3' flanking regions, to investigate how copy number variability
371 may arise. We found three blocks of highly conserved nucleotide sequence, a large central
372 one of approximately 100 nucleotides in length, and a 5' and 3' block both of around 55
373 nucleotides each. Homologous recombination occurs at regions of high similarity within
374 nucleotide sequences (Reviewed in 37). The minimum length requirement for efficient
375 recombination in *S. aureus* is unclear but stretches of 40-70 nucleotides have been reported
376 for other bacteria (38-40). Using RDP4 to predict recombination within *esaG* genes, we found
377 strong evidence for recombination, corresponding to events within the three homology blocks
378 we identified. Through the analysis of genome sequences from multiple closely related CC8
379 strains we found extensive evidence of recombination within *esaG* genes to alter copy number,
380 and in some instances the sequence of individual *esaG* genes.

381

382 Previous work has reported that the *S. aureus* tandem-like lipoproteins, encoded on the νSaα
383 island, also show extensive copy number variation across strains (41, 42). Similar analysis to
384 that reported here showed that each *lpl* gene shares a stretch of approximately 130
385 nucleotides of high sequence similarity in its central region. Recombination was demonstrated
386 to occur between the central conserved region of one gene and the same region of the
387 neighbouring gene (40), thus spanning the 3' portion of gene 1, the intergenic region and the
388 5' region of gene 2.

389

390 To investigate whether intergenic recombination might represent a general mechanism for the
391 evolution of T7SS immunity gene families, we examined the organisation of the *tipA*, *tipB* and
392 *tipC* immunity genes in Streptococci. It has previously been noted that *tipC* copy number is
393 highly variable in Streptococcal genomes (36), and our analysis identified that up to 15 copies
394 of *tipC* could be present. Examination of recombination events within *tipC* revealed that
395 intergenic recombination is also a feature, and that it primarily occurs between homologous

396 blocks of sequence identity at the start and end of the genes. The structure of TipC reveals
397 that it has seven beta strands forming a concave face, with three alpha helices made up from
398 the N- and C-terminal regions of the protein (53, Fig 7d). Analysis of colicin DNase toxins and
399 their immunity proteins has shown that sequence divergence between related toxins and
400 immunities tends to concentrate at the binding interface (43). In agreement with this, site-
401 directed mutagenesis has strongly implicated the concave face as the region that binds to
402 TelC, and this region of TipC shows the highest level of sequence divergence (Fig 7d; 36).
403 We speculate that for EsaG the regions of high sequence similarity are not directly involved in
404 toxin binding but may provide a structural framework on which amino acid substitutions in the
405 variable regions accumulate to alter the toxin binding specificity. Mechanisms to allow rapid
406 evolution of immunity proteins are likely to be essential to allow strains to quickly acquire
407 resistance to novel toxin variants. We have also demonstrated that strains can lose significant
408 numbers of immunity genes from such repertoires. This would be expected to reduce the
409 metabolic burden of synthesising a large number of similar proteins in an environment where
410 they provide no benefit. It is likely that there is a trade-off between carrying sufficient immunity
411 proteins to protect against competitors and the metabolic consequences of producing multiple
412 protein variants.
413
414 TspA is a second polymorphic toxin encoded by *S. aureus*. Protection from its membrane-
415 depolarising activity is provided by the immunity protein TsaI (19). TsaI is a polytopic
416 membrane protein predicted to have five transmembrane domains. As with *esaG*, multiple *tsaI*
417 genes are found in *S. aureus* genomes, however, our analysis has indicated that there is little
418 evidence of recombination between them. We found only a single block of high nucleotide
419 sequence similarity at the 5' end of *tsaI* genes. This encodes approximately the first 75 amino
420 acids of TsaI, which would encompass the first two transmembrane domains. At present it is
421 not known whether TsaI neutralises TspA though direct interaction, or through the
422 sequestering of a membrane-bound partner protein with which TspA must interact to facilitate
423 its insertion or folding (or a combination of both). In this context, membrane permeabilising

424 peptide bacteriocins produced by some Gram-positive bacteria require a membrane-bound
425 receptor, such as the membrane components of the mannose phosphotransferase system,
426 for their activity. In the case of the *Lactococcus lactis* lactococcin A bacteriocin, the immunity
427 protein LciA acts through formation of a complex with both the receptor protein and the
428 bacteriocin (44). By analogy it is possible that the first two transmembrane domains of TsaI
429 interact with a candidate receptor, constraining their sequence, whereas the remainder of the
430 immunity protein binds to the toxin and is therefore under diversifying selection. At present it
431 is unclear how *tsaI* gene clusters evolve, although we did note that there was evidence for
432 carriage of *tsaI* genes on Staphylococcal plasmids, which may facilitate gene movement within
433 and between strains (data not shown). Further work would be required to clarify the
434 mechanisms that drive *tsaI* diversity.
435
436 Fragments of a second copy of *tspA* are found within the *tsaI* gene cluster at the *tspA* locus in
437 many (but not all) strains of *S. aureus*. Interestingly, an Rhs toxin encoded by *Salmonella*
438 *enterica* serovar Typhimurium LT2 is also found as a second partial copy downstream of
439 primary toxin locus; in each case both *rhs* copies precede an immunity gene. Repeated
440 passage of the strain resulted in a fraction of the cells undergoing recombination between the
441 *rhs* copies, generating a mosaic gene between *rhs1* and *rhs2* (45). An alignment of the *tspA*
442 fragment (spanning the entire intergenic region between *tsaI10* and *tsaI11*) with full length
443 *tspA* indicated that the regions of homology fall at the 3' of *tspA*, although we did not find any
444 evidence of recombination between these nucleotide sequences in the datasets we analysed.
445 It should be noted that recombination between *tspA* and the *tspA* fragment would generate a
446 novel sequence variant for this toxin, as well as deleting a cluster of *tsaI* genes lying between
447 these regions, potentially helping to increase strain competitiveness under specific growth
448 conditions.

449 **FUNDING INFORMATION**

450 This work was supported by Wellcome through Investigator Award 110183/A/15/Z to T.P. and
451 Sir Henry Wellcome Postdoctoral Fellowship 218622/Z/19/Z to G.M. The NCTC3000 project,
452 which provided the 133 NCTC genome assemblies, was funded via Wellcome Trust grant
453 101503/Z/13/Z. S.R.G. is funded by the Newcastle-Liverpool-Durham BBSRC DTP2 Training
454 Grant, project reference number BB/M011186/1.

455

456 **ACKNOWLEDGEMENTS**

457 We thank Prof Jan-Maarten van Dijl (University of Groningen, NL) for providing us with strain
458 RN6390, and Prof Victor Torres and Dr. Alejandro Pironti for providing the genome sequence
459 of USA300 BKV_2.

460

461 **CONFLICT OF INTERESTS**

462 The authors declare that there are no conflicts of interest.

463

464 **AUTHOR CONTRIBUTIONS**

465 S.R.G. and T.P. conceptualised the study. S.R.G. carried out initial analysis. G.M. wrote
466 custom scripts and produced these outputs. J.D. wrote custom scripts for the analysis of the
467 NCTC dataset and provide the raw data for this. S.R.G. analysed the NCTC dataset. S.R.G.
468 created the figures. S.R.G. and T.P. wrote the manuscript. All authors approved the final
469 version of the manuscript.

470

471

472 **FIGURE LEGENDS**

473

474 **Figure 1. Homologues of *esaG* encoded at the *ess* locus in RN6390.** a. Illustration of the
475 3' of the RN6390 *ess* locus which encodes the T7SS nuclease toxin, EsaD, its cognate

476 immunity protein, EsaG and 11 further homologues of EsaG (numbered *esaG2* – *esaG12*).
477 Note that *esaG4* is shown in hatched shading because it is annotated as a pseudogene.
478 However, it does encode two predicted ORFs, EsaG4i and EsaG4ii. b. Sequence alignment
479 of EsaG homologues encoded by RN6390. The black boxes represent regions of high
480 sequence similarity based on this alignment.
481
482 **Figure 2. Recombination within the RN6390 *esaG* homologues.** a. Regions of high
483 sequence similarity across the RN6390 EsaG protein sequences (middle panel) and the
484 corresponding nucleotide sequences (bottom panel). The positions of homology blocks are
485 shown in grey shading and along with their relative positions along the gene sequence (top
486 panel). The basepair positions that define the conserved regions are taken from the nucleotide
487 sequences of *esaG1*. b. RDP4 was used to predict recombination events within the *esaG*
488 homologues encoded by RN6390. The identity of each gene is given in black at the left, with
489 regions of recombination labelled directly below, in the colour of the gene from which the
490 recombinant section originated. ci-iii. For each of the three variable regions of the RN6390
491 *esaG* homologues a maximum likelihood tree was generated in IQTREE and visualised and
492 annotated in iTOL. The variable regions consist of nucleotide 1-39, 96-251 and 358-465 of
493 *esaG* respectively d. Illustration of regions of high sequence similarity in the *esaG* homologues
494 in RN6390. Black bars represent conserved regions of the gene and the variable regions have
495 been assigned a colour and corresponding number. Homologous regions are coloured with
496 the same colour. Numbers were assigned based on the first gene in the series that had the
497 unique variable region. White hatched shading indicates a pseudogene.
498
499 **Figure 3. A recombination event in an epidemic lineage of USA300 results in loss of**
500 **part of an *esaG* cluster and generation of a novel *esaG* gene.** a. The *esaD* locus of
501 USA300 FPR3757 and USA300 BKV_2. The dashed lines represent the region that is missing
502 from the epidemic strain, USA300 BKV_2, when compared to the USA300 FPR3757 type
503 strain. b. Nucleotide sequence alignment for *SAUSA300_0295* and *SAUSA300_0299* from

504 USA300 FPR3757 and *esaG4* from USA300 BKV_2. Coloured blocks indicated homology
505 between *BKV_2 esaG4* and the genes with which it is aligned.
506
507 **Figure 4. Recombination between *esaG* paralogs results in variation at the *ess* locus in**
508 **CC8 *S. aureus* strains from a historical culture collection.** Two expansion events of the
509 *esaG* repertoire in the NCTC CC8 dataset were identified in a. NCTC10724 and b.
510 NCTC10702 as respective representatives. The solid black line depicts the predicted
511 recombination event to give the recombinant sequence depicted below. Five unique loss
512 events were also identified in this dataset, represented by strains c. NCTC9369, d.
513 NCTC10652, e. NCTC12232, f. NCTC13140 and g. NCTC13196. The dashed line represents
514 the region deleted in the recombinant. Genes were given an individual colour and numbered
515 based on *esaG1-12*, to make clear any recombination events.
516
517 **Figure 5 Recombination also occurs between *DUF5079* genes.** Recombination events
518 were detected between the two *DUF5079* genes at the *ess* locus in the NCTC *S. aureus* CC8
519 dataset. Recombination occurred a. within the *DUF5079* genes and b. in the intergenic regions
520 directly upstream of these, resulting in the loss of *esaG2-5* in several strains. c. An alignment
521 of the region spanning the 5' of *esaG1* to the 3' of *DUF5079-1* with the 5' of *esaG5* to the 3'
522 of *DUF5079-2*. Coloured bars below the alignment represent *esaG* (orange), the intergenic
523 region (white) and *DUF5079* (blue). d. All instances of the region spanning the 5' of *esaG* to
524 the 3' of *DUF5079* was manually extracted from the NCTC CC8 dataset, aligned using MAFFT
525 and visualised using Plotcon to identify regions of high similarity.
526
527 **Figure 6. Homologues of *tipC* encoded at the *telC* locus of *Streptococcus mitis* BCC08.**
528 a. Genetic arrangement of *tipC* genes in *S. mitis* BCC08. b. An alignment of the encoded TipC
529 homologues. The blue boxes represent regions of high sequence similarity, and the dashed
530 line at the N-terminus of the aligned sequences indicates a probable lipoprotein signal peptide.
531

532 **Figure 7. Recombination within the *S. mitis* BCC08 *tipC* homologues.** a. Regions of high
533 sequence similarity across the *S. mitis* BCC08 TipC protein sequences (middle panel) and the
534 corresponding nucleotide sequences (bottom panel). The positions of homology blocks are
535 shown in grey shading, with their relative positions along the gene sequence (top). The first
536 18 amino acids of TipC form a predicted lipoprotein signal sequence which is indicated by pale
537 grey shading. The basepair positions that define the conserved regions are taken from the
538 nucleotide sequences of *D8786_RS05940*. b. RDP4 was used to predict recombination events
539 within the *tipC* homologues. The identity of each gene is given in black at the left, with regions
540 of recombination labelled directly below, in the colour of the gene from which the recombinant
541 section originated. c. A maximum likelihood tree was generated for *tipC* homologues in
542 IQTREE and visualised and annotated in iTOL. d. The conserved (cyan) and variable (orange)
543 regions of TipC were mapped to the crystal structure of *S. intermedius* TipC2 (pdb:6DHX; 53).
544
545
546
547 **References**

548 1. **Hsu T, Hingley-Wilson SM, Chen B, Chen M, Dai AZ *et al.*** The primary mechanism of
549 attenuation of bacillus Calmette-Guerin is a loss of secreted lytic function required for invasion
550 of lung interstitial tissue. *Proc Natl Acad Sci U S A* 2003;100:12420-12425.
551
552 2. **Pym AS, Brodin P, Majlessi L, Brosch R, Demangel C, *et al.*** Recombinant BCG
553 exporting ESAT-6 confers enhanced protection against tuberculosis. Nat Med 2003;9:533-
554 539.
555
556 3. **Stanley SA, Raghavan S, Hwang WW, Cox JS.** Acute infection and macrophage
557 subversion by *Mycobacterium tuberculosis* require a specialized secretion system. *Proc Natl*
558 *Acad Sci U S A* 2003;100:13001-13006.
559
560 4. **Akpe San Roman S, Facey PD, Fernandez-Martinez L, Rodriguez C, Vallin C *et al*.** A
561 heterodimer of EsxA and EsxB is involved in sporulation and is secreted by a type VII secretion
562 system in *Streptomyces coelicolor*. *Microbiology* 2010;156:1719-1729.
563

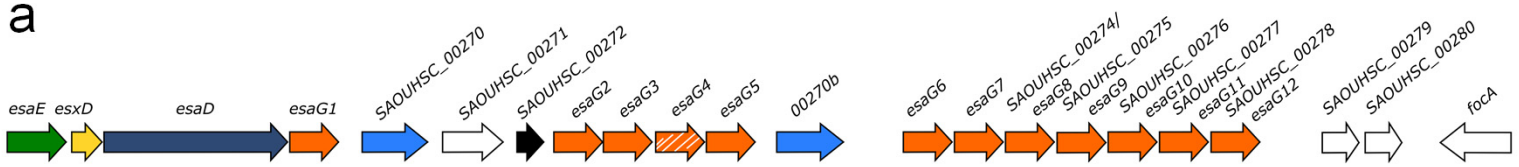
564 5. **Burts ML, Williams WA, DeBord K, Missiakas DM.** EsxA and EsxB are secreted by an
565 ESAT-6-like system that is required for the pathogenesis of *Staphylococcus aureus* infections.
566 *Proc Natl Acad Sci U S A* 2005;102:1169-1174.
567
568 6. **Bunduc CM, Fahrenkamp D, Wald J, Ummels R, Bitter W *et al*.** Structure and dynamics
569 of a mycobacterial type VII secretion system. *Nature* 2021;593:445-448.
570
571 7. **Beckham KSH, Ritter C, Chojnowski G, Ziemianowicz DS, Mullapudi E *et al*.** Structure
572 of the mycobacterial ESX-5 type VII secretion system pore complex. *Sci Adv*
573 2021;7:eabg9923.
574
575 8. **Abdallah AM, Gey van Pittius NC, Champion PA, Cox J, Luirink J, *et al.*** Type VII
576 secretion--mycobacteria show the way. *Nat Rev Microbiol* 2007;5:883-891.
577
578 9. **Zoltner M, Ng WM, Money JJ, Fyfe PK, Kneuper H, *et al.*** EssC: domain structures inform
579 on the elusive translocation channel in the Type VII secretion system. *Biochem J* 2016;
580 473:1941-1952.
581
582 10. **Bowman L, Palmer T.** The Type VII Secretion System of *Staphylococcus*. *Annu Rev*
583 *Microbiol* 2021;75:471-494.
584
585 11. **Tran HR, Grebenc DW, Klein TA, Whitney JC.** Bacterial type VII secretion: An important
586 player in host-microbe and microbe-microbe interactions. *Mol Microbiol* 2021;115:478-489.
587
588 12. **Whitney JC, Peterson SB, Kim J, Pazos M, Verster AJ, *et al.*** A broadly distributed toxin
589 family mediates contact-dependent antagonism between gram-positive bacteria. *Elife*. 2017;
590 6:e26938.
591
592 13. **Chatterjee A, Willett JLE, Dunny GM, Duerkop BA.** Phage infection and sub-lethal
593 antibiotic exposure mediate *Enterococcus faecalis* type VII secretion system dependent
594 inhibition of bystander bacteria. *PLoS genet* 2021;17:e1009204.
595
596 14. **Bowran K, Palmer T.** Extreme genetic diversity in the type VII secretion system of Listeria
597 monocytogenes suggests a role in bacterial antagonism. *Microbiology* 2021;167: doi:
598 10.1099/mic.0.001034
599

600 15. **Cao Z, Casabona MG, Kneuper H, Chalmers JD, Palmer T.** The type VII secretion
601 system of *Staphylococcus aureus* secretes a nuclease toxin that targets competitor bacteria.
602 *Nat Microbiol* 2016;2:16183.
603
604 16. **Warne B, Harkins CP, Harris SR, Vatsiou A, Stanley-Wall N, *et al.*** The Ess/Type VII
605 secretion system of *Staphylococcus aureus* shows unexpected genetic diversity. *BMC*
606 *Genomics* 2016;17:222.
607
608 17. **Baek KT, Frees D, Renzoni A, Barras C, Rodriguez N *et al*.** Genetic variation in the
609 *Staphylococcus aureus* 8325 strain lineage revealed by whole-genome sequencing. PLoS
610 ONE. 2013;8:e77122.
611
612 18. **Ohr RJ, Anderson M, Shi M, Schneewind O, Missiakas D.** EssD, a Nuclease Effector
613 of the *Staphylococcus aureus* ESS Pathway. *J Bacteriol* 2016;199:e00528-16.
614
615 19. **Ulhuq FR, Gomes MC, Duggan GM, Guo M, Mendonca C, *et al.*** A membrane-
616 depolarizing toxin substrate of the *Staphylococcus aureus* Type VII secretion system mediates
617 intra-species competition. *Proc Natl Acad Sci U S A* 2020;117:20836-20847.
618
619 20. **Ross BD, Verster AJ, Radey MC, Schmidtke DT, Pope CE *et al*.** Human gut bacteria
620 contain acquired interbacterial defence systems. *Nature* 2019;575:224-228.
621
622 21. **Katoh K, Misawa K, Kuma K, Miyata T.** MAFFT: a novel method for rapid multiple
623 sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;30:3059-3066.
624
625 22. **Edgar RC.** MUSCLE: multiple sequence alignment with high accuracy and high
626 throughput. *Nucleic Acids Res* 2004;32:1792-1797.
627
628 23. **Martin DP, Murrell B, Golden M, Khoosal A, Muhire B.** RDP4: Detection and analysis
629 of recombination patterns in virus genomes. *Virus Evol* 2015;1:vev003.
630
631 24. **Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T.** trimAl: a tool for automated
632 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972-1973.
633
634 25. **Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD *et al*.** IQ-TREE 2:
635 New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol
636 Evol 2020;37:1530-1534.

637
638 26. **Letunic I, Bork P.** Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree
639 display and annotation. *Nucleic Acids Res* 2021;49:W293-W6.
640
641 27. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** Basic local alignment search
642 tool. *J Mol Biol* 1990;215:403-410.
643
644 28. **Guy L, Kultima JR, Andersson SG.** genoPlotR: comparative gene and genome
645 visualization in R. *Bioinformatics* 2010;26**:**2334-2335.
646
647 29. **RStudio Team.** RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL
648 http://www.rstudio.com/ 2020.
649
650 30. **Saha CK, Sanches Pires R, Brolin H, Delannoy M, Atkinson GC.** FlaGs and webFlaGs:
651 discovering novel biology through the analysis of gene neighbourhood conservation.
652 *Bioinformatics* 2021;37:1312-1314.
653
654 31. **Eddy SR.** Accelerated profile HMM searches. *PLoS Comput Biol* 2011;7:e1002195.
655
656 32. **Copin R, Sause WE, Fulmer Y, Balasubramanian D, Dyzenhaus S *et al*.** Sequential
657 evolution of virulence and resistance during clonal spread of community-acquired methicillin-
658 resistant Staphylococcus aureus. *Proc Natl Acad Sci U S A* 2019;116:1745-1754.
659
660 33. **Seybold U, Kourbatova EV, Johnson JG, Halvosa SJ, Wang YF *et al*.** Emergence of
661 community-associated methicillin-resistant *Staphylococcus aureus* USA300 genotype as a
662 major cause of health care-associated blood stream infections. *Clin Infect Dis* 2006;42:647-
663 656.
664
665 34. **Dicks J, Turnbull JD, Russell J, Parkhill J, Alexander S.** Genome Sequencing of a
666 Historic *Staphylococcus aureus* Collection Reveals New Enterotoxin Genes and Sheds Light
667 on the Evolution and Genomic Organization of This Key Virulence Gene Family. *J Bacteriol*
668 2021;203:e00587-20.
669
670 35. **Sullivan MJ, Petty NK, Beatson SA.** Easyfig: a genome comparison visualizer.
671 *Bioinformatics* 2011; 27:1009-1010.
672

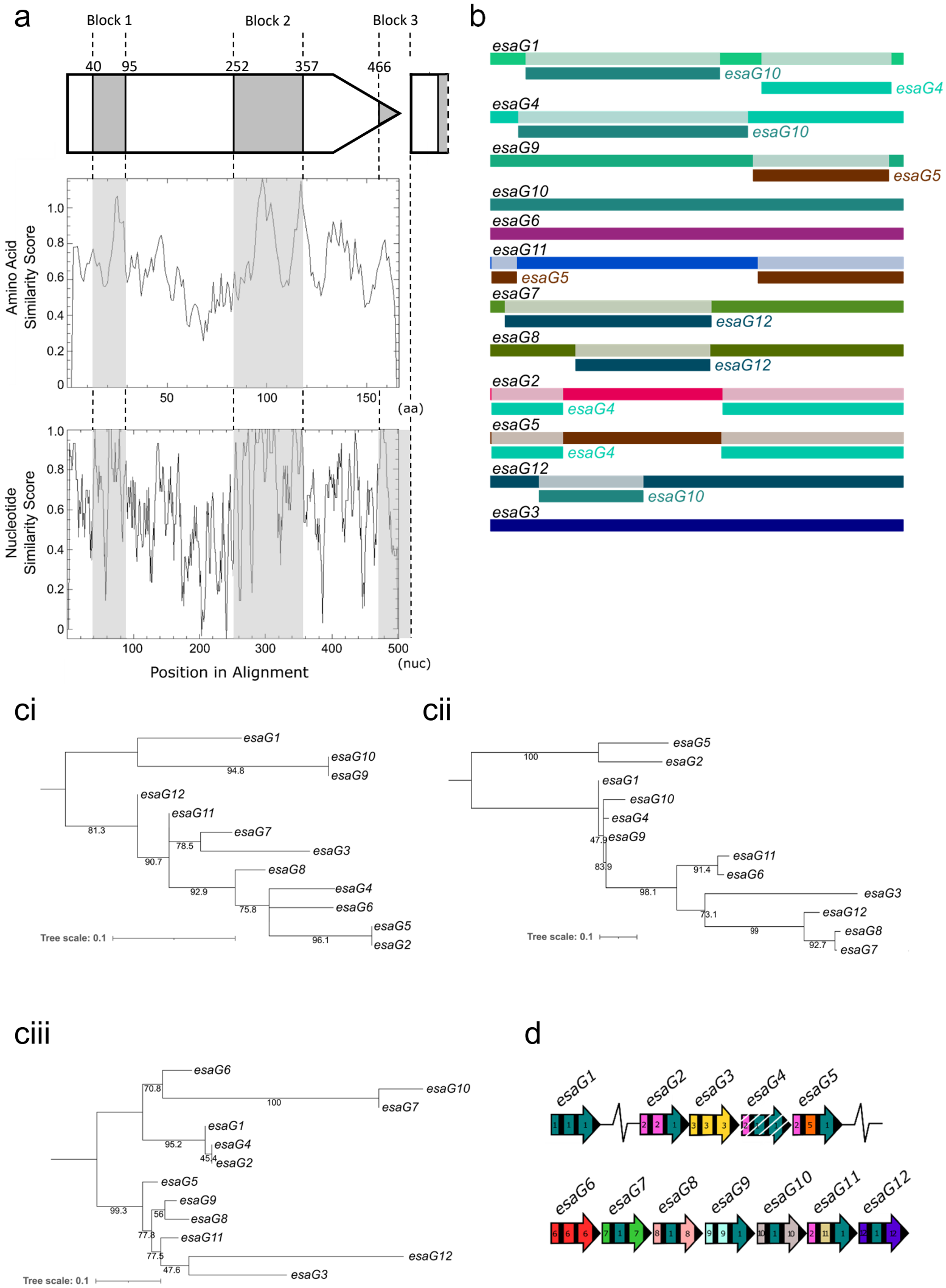
673 36. **Klein TA, Pazos M, Surette MG, Vollmer W, Whitney JC.** Molecular Basis for Immunity
674 Protein Recognition of a Type VII Secretion System Exported Antibacterial Toxin. *J Mol Biol*
675 2018;430:4344-4358.
676
677 37. **Michel B, Leach D.** Homologous Recombination-Enzymes and Pathways. *EcoSal Plus*
678 2012;5: doi: 10.1128/ecosalplus.7.2.7.
679
680 38. **Khasanov FK, Zvingila DJ, Zainullin AA, Prozorov AA, Bashkirov VI.** Homologous
681 recombination between plasmid and chromosomal DNA in *Bacillus subtilis* requires
682 approximately 70 bp of homology. *Mol Gen Genet* 1992;234:494-497.
683
684 39. **Shen P, Huang HV.** Homologous recombination in *Escherichia coli*: dependence on
685 substrate length and homology. *Genetics* 1986;112:441-457.
686
687 40. **Fujitani Y, Yamamoto K, Kobayashi I.** Dependence of frequency of homologous
688 recombination on the homology length. *Genetics* 1995;140:797-809.
689
690 41. **Tsuru T, Kawai M, Mizutani-Ui Y, Uchiyama I, Kobayashi I.** Evolution of paralogous
691 genes: Reconstruction of genome rearrangements through comparison of multiple genomes
692 within *Staphylococcus aureus*. *Mol Biol Evol* 2006;23:1269-1285.
693
694 42. **Tsuru T, Kobayashi I.** Multiple genome comparison within a bacterial species reveals a
695 unit of evolution spanning two adjacent genes in a tandem paralog cluster. *Mol Biol Evol*
696 2008;25:2457-2473.
697
698 43. **Ruhe ZC, Low DA, Hayes CS.** Polymorphic Toxins and Their Immunity Proteins:
699 Diversity, Evolution, and Mechanisms of Delivery. *Annu Rev Microbiol* 2020;74:497-520.
700
701 44. **Diep DB, Skaugen M, Salehian Z, Holo H, Nes IF.** Common mechanisms of target cell
702 recognition and immunity for class II bacteriocins. *Proc Natl Acad Sci U S A* 2007;104:2384-
703 2389.
704
705 45. **Koskiniemi S, Garza-Sánchez F, Sandegren L, Webb JS, Braaten BA *et al*.** Selection
706 of orphan Rhs toxin expression in evolved *Salmonella enterica* serovar Typhimurium. *PLoS*
707 *Genet* 2014;10:e1004255.

a

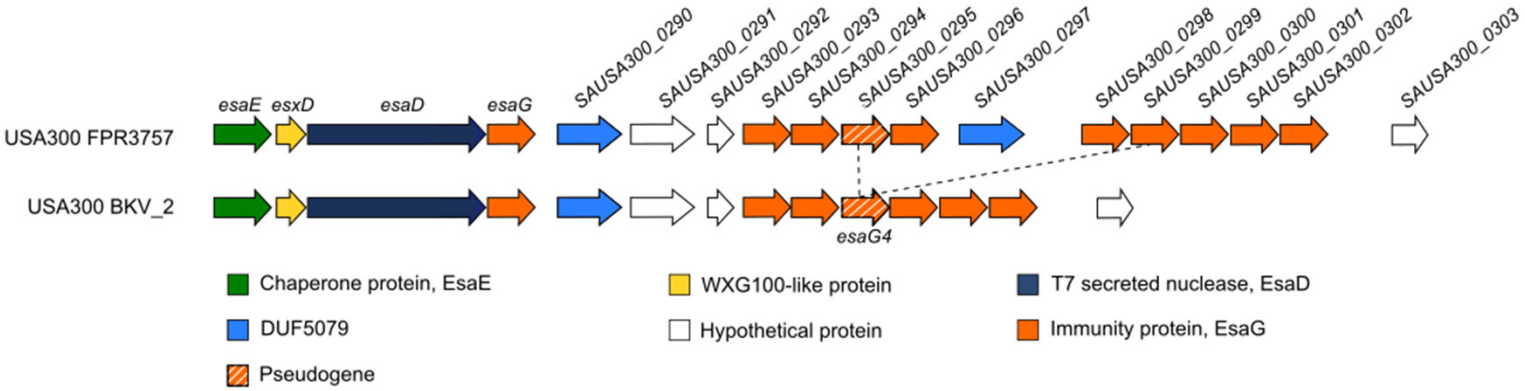



b

```
EsaG3      1  MNFEEKLSEMYNKIANEINEMIPVEWEKVYVIAYVDDGGGQVIFYYTKPRNDELYYYSSI
EsaG5      1  MTFEEKLSEMYSEIANKIISMIPVEWEKVYAMAYIDEECGEVFYNYTEPSSDELFYYTSV
EsaG6      1  MTFEEKLSEIYNKIANEISGMIPVEWDQVFTIAYVNDRGGEIVFNYTKPGSDELNYYTYI
EsaG12     1  MTFEEKLSQMYNEIANEISGMIPVEWEKVYTIAYVDDEGGEVVFNYTKPGSEDLNYYSDI
EsaG7      1  MTFEEKLSQMYNKIASEISGMIPVEWEQVFTIAYVTDQAGEVIFNYTKPGSDELNYYSDI
EsaG8      1  MTFEGKLSQMYNEIANEISGMIPVEWEQVFTIAYVTDQAGEVIFNYTKPDSDELNYYSDI
EsaG2      1  MTFEEKLSEMYSEIANKISSMIPVEWEKVYAMAYIDDGGGEVFYYYTEPGSNELYYYTSV
EsaG11     1  MTFEEKLSQMYNEIANEISGMIPVEWENIYTIAYVTDQGGEVIFNYTKPGSDELNYYTYI
EsaG10     1  MTFEEKISKLYNEIANEISSMIPVEWEKVYTMAYIDDGGGEVFFNYTKPGSEDLNYYTDI
EsaG4i     1  MIFEEKLNEMYNEIANKISSMIPVEWEKVYTMAYIDDGGGEVFFNYTKQR----------
EsaG9      1  MTFEEKISKLYNEIANEISSMIPVEWEKVYTMAYIDDGGGEVFFNYTKPGSDDLNYYTDI
EsaG1      1  MTFEEKLSKIYNEIANEISSMIPVEWEKVYTMAYIDDGGGEVFFNYTKPGSDDLNYYTNI
EsaG4ii    1  ------------------------------------------------------------

EsaG3     61  VEDYNVLEEIFDDLWMELYRSFKKLRNIFKEESLEPWTSCEFDFTKEDKLKVSFDYIDWK
EsaG5     61  IKKYNLLKSSFMDSVYELHDQFEELREVFIEEGLEPWTSCEFDFTKEDKLKVSFDYIDWI
EsaG6     61  PREYNVSEKVFYDLWTDLYRLFKKLRNAFKEEDLEPWTSCEFDFTRDGKLNVVFDYVDWM
EsaG12    61  PKDCNVSKDIFKNSWFKVYRMFDELRETFKKEDLEPWTSCEFDFTRKGNLKVSFDYIDWI
EsaG7     61  PKDCNVSKDIFKNSWFKVYRMFDELRETFKEEELEPWTSCEFDFTRDGKLNVSFDYIDWV
EsaG8     61  PKDCNVSKDIFKNSWFKVYRMFDELRETFKEEGLEPWTSCEFDFTRDGKLNVSFDYIDWI
EsaG2     61  LNKYDISESEFMDSVYELYKQFQNLRNLFKEEGHEPWTSCEFDFTREGELKVSFDYIDWI
EsaG11    61  PREYNVSEKVFYDLWTDLYRLFKKLRETFKEEGLEPWTSSEFDFTSEGKLKVSFDYIDWI
EsaG10    61  PKEYNVSVQVFDDLWMDLYDLFKNLRNLFKEEGLEPWTSCEFDFTRDGKLNVSFDYIDWA
EsaG4i        ------------------------------------------------------------
EsaG9     61  PKEYNISVQVFDDLWMDLYDLFEELRDLFKEEGLEPWTSCEFDFTSEGKLKVSFDYIDWI
EsaG1     61  PKEYNISVQVFDDLWMDLYDLFEELRDLFKEEDLEPWTSCEFDFTREGELKVSFDYIDWI
EsaG4ii    1  -------MQVFDDLWMDLYDLFEELRNLFKEEGLEPWTSCEFDFTREGKLKVSFDYIDWI

EsaG3    121  NTEFDQLGQENYYNYRKFGIIPEMEYEMEEVKQIEQYIKEQEEAEQ
EsaG5    121  NTEFDQLGRENYYMYKKFGVLPETEYEMEEVKEIEQYIKEQEE---
EsaG6    121  NSEFGPIAKENYYMYKKFGVLPETEYEINKVKEIDQYVKEQDEAEL
EsaG12   121  KLGFGPSGKENYYMYKKFGILPDMEYEMEEIRAVEKYVKEQE----
EsaG7    121  NSEFGPMGREHYYMYKKFGIWPEKEYAINWVEKIKDYVKEQDEAEL
EsaG8    121  NTEFDQLGRQNYYMYKKFGVIPEMEYEMEEVKEIEQYIKEQEEAEQ
EsaG2    121  NSEFGQVGRQNYYKYRKFGILPETEYEINKVKEIEQYVKEQEEAEQ
EsaG11   121  NTEFDQLGRENYYMYKKFGVLPEMEYEMEEVKEIEQYIKEQDEAEL
EsaG10   121  NSEFGQMGREHYYMYKKFGIWPEKEYAINWVKKIKDYVKEQDEAEL
EsaG4i        ------------------------------------------------------------
EsaG9    121  NTEFDQLGRENYYMYKKFGVLPEMEYEMEEIKEIDQYIKEQDEAEI
EsaG1    121  NSEFGQIGRQNYYKYRKFGILPETEYEINKVKEIEQYIKELEE---
EsaG4ii   54  NSEFGQVGRQNYYKYRKFGILPETEYEINKVKEIEQYIKEQDEAEL
```

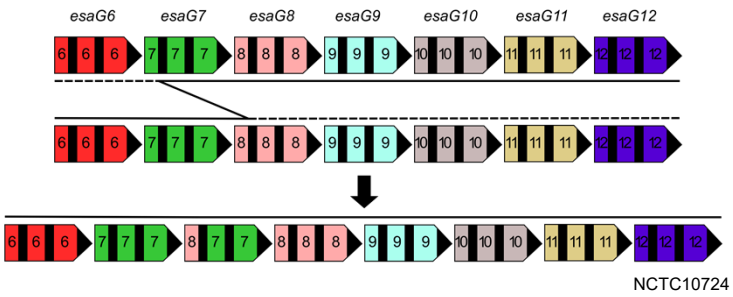

a
Block 1
Block 2
Block 3
40
95
252
357
466
Amino Acid Similarity Score
1.0
0.8
0.6
0.4
0.2
0
50
100
150
(aa)
Nucleotide Similarity Score
100
200
300
400
500
(nuc)
Position in Alignment
b
esaG1
esaG10
esaG4
esaG4
esaG10
esaG9
esaG5
esaG10
esaG6
esaG11
esaG5
esaG7
esaG12
esaG8
esaG12
esaG2
esaG4
esaG5
esaG4
esaG12
esaG10
esaG3
ci
esaG1
esaG10
esaG9
94.8
esaG12
esaG11
esaG7
esaG3
81.3
78.5
90.7
esaG8
92.9
esaG4
esaG6
75.8
esaG5
esaG2
96.1
Tree scale: 0.1
cii
esaG5
esaG2
100
esaG1
esaG10
esaG4
esaG9
47.9
83.9
esaG11
esaG6
91.4
98.1
esaG3
73.1
esaG12
99
esaG8
esaG7
92.7
Tree scale: 0.1
ciii
esaG6
esaG10
esaG7
70.8
100
esaG1
esaG4
esaG2
95.2
45.4
esaG5
esaG9
esaG8
esaG11
esaG12
esaG3
99.3
56
77.8
77.5
47.6
Tree scale: 0.1
d
esaG1
esaG2
esaG3
esaG4
esaG5
esaG6
esaG7
esaG8
esaG9
esaG10
esaG11
esaG12

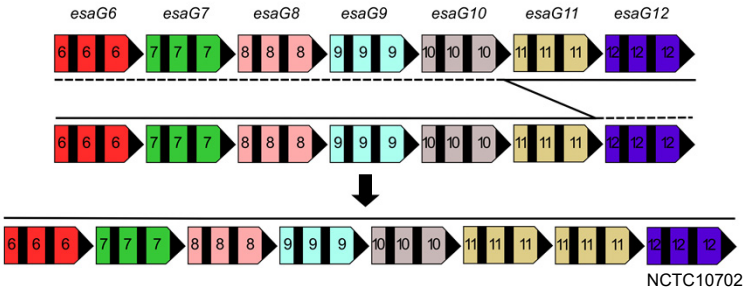
a

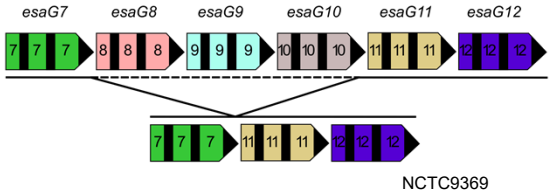


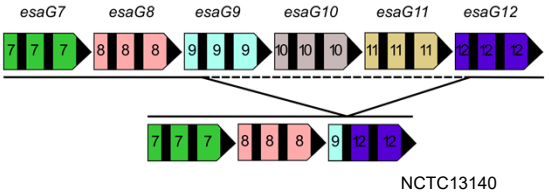

b

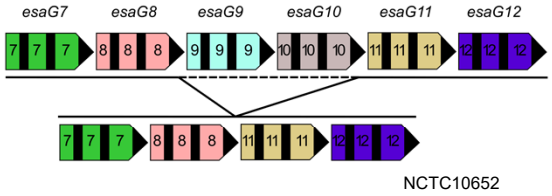
```
VDXVD633b35<8    1  ATGATTTTCGAAGAAAAACTAAATGAAATGTACAACGAGATTGCGAATAAAATTAGTAGC
ENYb5#hvdJ7      1  ATGATTTTCGAAGAAAAACTAAATGAAATGTACAACGAGATTGCGAATAAAATTAGTAGC
VDXVD633b35<<    1  ATGACTTTCGAAGAGAAACTAAGTCAAATGTACAATAAAATTGCAAGTGAGATCAGTGGA

VDXVD633b35<8   61  ATGATACCAGTAGAATGGGAAAAGGTATATACAATGGCTTATATAGATGATGGAGGAGGT
ENYb5#hvdJ7     61  ATGATACCAGTAGAATGGGAAAAGGTATATACAATGGCTTATATAGATGATGGAGGAGGT
VDXVD633b35<<   61  ATGATACCAGTAGAATGGGAGCAAGTATTTACAATAGCTTATGTAACTGATCAAGCTGGG

VDXVD633b35<8  121  GAAGTATTCTTTAATTATACTAAAC-----AGCGATGAATTGAATTATTACACCGATATA
ENYb5#hvdJ7    121  GAAGTATTCTTTAATTATACTAAAC-----AGCGATGAATTGAATTATTACACCGATATA
VDXVD633b35<<  121  GAAGTCATTTTTAATTATACTAAACCAGGTAGTGATGAATTAAATTATTATTCAGACATA

VDXVD633b35<8  176  CCTAAGGAGTATAACATTTCTGTGCAAGTATTTGATGATTTATGGATGGATTTATATGAT
ENYb5#hvdJ7    176  CCTAAGGAGTATAACATTTCTGTGCAAGTATTTGATGATTTATGGATGGATTTATATGAT
VDXVD633b35<<  181  CCTAAAGATTGCAATGTCTCAAAAGATATTTTTAAGAATTCATGGTTTAAAGTTTATCGA

VDXVD633b35<8  236  TTGTTTGAGGAATTAAGAAATTTATTTAAAGAAGAAGGACTAGAACCATGGACATCATGC
ENYb5#hvdJ7    236  TTGTTTGAGGAATTAAGAAATTTATTTAAAGAAGAAGGACTAGAACCATGGACATCATGC
VDXVD633b35<<  241  ATGTTTGATGAGTTAAGAGAAACTTTTAAAGAAGAAGAGCTTGAACCATGGACATCATGC

VDXVD633b35<8  296  GAATTTGATTTTACAAGAGAAGGTAAATTAAAAGTTTCATTTGATTATATTGATTGGATA
ENYb5#hvdJ7    296  GAATTTGATTTTACAAGAGAAGGTAAATTAAAAGTTTCATTTGATTATATTGATTGGGTG
VDXVD633b35<<  301  GAATTTGACTTTACAAGAGATGGCAAATTGAATGTATCTTTTGATTATATTGATTGGGTG

VDXVD633b35<8  356  AATTCAGAATTTGGTCAAGTAGGTCGACAAAATTACTATAAGTATAGAAAATTTGGAATT
ENYb5#hvdJ7    356  AATTCAGAATTTGGACCAATGGGAAGAGAACATTATTATATGTATAAAAATTTGGAATT
VDXVD633b35<<  361  AATTCAGAATTTGGACCAATGGGAAGAGAACATTATTATATGTATAAAAATTTGGAATT

VDXVD633b35<8  416  TTACCAGAAACGGAATATGAAATTAATAAAGTTAAAGAAATCGAGCAATATATTAAAGAG
ENYb5#hvdJ7    416  TGGCCTGAAAAAGAATATGCCATAAATTGGGTTGAAAAAATAAAAGATTATGTTAAAGAG
VDXVD633b35<<  421  TGGCCTGAAAAAGAATATGCCATAAATTGGGTTGAAAAAATAAAAGATTATGTTAAAGAG

VDXVD633b35<8  476  CAAGATGAAGCTGAACTATAG
ENYb5#hvdJ7    476  CAAGATGAAGCTGAACTATAG
VDXVD633b35<<  481  CAAGATGAAGCTGAACTATAG
```

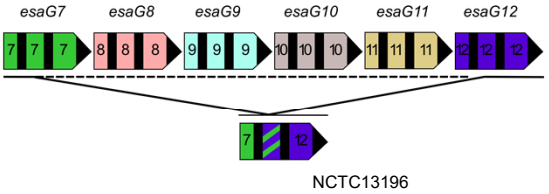

a
esaG6
esaG7
esaG8
esaG9
esaG10
esaG11
esaG12
NCTC10724

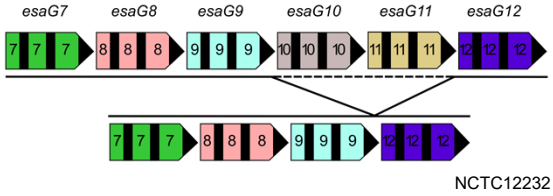

b
esaG6
esaG7
esaG8
esaG9
esaG10
esaG11
esaG12
NCTC10702

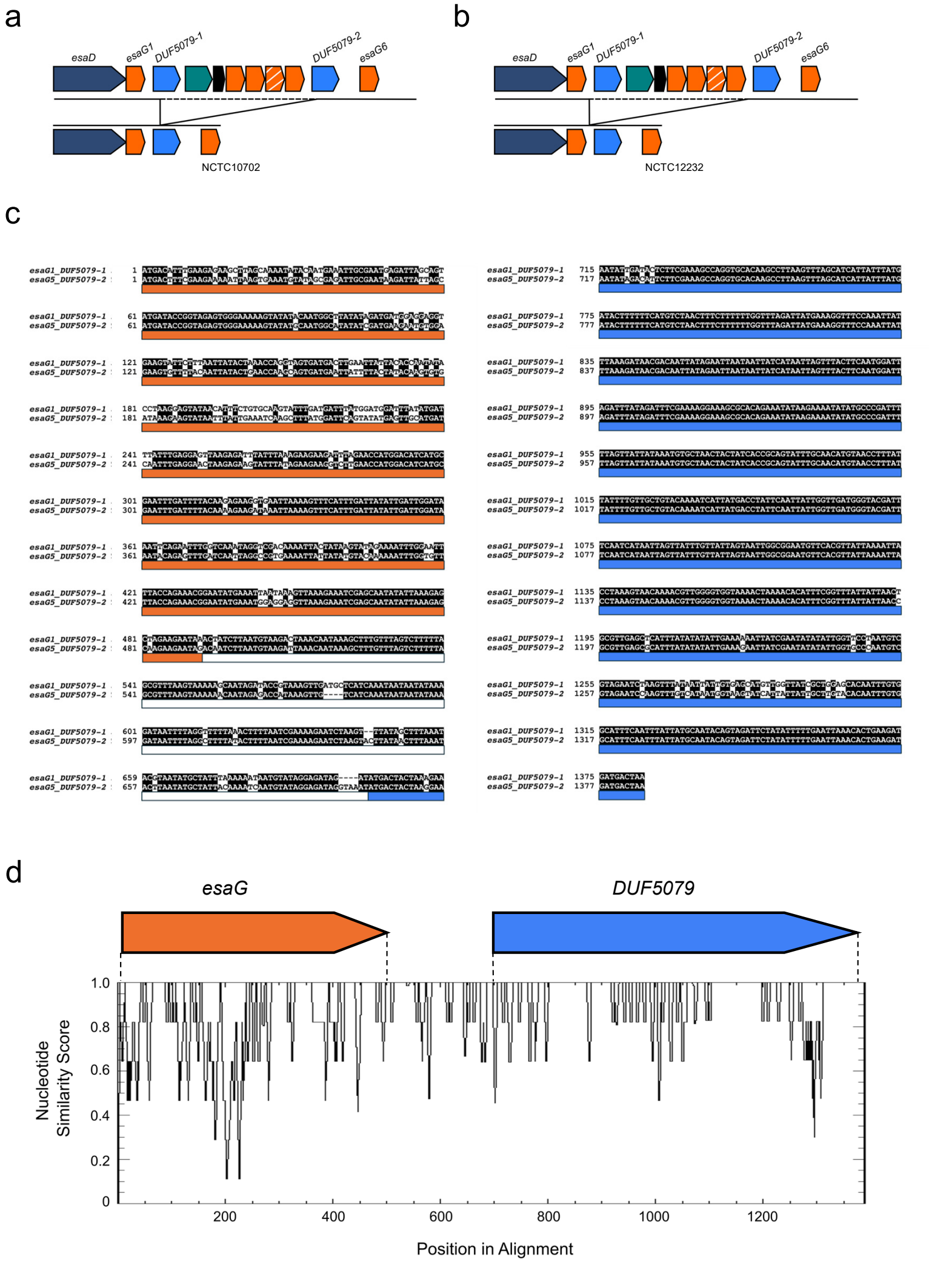

c
esaG7
esaG8
esaG9
esaG10
esaG11
esaG12
NCTC9369


f
esaG7
esaG8
esaG9
esaG10
esaG11
esaG12
NCTC13140

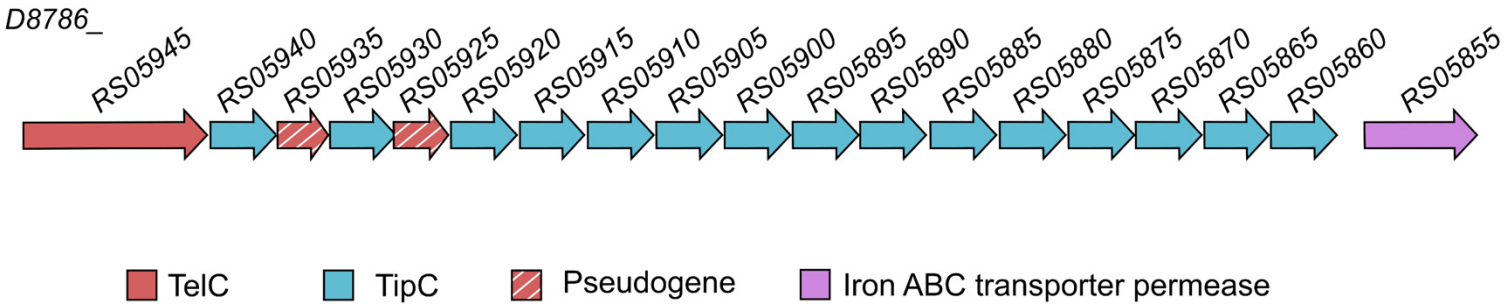

d
esaG7
esaG8
esaG9
esaG10
esaG11
esaG12
NCTC10652


g
esaG7
esaG8
esaG9
esaG10
esaG11
esaG12
NCTC13196


e
esaG7
esaG8
esaG9
esaG10
esaG11
esaG12
NCTC12232

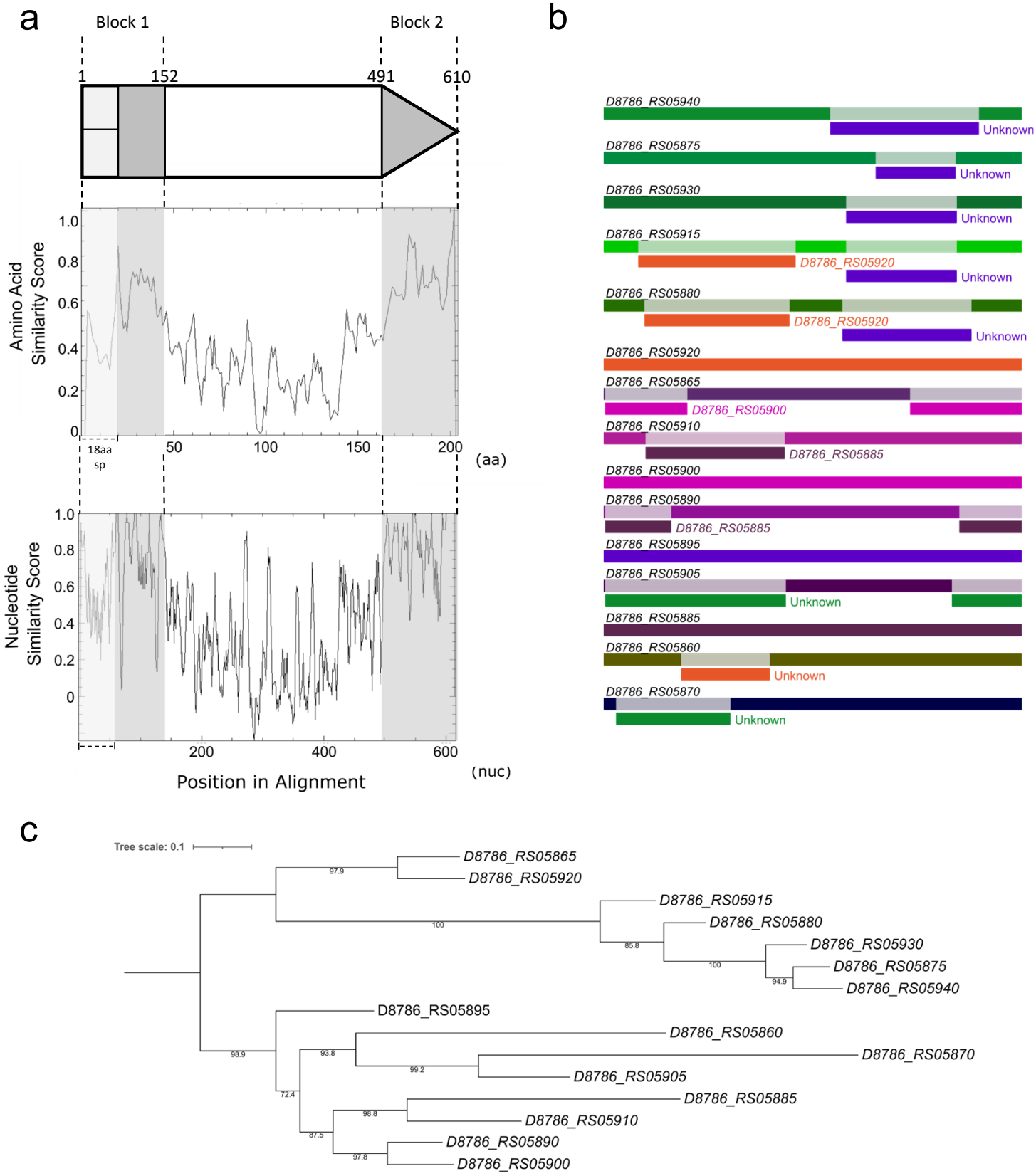

a
esaD
esaG1
DUF5079-1
DUF5079-2
esaG6
NCTC10702
b
esaD
esaG1
DUF5079-1
DUF5079-2
esaG6
NCTC12232
c
esaG1_DUF5079-1
esaG5_DUF5079-2
d
esaG
DUF5079
Nucleotide Similarity Score
Position in Alignment

a




b

```
D8786_RS05940   1  MRKIFSLLALVLIVVLFCFYFFPKQPKNIFDEIYQETEKTYRTNNILRYIDGFKISPGWP
D8786_RS05930   1  MKKILGVLTIVVLLVSVYFYFFPKQPKNIFDEIYQETEKTYRSNNVLRHIDGFKIRPDWP
D8786_RS05875   1  MKKILGVLTIIVLLVSVCFYFFPKQPKNIFDEIYQETEKTYRTNNILRHIDGFKIRAGWS
D8786_RS05915   1  MKKILGLVALIVLIVSSCFYFFIHQPKNIFDEIYQETEKTYRTNNILRNIDGFKIRAGWP
D8786_RS05880   1  MKKILGVLTIIVLLVLVCFYFFVRQPKNIFDEIYQETEKTYRSNNILRNIDGFKIRPDWP
D8786_RS05870   1  MKKILGVLTILVLLVSVCLYFFTSQPKNIFDEIYQQTEKTYLGNNIFNQLKDVEVHK-YQ
D8786_RS05860   1  MKKILGLIALIALIVLSCFYFFIHQPKNIFDEIYQETEKSYQVDNIFGKNEDIIVRPVWP
D8786_RS05920   1  MKKTLSILAIVIIISSCFYFFSRQPKNIFDEIYQETEKTYRTNNILRKIDGFEISPGWP
D8786_RS05865   1  MKKILGLVALILLIVLSCFYFFIHQPKNIFDEIYQETEKTYRTNNILRNIEGFEIDDVWP
D8786_RS05885   1  MKKILGLVALFVIIISSCFYFFPKQPKNIFDEIYQETEKTYRSNNILRNIDGFKIRAVWP
D8786_RS05905   1  MKKILGLVALFVIIVLSCFYFFIQQPKNIFDEIYQETEKTYLGNNVFNNLKDVEVHK-YQ
D8786_RS05895   1  MKKILGLVALFVLIILSCFYFFPKQPKNIFDEIYQETEKTYQTNNILRNIEGFDIKPVWP
D8786_RS05910   1  MKKILGLVALVLLVVLFCFYFFPKQPNNIFDEIYQETEKTYHTNNILRNIDGFKIRPDWP
D8786_RS05900   1  MKKILGLVALILLIVLSCFYFFTNHPKNIFDEIYQETEKTYRSNNILRNIDGFKISPGWP
D8786_RS05890   1  MKEILGLVALFVIIISSCFYFFIHQPKNIFDEIYQETEKTYRSNNILRNIEGFEIRPDWP

D8786_RS05940  61  SDDPNISYTPFGKY--ETLPKGYSDITIDLNFGKGIKGVLILFERKTNSNITLWYSAHYN
D8786_RS05930  61  SDDPNISYTPFGKY--ETLTKGYSDITIHFNFGEGIKGVSIRFERKTNSNITLWYSAHYN
D8786_RS05875  61  SDDPNISYTPFGKY--ETLPKGYSDITIDLNFGKGIKGVLILFERKTNSNITLWYSAHYN
D8786_RS05915  61  NDSEYFAYTPSGKY--QTHPEGYKDISIGFNFGSGIKGMTILFEKRINSDITLWYSAHYN
D8786_RS05880  61  NDGEYFAYTPSGKY--QTHPEGYKDISISFNFGEGIKGMTIRFEKRINSDITLWYSAHYN
D8786_RS05870  60  EYDDDSKFYSYVLYQKQKTPHDYKNIDLIFHFTKDTNTVFVXFEKELSNGLRIFIFGRYL
D8786_RS05860  61  SDNESLKYSPHVTY-KDNIFNNYCKVVLGFNFQNLEQTSFIRFDKYIDSGVRIRIRTVYS
D8786_RS05920  61  SDGEYFKYSPLGKY--KTLPEGYLELRVGFNFEKAYSKMFVSVERKIADGTKIWMISKYN
D8786_RS05865  61  SDGDYFKYSPLGKY--KTLPEDYLELRVGFNFEKAYSKMFVSVERKIADGTKLWMVNKYN
D8786_RS05885  61  SDDPNILYTPFGLYNKEKTPSDYSEIEIGFNFQNSQKVLFIYSERRLSSNVSVKVWGSYN
D8786_RS05905  60  EYSDDSKFYPSIVYERNALPDSYEKIAIDFNFKSEAQLSLISFEKLIESNVSIKMWTVYS
D8786_RS05895  61  SDGEFLRYTPSGNY--KNIPEGYLELRIGFSFYSEDEIGSISFEKRIESNVNIKMWTVYS
D8786_RS05910  61  SDDPNILYTPFGLYNKEKTPSDYSEIEIGFNFEYTSQLSSVSFERQVGLNTRVRIWNKYT
D8786_RS05900  61  SDGEYSKYYPFGTYNKDHTPEEYFEVEISFNFATKTQLSSISFEKRIGPNVRVRLWNKYI
D8786_RS05890  61  SDGEYSKYYPFGTYNKDHTPEEYFEIEIGFNFTTKTQLSSISFEKRIGPNVRVRIWDNYT

D8786_RS05940 119  IKKKVLKKELAIFEEPRKPGQFIDDEEKVREYLRKNNISKEELEKDYDEIINQKVLKDWC
D8786_RS05930 119  LQKKVLKKKLAIFEEPRKPGQFIDDEEKVREYLRKNNISKEELEKDFDEIVNQKVLKDWC
D8786_RS05875 119  MQKKVLKKELAIFEEPRKPGQFIDDEEKVREYLRKNNISKEELEKDYDEIINQKVLKDWC
D8786_RS05915 119  IKKKVLQKEIAIFEEPRQPGQYLEDEEKIREYLRKYNISKEELEKDYDEIVNQKVLKDWC
D8786_RS05880 119  MQKKVLKKGLAIFEEPRKPGQYLDDEEKVRNYLKKYNITKEDLEKDYDEIVNQKVLKDWC
D8786_RS05870 120  TKEKLFQKNVQLLINQDGTDKSIEDEAQVKSYLEQYGITAKDLDSYYDEIVNQKVLKDWC
D8786_RS05860 120  YLEKSLKKEVEVIVIKGNSENYIDNESQVKSYLEQYGITAKDLDSYYDEIVNQKVLKDWC
D8786_RS05920 119  PNTKTITKSIQIVL-SGNEDSYIEDEAQVKSYLEKYGITAKDLDSYYDEIVNQKVLKDWC
D8786_RS05865 119  PNTKMITKSIQIVL-SGKEDSYIEDEAQVKSYLEQYGITAKDLDSYYDEIVNQKVLKDWS
D8786_RS05885 121  YKLGVLSKEVTIIKKENNSKVYIDEQSEVKSYLEQYGITTKDLGSYYDDIVNQRVLKDWC
D8786_RS05905 120  HKERSLKKFVKIVLKKAGTKNYIEDEAQVKSYLEQYGITAKDLDSYYDEIVNQKVLKDWC
D8786_RS05895 119  HKERNLKKFVKIGIKKADTETYIEDEAQVKSYLEKYGITAKDLDSYYDEIVNQKVLKDWC
D8786_RS05910 121  YQDRTLKKSVRVVLKKADTETYIEDEAQVKSYLEQYGITAKDLDSYYDEIVNQKVLKDWC
D8786_RS05900 121  RKDRVLKKFVKIGLKKADTETYIEDEAQVKSYLEQYGITAKDLDSYYDEIVNQKVLKDWC
D8786_RS05890 121  RKDRVLKKFVKIGLKKADTETYIEDEAQVKSYLEQYGITAKDLDSYYDEIVNQKVLKDWC

D8786_RS05940 179  TIYDSKYSPSNYGEVKIETQWENW
D8786_RS05930 179  SIYDSKYSPSNYGEVKIETQWENW
D8786_RS05875 179  SIYDSKYSSGNYGEVKVETQWENW
D8786_RS05915 179  SIYDSKYSPSNYGEVKIETQWENW
D8786_RS05880 179  SIYDSKYSPSNYGDVKIETQWENW
D8786_RS05870 180  TIYDSKYSPSNYGDVKVETQWENW
D8786_RS05860 180  SIYDSKYSPSNYGEVKVETQWENW
D8786_RS05920 178  TIYDSKYSPSNYGDVKIETQWENW
D8786_RS05865 178  SIYDSKYSPSNYGDIKIETQWENW
D8786_RS05885 181  SIYDSKYSPSNYGDVKIETQWENW
D8786_RS05905 180  TIYDSKYLPSNYGDVKVETQWENW
D8786_RS05895 179  SIYDSKYSPSNYGEVKVETQWENW
D8786_RS05910 181  TIYDSKYSPSNYGEVKVETQWENW
D8786_RS05900 181  TIYESKYSPSNYGEVKVETQWENW
D8786_RS05890 181  SIYDSKYSPSNYGEVKVETQWENW
```


a
Block 1
Block 2
1
152
491
610
Amino Acid Similarity Score
1.0
0.8
0.6
0.4
0.2
0
18aa sp
50
100
150
200
(aa)
Nucleotide Similarity Score
200
400
600
(nuc)
Position in Alignment
b
D8786_RS05940
Unknown
D8786_RS05875
Unknown
D8786_RS05930
Unknown
D8786_RS05915
D8786_RS05920
Unknown
D8786_RS05880
D8786_RS05920
Unknown
D8786_RS05920
D8786_RS05865
D8786_RS05900
D8786_RS05910
D8786_RS05885
D8786_RS05900
D8786_RS05890
D8786_RS05885
D8786_RS05895
D8786_RS05905
Unknown
D8786_RS05885
D8786_RS05860
Unknown
D8786_RS05870
Unknown
c
Tree scale: 0.1
D8786_RS05865
D8786_RS05920
D8786_RS05915
D8786_RS05880
D8786_RS05930
D8786_RS05875
D8786_RS05940
D8786_RS05895
D8786_RS05860
D8786_RS05870
D8786_RS05905
D8786_RS05885
D8786_RS05910
D8786_RS05890
D8786_RS05900
97.9
100
85.8
100
94.9
98.9
93.8
99.2
72.4
98.8
87.5
97.8
d