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

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

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

*To whom correspondence should be addressed.

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

Tel +44 191 208 3219

Running title: Expansion of T7 immunity gene families in firmicutes

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. During genome sequence analysis of closely related *S. aureus* strains we noted that there had been a deletion of six consecutive *esaG* copies in one lineage. To investigate this further, we analysed the sequences of the tandem *esaG* genes and their encoded proteins. We identified three blocks of high sequence homology 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 of *esaG* genes from *S. aureus* genomes. 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 homology 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 block of homology 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.

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

38 **INTRODUCTION**

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

47

48 While the T7SSa is heavily linked with Mycobacterial virulence, there is growing evidence that
49 the T7SSb plays an important role in bacterial antagonism (10, 11). In *Streptococcus*
50 *intermedius*, three T7SSb-secreted antibacterial effectors have been identified, including TelB,
51 an NADase, and TelC a lipid II phosphatase (12). The *Enterococcus faecalis* T7SSb also
52 mediates contact-dependent inhibition of some firmicute bacteria, and a bioinformatic analysis
53 in *Listeria monocytogenes* has identified over 40 potential antibacterial substrates of the T7SS
54 (13, 14).

55

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

66 island located at the 3' end of the *ess*/*T7SS* locus (10, 15, 16). These additional *esaG* copies
67 are genetically diverse, but share some core regions of similarity within the encoded amino
68 acid sequences (17, 18). Strings of up to nine non-identical *esaG* genes are also found in a
69 similar genomic location in *essC2*, *essC3* and *essC4* strains (10).
70
71 TspA is a second antibacterial toxin secreted by the *S. aureus* T7SS. TspA has a C-terminal
72 membrane-depolarising domain, and immunity from intoxication is provided by the membrane-
73 bound TsaI protein (19). The *tspA-tsaI* locus is encoded distantly from the T7 gene cluster,
74 and is found across all four *essC* variant strains. Similar to EsaD, the TspA toxin domain is
75 polymorphic, and all strains encode clusters of TsaI variants (between two and sixteen copies)
76 directly downstream of *tspA* (19).
77
78 It is common for bacteria to encode repertoires of immunity proteins for protection against
79 polymorphic effector proteins. For example, bacteria in the gut accumulate immunity genes in
80 genomic islands that provide protection against type VI secretion system effectors (20). This
81 includes orphan immunity genes in strains that do not encode the cognate effector protein, as
82 we observe for *esaG* genes in *S. aureus essC2*, *essC3* and *essC4* strains that do not contain
83 *esaD*. While it is common for immunity islands to carry many predicted immunity genes, it is
84 less common to see so many homologues of the same immunity gene clustered together. At
85 present, little is known about the origin of the T7SS immunity repertoires.
86
87 The *Staphylococcus aureus* strain NCTC8325, and its derivatives, are commonly used as a
88 model strains for laboratory studies (21). NCTC8325 was initially isolated from a corneal ulcer
89 and has been used extensively for the propagation of the lysogenic bacteriophage
90 *Staphylococcus* virus 11 (Previously Phage 47) (22). Since NCTC8325 also carries two other
91 lysogens in its genome, *Staphylococcus* viruses 12 and 13 (23) UV-induced curing was used
92 to rid the strain of prophage, to give strain RN450 (24) (Fig 1a). RN450 has been used to

93 construct many laboratory strains over the years, including RN6390, which is one of the key
94 strains we have employed for T7SS secretion studies (e.g. (25, 26)).
95
96 During our analysis, we found that there were two separate genome sequences reported for
97 NCTC8325. When comparing the first sequence, published in 2006 (GenBank accession
98 number CP000253; (27)), with the later sequence (uploaded in 2018 by the Wellcome Trust
99 Sanger Institute to GenBank; accession number LS483365) we noted that there were a
100 number of differences, including in *esaG* repertoire encoded at the 3' end of the *ess* gene
101 cluster (Fig 1b). In this study we used whole genome sequence analysis to identify which of
102 these two genome sequences was most closely related to RN6390. We subsequently used
103 gene phylogeny and cluster analyses to identify the processes that drive the accumulation of
104 T7SS immunity genes. Our findings indicate that intergenic recombination is a major factor in
105 the evolution of *esaG* genes. We also noted that a similar process drives the evolution of *tipC*
106 genes, which encode Streptococcal TelC immunity proteins (12). By contrast, while there is
107 some evidence of intergenic recombination within *tsaI* clusters, it does not appear to be the
108 main mechanism of evolution for this gene family.

109 **METHODS**

110

111 **Strains, genome sequencing and phylogenic analysis of prophage**

112 Strain RN6390 was obtained from Professor Jan Maarten van Dijl (University of Groningen,
113 NL). All other strains were obtained from Professor José Penadés (Imperial College, UK). All
114 genome sequencing was carried out by MicrobesNG (Birmingham, UK) with enhanced
115 genome sequencing for RN6390 and standard whole genome service for RN25 and RN450;
116 sequences are available at NCBI under accession numbers CP090001.1,
117 JAJSOX000000000.1 and JAJSOY000000000.1, respectively. Whole genome alignment was
118 executed using progressiveMauve (28). SNP calling was carried out using Snippy v4.6.0 (29).
119 To determine the taxonomy of the phage identified in RN6390 strain, its nucleotide sequence
120 was submitted to VIPTree browser with default parameters (30). From the resulting proteomic
121 tree, phage genomes associated with *S. aureus* species were selected to generate the tree in
122 Fig 2.

123

124 **Gene and protein alignment**

125 Nucleotide sequences were obtained from NCBI and aligned with MAFFT v7.489 (31). Amino
126 acid sequences were obtained from NCBI and aligned using MUSCLE v3.8.1551 (32). To
127 construct similarity plots for both nucleotide and amino acid sequences, Plotcon
128 (https://www.bioinformatics.nl/cgi-bin/emboss/plotcon) was executed using aligned
129 sequences, with a window size of 5. For *esaG4,* the full pseudogene was used for the
130 nucleotide alignment. For the alignment of the EsaG4 amino acid sequence, the two predicted
131 open reading frames (ORFs) were used. For Plotcon analysis on large input sequences,
132 alignments were manually curated to remove partial sequences and pseudogenes.

133

134 **Recombination Prediction**

135 Aligned nucleotide sequences were opened in the RDP4 software (33) and Run All selected.
136 TrimAl v1.2 (34) was used to remove the unaligned 5' region of *tsaI* genes before RDP4
137 analysis was carried out.
138
139 **Gene phylogeny**
140 Maximum likelihood trees for nucleotide sequences were built with IQTREE v 2.1.4 (35), with
141 1000 ultrafast bootstraps. Trees were visualised and annotated using iTOL (36).
142
143 **Comparison of genetic loci and gene cluster analysis**
144 T7SS-encoding loci from each variant were subjected to pairwise comparisons with the
145 RN6390 *T7SS* locus using BLAST (37). A similar approach was used to compare
146 *Staphylococcus* phage RN6390 with *Staphylococcus* phage SAP40 genomes. The report
147 produced by BLAST was used as a comparison file in genoPlotR package (38) within RStudio
148 (39) to plot regions of similarities between the two loci. FlaGs (40) was used to assess the
149 variation in number of *tsaI* repeats across *S. aureus* strains. Examples were selected to
150 represent diversity at this locus.
151
152 **Construction and analysis of a plasmid database**
153 A database containing all available plasmid sequences was built from PLSDB (https://ccb-
154 microbe.cs.uni-saarland.de/plsdb/plasmids/) (41). Subsequently, Hmmsearch from the
155 HMMER suite (v 3.3.2) (42) was used to identify *esaG* or *tsaI* genes encoded within the
156 database. For each *esaG* or *tsaI* copy identified, efetch from the entrez-utilities (43) was used
157 to retrieve the specific assembly ID where the genes were encoded. EsaG and TsaI proteins
158 identified as encoded on plasmids together with their corresponding assembly ID, were then
159 analysed using FlaGs to define the specific gene neighbourhood of *esaG* and *tsaI* on these
160 assemblies.

161 **RESULTS**

162

163 **Analysis of the RN6390 genome.**

164 Two genome sequences are available for NCTC8325, here designated NCTC8325-Oklahoma
165 (GenBank accession number CP000253) and NCTC8325-Sanger (GenBank accession
166 number LS483365). The genomes of NCTC8325-Oklahoma and NCTC8325-Sanger were
167 aligned to identify differences between the strains (Table S1). Following SNP calling and
168 manual curation we identified a total of 78 SNPs and other small polymorphisms between
169 these two strains. We also identified a duplication of 16S and 23S ribosomal RNA genes in
170 NCTC8325-Oklahoma relative to NCTC8325-Sanger (Table S1). A further notable difference
171 is that NCTC8325-Sanger carries six additional *esaG* genes, relative to NCTC8325-
172 Oklahoma, at the *ess* cluster (Table S1, Fig 1b).

173

174 In order to identify which NCTC8325 strain is the likely precursor strain of RN6390, we carried
175 out whole genome sequencing of RN6390 (GenBank accession number CP090001.1) and
176 aligned it with the genomes of NCTC8325-Oklahoma and NCTC8325-Sanger, respectively.
177 Differences between RN6390 and each NCTC8325 strain are recorded in SNP tables (Tables
178 S2 and S3). As expected, a major difference found between both NCTC8325 genomes and
179 RN6390 is the presence of three prophages in the NCTC8325 genomes which were
180 successively cured out during the construction of the RN6390 progenitor, RN450 (Fig 1a). A
181 prophage, *Staphylococcus* phage 6390 (φ6390) has been identified previously in the genome
182 of RN6390, integrated in the *rpmF* gene (44, 45). This is at a different locus from the three
183 prophage loci found in the two NCTC8325 genome sequences. To determine whether φ6390
184 is related to any of the prophages cured from NCTC8325, the genome sequence was
185 extracted and used to determine its taxonomy. As shown in Fig 2, φ6390 is a distinct prophage
186 to those cured from NCTC8325. Analysis of φ6390 using blastn gave an almost identical
187 match to *Staphylococcus* phage SAP40 (GenBank accession number: MK801683.1, 99%

188 coverage and 99.98% identity). As φ6390 is not present in RN4220, this suggests that it was
189 introduced during genetic manipulation steps performed following generation of RN450.
190
191 Further analysis of the RN6390 genome sequence revealed that it aligned more closely to
192 NCTC8325-Sanger than NCTC8325-Oklahoma, and also carries the six additional *esaG*
193 genes found in NCTC8325-Sanger. This suggests that NCTC8325-Sanger is the likely
194 progenitor of the 8325 lineage. To confirm this, we sequenced the genomes of the
195 intermediate strains RN25 and RN450 (Genbank accession numbers JAJSOX000000000.1
196 and JAJSOY000000000.1, respectively), as they are directly on the lineage of RN6390. Both
197 of these strains share the additional copies of *esaG* and other genomic features of NCTC8325-
198 Sanger, confirming that this strain is likely to be the true parent of RN6390.
199
200 The origin of the NCTC8325-Oklahoma strain from GenBank accession CP000253 is
201 described as the 'University of Oklahoma Health Sciences Center'. This strain is clearly closely
202 related to NCTC8325-Sanger, but may potentially have accumulated mutations through serial
203 passaging within a laboratory setting. Strikingly, the biggest difference between the two strains
204 apart from the ribosomal RNA gene duplication is the loss of the six *esaG* homologues.
205
206 **Intergenic recombination is a driving force in the evolution of *esaG* genes.**
207 To determine how the six *esaG* copies may have been lost in NCTC8325-Oklahoma, we
208 examined this region of the chromosome in more detail. RN6390 encodes 12 homologues of
209 *esaG*, all of them at the *ess* gene cluster. The first copy is found directly downstream of *esaD*
210 and encodes the cognate immunity protein for the nuclease toxin (Fig 3a)(15). We have named
211 this *esaG1*. This is followed by three genes encoding hypothetical proteins (SAOUHSC_00270
212 _00271 and _00272), before a stretch of four further *esaG* homologues (*esaG2* - *esaG5*). This
213 is followed by a further homologue of *SAOUHSC_00270* (annotated as *SAOUHSC_00270b*
214 on Fig 3a, sharing 96.49% identity with *SAOUHSC_00270*) and then an additional seven *esaG*
215 homologues (*esaG6* – *esaG12*; Fig 3a). Note that while *esaG4* is annotated as a pseudogene,

216 due to a short additional stretch of nucleotides close to the centre of the gene introducing a
217 premature stop and subsequent start codon, it actually encodes two smaller ORFs (Fig 3a). A
218 genome alignment using progressiveMauve predicts that the genomic deletion in NCTC8325-
219 Oklahoma spans *esaG2 – esaG7.*
220
221 To assess variation between the 12 homologues of EsaG, the amino acid sequences were
222 aligned (Fig 3b), and regions of homology between the proteins were analysed using Plotcon
223 (Fig 4a). Numerous regions of sequence homology were observed, including two major blocks
224 of high sequence homology at amino acids 13-31 and 84-119. To assess whether these were
225 general features of *S. aureus* EsaG proteins, we downloaded approximately 4,000 *S. aureus*
226 EsaG sequences from RefSeq and collectively analysed them using Plotcon. Fig S1a indicates
227 that a similar profile of homology is seen across all EsaG proteins.
228
229 We next examined the intergenic regions between the RN6390 *esaG* homologues. Strikingly,
230 we noted that most of the intergenic regions were of a very similar length, other than when
231 they directly preceded a non-*esaG* gene (for example *SAOUHSC_00270b*). They also share
232 a high degree of homology (Fig S2a). When we undertook Plotcon analysis on the *esaG*
233 genes, including the 3' intergenic regions (Fig 2a, Fig S2b), we noted the same two major
234 blocks of homology that we had seen from the amino acid analysis, but in addition a third block
235 encompassing the end of the gene and the downstream intergenic region (Fig 3a, Fig S2b,
236 Fig S3).
237
238 Given the substantial levels of homology between the *esaG* genes, we used the recombination
239 prediction software, RDP4, to determine whether there had been recombination events
240 between the genes. The RDP4 output, shown in Table S4 and summarised Fig 4b, predicts
241 with high significance that there have been extensive recombination events within most (but
242 not all) of these genes. As shown in Fig 4b, recombination appears to occur at the three points
243 within the genes that correspond to the regions of high nucleotide sequence homology (Fig

244 4b, Table S4). To support these findings, we constructed a maximum-likelihood tree for the
245 12 *esaG* genes (Fig 4c). No recombination is predicted within *esaG3*, which is genetically
246 distant from other *esaG* homologues. Conversely, *esaG2* and *esaG5* appear to vary only in
247 their central regions, and these cluster closely on the tree (Fig 4c), consistent with the RDP4
248 output.
249
250 Based on the RDP4 results, we built a schematic representation showing the homologous
251 regions of the *esaG* genes that are likely involved in the recombination events (Fig 4d).
252 Specific regions of the genes seem to share high homology with others, for example, the mid-
253 section of many of the homologues share high homology to equivalent sections of *esaG1*
254 (coloured teal). Conversely, genes such as *esaG3* and *esaG6* appear to be much more
255 diverse. Based on our observations, it is probable that a recombination event between the first
256 conserved regions of *esaG2* and *esaG8* was responsible for the loss of the seven genes at
257 this locus in NCTC8325-Oklahoma. This is corroborated by an alignment of the nucleotide
258 sequences of *esaG2* and *esaG8 with SAOUHSC_00274* (Fig S2c). The alignment indicates
259 that *SAOUHSC_00274* is mosaic comprising the 5' region of *esaG2* with the middle and 3'
260 sections of *esaG8*, consistent with recombination at homology block 1 (Fig S2d).
261
262 ***esaG* recombination in an epidemic strain of *S. aureus***
263 USA300 is a methicillin-resistant *S. aureus essC1* strain, and a dominant cause of community-
264 acquired *S. aureus* infection in the USA (46). A recent study analysed the community spread
265 and evolution of a USA300 variant during a New York outbreak (47). Comparing the whole-
266 genome sequences of the epidemic lineage showed that these strains carry only seven *esaG*
267 genes in comparison to the ten copies in the closely related USA300 FPR3757. Using strain
268 BKV_2 as a representative of the outbreak lineage, we aligned the region spanning from *esaE*
269 to *SAUSA300_0303* with USA300 FPR3757. The alignment showed almost complete
270 sequence identity, other than in a region spanning from the middle of *SAUSA300_0295* to the
271 middle of *SAUSA300_0299* (Fig S4a) which was absent from BKV_2. When the nucleotide

272 sequences of *SAUSA300_0295* and *SAUSA300_0299* were aligned with *esaG4* of BKV_2,
273 *esaG4* was again seen to be mosaic, with most of the gene being identical to
274 *SAUSA300_0295* but the 3' end showing 100% identity to *SAUSA300_0299* (Fig S4b). This
275 is consistent with recombination between homology block 2 of *SAUSA300_0295* and
276 *SAUSA300_0299*, with loss of the intervening DNA.
277
278 ***esaG* diversity across *S. aureus essC2*, *essC3* and *essC4* variants.**
279 Although *S. aureus essC2*, *essC3* and *essC4* variants do not encode EsaD, these strains all
280 accumulate *esaG* genes at the 3' end of their T7SS-encoding loci. To ascertain whether the
281 homologues of *esaG* encoded in these strains are close relatives of those found in RN6390,
282 we analysed the *esaG* genes encoded in strains ST398, MRSA252 and HO 5096 0412, as
283 representatives of *essC2*, *essC3* and *essC4* variants, respectively. The sequence of the *T7SS*
284 loci for each of these strains was subjected to pairwise comparison, using BLAST, with the
285 RN6390 locus, and further analysed using genoPlotR (48) to produce a graphic representation
286 of the alignment of these regions (Fig 5). In this output, regions of homology are highlighted
287 by red-connecting blocks, and the colour intensity of these blocks reflects the percentage
288 identity found between the two compared regions.
289
290 As shown in Fig 5a, the *essC2* strain, ST398, shares higher homology with the RN6390 *T7SS*
291 cluster than either MRSA252 or HO 5096 0412. This strain harbours two intact copies of *esaG*
292 (*SAPIG_0310* and *_0311*), and one pseudogene (covering the two small genes *SAPIG_0314-*
293 *0315*, and herein referred to as a single *SAPIG_0314* pseudogene). We used RDP4 analysis
294 with *esaG1-12* from RN6390, and phylogenetic tree construction, to determine whether there
295 were any regions of shared homology. The analysis showed that *SAPIG_0314-0315* has the
296 highest homology to *esaG12* (Fig 5a) and appears to cluster in the same branch of the
297 phylogenetic tree (Fig 5d). No recombination events are predicted for this pseudogene,
298 suggesting this is most likely a copy of *esaG12* that has accrued mutations (Fig S5a).
299 Conversely, *SAPIG_0310* and *SAPIG_0311* have much lower homology to the *esaG*

300 homologues in RN6390 (Fig 5a), although recombination events are predicted with *esaG4* for
301 *SAPIG_0310* and with *SAPIG_0314-315* for *SAPIG_0311* (Fig S5a).
302
303 The *essC3* strain, MRSA252, has four tandem *esaG* homologues at its *ess* locus, *SAR_0293*
304 - *SAR_0296* (Fig 5b). These are predicted by BLAST to share the highest homology with
305 RN6390 *esaG9-esaG12*. However, these homologues do not cluster together in the
306 phylogenetic tree (Fig 5d), and for example *SAR_0295*, predicted by genoPlotR to be a
307 homologue of *esaG11*, clusters with *esaG3.* Recombination events are also detected for all of
308 the *esaG* genes in MRSA252 (Fig S5b).
309
310 The *essC4* strain, HO 5096 0412, harbours two *esaG* copies, *SAEMRSA15_02570* and
311 *SAEMRSA15_02580*, which are predicted by BLAST to share highest homology with *esaG11*
312 and *esaG12*, respectively (Fig 5c). However, phylogenetic analysis indicates that
313 *SAEMRSA15_02580* clusters with *esaG11* as opposed to *esaG12*, and *SAEMRSA15_02570*
314 does not cluster closely with any of the RN6390 *esaG* genes (Fig 5d). RDP4 analysis also
315 predicts recombination events for both genes in HO 5096 0412 (Fig S5c).
316
317 In summary, whilst *SAPIG_0314-0315* is a copy of *esaG12*, the remaining *esaG* homologues
318 found in the three representative strains of *essC* variants 2, 3 and 4 all appear to be
319 recombinants. Some of the recombination events are between homologues present in these
320 strains (Tables S5 - S7), suggestive of common ancestry. However, for other recombination
321 events, the parent *esaG* gene is unknown and is not present in any of these representative
322 strains.
323
324 **Accumulation of *tsaI* genes in the RN6390 genome.**
325 TspA is a T7SS-secreted antibacterial toxin that is highly conserved across all *essC* variant
326 strains (19). In all of these strains it is encoded away from the T7SS gene cluster, at a genomic
327 location bounded by *SAOUHSC_00583* and *iolS* (*SAOUHSC_00603*). The toxic activity of

328 TspA is neutralised by TsaI, a membrane protein of the DUF443 family (19). Multiple copies
329 of *tsaI* genes are encoded downstream of *tspA* (Fig 6, Fig S6), and in RN6390 there are 11
330 copies, *tsaI1* – *tsaI11* (Fig 6a). A small pseudogene, encoded by *SAOUHSC_00600* shares
331 homology to part of the toxin region of TspA and is also found at this locus.
332
333 We wondered whether recombination events, similar to those we have observed for *esaG*
334 genes, also contributed to the evolution of *tsaI* genes. An alignment of the amino acid
335 sequences for RN6390 TsaI proteins (Fig 6b), shows that there is much greater sequence
336 variability between these proteins than the EsaG homologues. This is particularly apparent in
337 the C-terminal region of the protein, with only limited sequence identity observed between TsaI
338 homologues. Similar variability was observed in a representative alignment of around 3000
339 TsaI sequences (Fig S1b). Much greater variability was also observed in the *tsaI* intergenic
340 regions, in both length and DNA sequence compared with *esaG* (Fig 6c).
341
342 To identify the regions of highest homology at both the protein and DNA level, we ran Plotcon
343 analysis of the 11 *tsaI* genes and their encoded ORFs. Unlike EsaG, TsaI homologues have
344 only a single region of high similarly, covering approximately the first 75 amino acids, which is
345 also mirrored at the DNA level (Fig 7a, Fig S7). As only one block of high homology is detected
346 and there is a high degree of sequence variability in the *tsaI* intergenic regions, recombination
347 within individual genes is unlikely. To analyse this, we used RDP4 to predict recombination
348 events within the 11 homologues of *tsaI* genes (Fig 7b). Far fewer potential recombination
349 events were predicted than for *esaD* genes, and with lower probability, which could arise from
350 evolutionary processes other than recombination (Table S8).
351
352 ***tsaI* genes are found on Staphylococcal plasmids**
353 Horizontal gene transfer (HGT) is a major mechanism for the movement of genomic material
354 between bacteria (49). Plasmids are one of the key drivers of HGT and help to mediate the
355 spread of resistance genes among bacterial populations (e.g. 50, 51). To investigate whether

356 the *esaG* and/or *tsaI* toxin resistance genes may also be disseminated by plasmids, we
357 constructed a database of bacterial plasmids as described in the methods section and
358 interrogated this for the presence of *tsaI* and *esaG* genes. We identified a single
359 Staphylococcal plasmid carrying an *esaG* gene and five Staphylococcal plasmids that encode
360 one or more TsaI homologues (Fig 8). Two of the five plasmids encoding TsaI (pCAPBN21
361 and an unnamed plasmid from *Staphylococcus caprae* 26D) also encode a full-length TspA
362 along with the two WXG100-like proteins (DUF5344 and DUF3958 family proteins) that have
363 been proposed to serve as TspA-specific T7SS targeting factors (10). A further two plasmids
364 (pSB1-57-a and an unnamed plasmid from *Staphylococcus warneri* SWO) code for a fragment
365 of TspA alongside TsaI, with the SWO plasmid also encoding two further *tsaI* genes. An
366 unnamed plasmid from *Staphylococcus simulans* MR1 encodes an orphan *tsaI* with no
367 detectable *tspA* remnant. The SWO plasmid is particularly interesting as this carries further
368 T7-related genes including the *esaG* we identified along with a portion of *esaD*, and a fragment
369 of an HNH-nuclease gene along with a SM1/KNR4 protein encoding gene (a family implicated
370 as a nuclease immunity protein (52) and found in Staphylococcal T7SS immunity gene islands
371 (10)). For three of the plasmids, the immunity genes are close to recombinase genes, which
372 may provide a mechanism for their accumulation, and four of them carry nearby *IS* elements
373 that could facilitate their transfer to the chromosome without the need for homologous
374 recombination.
375
376 **Intergenic recombination between *tipC* immunity genes in *Streptococcus*.**
377 Three T7SS-secreted antibacterial toxins have been identified in *Streptococcus intermedius*.
378 TelA and TelB are both cytoplasmic-acting toxins neutralised by the TipA and TipB immunity
379 proteins, respectively (12). Our genome analysis indicates that strains generally encode only
380 a single *tipA* gene, while *tipB* is found in up to five copies. The third *S. intermedius* toxin is
381 TelC, a lipid II phosphatase. Protection from TelC toxicity is provided by TipC, a membrane-
382 bound immunity protein that faces the extracellular space (12, 53). Klein *et al*. reported that
383 the number of *tipC* genes encoded at the *telC* locus was highly variable between strains (53).

384 We used gene neighbourhood analysis across Streptococcal genomes to compare the
385 number of *tipC* genes present at the *telC* gene cluster, identifying 15 of them in a strain of
386 *Streptococcus mitis* BCC08 (Fig 9a).
387
388 Alignment of the *S. mitis* BCC08 TipC sequences and their encoding DNA (Fig 9b, Fig S8)
389 showed two regions of high sequence conservation close to the start and end, with a central
390 region of much higher sequence variability (Fig 10a). Using RDP4 to screen for recombination,
391 at least five recombination events were predicted between these genes (Fig 10b, Table S9),
392 in each case almost certainly through the blocks of high homology we identified. To analyse
393 this further we constructed a maximum likelihood tree to compare *tipC* homology. Genes
394 *D8786_RS05910* and *D8786_RS0585* cluster closely in this tree, which corresponds to the
395 recombination event predicted between these two genes (Fig 10b, Table S9). Likewise
396 *D8786_RS05920*, which is predicted to be the major parent to *D8786_RS05865* (Table S9)
397 also clusters phylogenetically with this gene. We conclude that similar to *esaG*, intergenic
398 recombination drives the evolution of *tipC* immunity gene repertoires.

399 **DISCUSSION**

400 Through comparative genome analysis we noted that the *esaG* copy number is distinctly
401 different in closely related strains of *S. aureus*, a finding that had also been previously
402 described within the NCTC8235 lineage (17). To investigate how copy number variability may
403 arise, we undertook sequence analysis of EsaG proteins and their encoding DNA, including
404 their 3' flanking regions. We found three blocks of highly conserved nucleotide sequence, a
405 large central one of approximately 100 nucleotides in length, and a 5' and 3' block both of
406 around 55 nucleotides each. Homologous recombination occurs at regions of high homology
407 within nucleotide sequences (Reviewed in 54). The minimum length requirement for efficient
408 recombination in *S. aureus* is unclear, but stretches of 40-70 nucleotides have been reported
409 for other bacteria (e.g. 55 - 57). Using RDP4 to predict recombination within *esaG* genes, we
410 found strong evidence for recombination, corresponding to events within each of the three
411 homology blocks we identified. Furthermore, our analysis revealed that the loss of six *esaG*
412 genes in NCTC8325-Oklahoma arose from recombination across homology block 1, whereas
413 three *esaG* genes have been lost in an outbreak strain of USA300 through recombination
414 across homology block 2.

415

416 Previous work has reported that the *S. aureus* tandem-like lipoproteins, encoded on the νSaα
417 island, also show extensive copy number variation across strains (58, 59). Similar analysis to
418 that reported here showed that each *lpl* gene shares a stretch of approximately 130
419 nucleotides of high homology in its central region. Recombination was demonstrated to occur
420 between the central conserved region of one gene and the same region of the neighbouring
421 gene (57), thus spanning the 3' portion of gene 1, the intergenic region and the 5' region of
422 gene 2.

423

424 To investigate whether intergenic recombination might represent a general mechanism for the
425 evolution of T7SS immunity gene families, we examined the organisation of the *tipA*, *tipB* and
426 *tipC* immunity genes in Streptococci. It has previously been noted that *tipC* copy number is

427 highly variable in Streptococcal genomes (53), and our analysis identified that up to 15 copies
428 of *tipC* could be present. Examination of recombination events within *tipC* revealed that
429 intergenic recombination is also a feature, and that it primarily occurs between homologous
430 blocks of sequence identity at the start and end of the genes. The structure of TipC reveals
431 that it has seven beta strands forming a concave face, with three alpha helices made up from
432 the N- and C-terminal regions of the protein (53, Fig 10d). Analysis of colicin DNase toxins
433 and their immunity proteins has shown that sequence divergence between related toxins and
434 immunities tends to concentrate at the binding interface (60). In agreement with this, site-
435 directed mutagenesis has strongly implicated the concave face as the region that binds to
436 TelC, and this region of TipC shows the highest level of sequence divergence (Fig 10d; 53).
437 We speculate that for EsaG the regions of high homology are not directly involved in toxin
438 binding, but may provide a structural framework on which amino acid substitutions in the
439 variable regions accumulate to alter the toxin binding specificity. Mechanisms to allow rapid
440 evolution of immunity proteins are likely to be essential to allow strains to rapidly acquire
441 resistance to novel toxin variants.
442
443 TspA is a second polymorphic toxin encoded by *S. aureus*. Protection from its membrane-
444 depolarising activity is provided by the immunity protein TsaI (19). TsaI is a polytopic
445 membrane protein predicted to have five transmembrane domains. As with *esaG*, multiple *tsaI*
446 genes are found in *S. aureus* genomes, however, our analysis has indicated that there is little
447 evidence of recombination between them. We found only a single block of high nucleotide
448 sequence homology at the 5' end of *tsaI* genes. This encodes approximately the first 75 amino
449 acids of TsaI, which would encompass the first two transmembrane domains. At present it is
450 not known whether TsaI neutralises TspA though direct interaction, or through the
451 sequestering of a membrane-bound partner protein with which TspA must interact to facilitate
452 its insertion or folding (or a combination of both of these). In this context, membrane
453 permeabilising peptide bacteriocins produced by some Gram-positive bacteria require a
454 membrane-bound receptor, such as the membrane components of the mannose

455 phosphotransferase system, for their activity. In the case of the *Lactococcus lactis* lactococcin
456 A bacteriocin, the immunity protein LciA acts through formation a complex with both the
457 receptor protein and the bacteriocin (61). By analogy it is possible that the first two
458 transmembrane domains of TsaI interact with a candidate receptor, constraining their
459 sequence, whereas the remainder of the immunity protein binds to the toxin and is therefore
460 under diversifying selection. At present it is unclear how *tsaI* gene clusters evolve, although
461 we did note that there was evidence for carriage of *tsaI* genes on Staphylococcal plasmids,
462 which may facilitate gene movement within and between strains. Further work would be
463 required to clarify the mechanisms that drive *tsaI* diversity.

464 **ACKNOWLEDGEMENTS**

465 We thank Prof Jan-Maarten van Dijl (University of Groningen, NL) for providing us with strain
466 RN6390, and Prof José Penadés (Imperial College, UK) for providing all of the other strains
467 in this study. We thank Prof Victor J. Torres and Dr Alejandro Pironti (NYU Langone, USA) for
468 providing the genome sequence of USA300 BKV_2. This work was supported by Wellcome
469 through Investigator Award 110183/A/15/Z to TP and Sir Henry Wellcome Postdoctoral
470 Fellowship 218622/Z/19/Z to GM. SG is funded by the Newcastle-Liverpool-Durham BBSRC
471 DTP2 Training Grant, project reference number BB/M011186/1.

472

473 **FIGURE LEGENDS**

474

475 **Figure 1. Lineage of NCTC8325 daughter strains.** a. A selection of daughter strains of the
476 NCTC8325 lineage and how they were generated. RN25 (also called 8325-3) was generated
477 from NCTC8325 following UV curing of prophages 11 and 12 (23). RN450 (also called 8325-
478 4) was generated from RN25 by a second round of UV exposure, to cure prophage 13 (23).
479 Methylnitronitrosoguanidine (MNNG)-mediated mutagenesis of RN450 was used to generate
480 RN4220, which is restriction deficient and capable of accepting foreign DNA (62). RN450 was
481 separately transduced with pRN3032, a Tn551 donor plasmid to generate RN1478 (63)
482 ISP479 is a cadmium-resistant revertant of RN1478 (64). RN6390 was generated from ISP479
483 by curing of pRN3032 (24). b. The genetic layout of the *ess* locus in NCTC8325- Oklahoma
484 and NCTC8325-Sanger. The dashed lines and bar represent the region that is missing from
485 NCTC8325-Oklahoma, as identified in this study.

486

487 **Figure 2. Analysis of *Staphylococcal* phage 6390.** A phylogenetic tree of φ6390 against a
488 database of *Staphylococcal* phages generated using VIPtree (29). The positions of φ6390 and
489 *Staphylococcal* phages 11, 12 and 13 (excised from the NCTC8325 parent strain during
490 construction of RN6390) on the tree are indicated with red stars.

491

492 **Figure 3. Homologues of *esaG* encoded at the *ess* locus in RN6390.** a. Illustration of the
493 3' of the RN6390 *ess* locus which encodes the T7SS nuclease toxin, EsaD, its cognate
494 immunity protein, EsaG and 11 further homologues of EsaG (numbered *esaG2* – *esaG12*).
495 Genes *esaG2* to *esaG7* are indicated by a black bar, and are absent from NCTC8325-
496 Oklahoma, while the dotted line indicates the actual region lost by recombination. Note that
497 *esaG4* is shown in hatched shading because it is annotated as a pseudogene. However, it
498 does encode two predicted ORFs, EsaG4i and EsaG4ii. b. Sequence alignment of EsaG
499 homologues encoded by RN6390. The black boxes represent regions of high homology based
500 in this alignment.

501

502 **Figure 4. Recombination within the RN6390 *esaG* homologues.** a. Regions of high
503 homology across the RN6390 EsaG protein sequences (middle panel) and the corresponding
504 nucleotide sequences (bottom panel). The positions of blocks of high homology are shown in
505 grey shading and along with their relative positions along the gene sequence (top panel). The
506 basepair positions that define the conserved regions are taken from the nucleotide sequences
507 of *esaG1*. b. RDP4 was used to predict recombination events within the *esaG* homologues
508 encoded by RN6390. The identity of each gene is given in black at the left, with regions of
509 recombination labelled directly below, in the colour of the gene from which the recombinant
510 section originated. c. A maximum likelihood tree was generated for RN6390 *esaG* homologues
511 in IQTREE and visualised and annotated in iTOL. d. Illustration of regions of homology in the
512 *esaG* homologues in RN6390. Black bars represent conserved regions of the gene and the
513 variable regions have been assigned a colour and corresponding number. Homologous
514 regions are coloured with the same colour. Numbers were assigned based on the first gene
515 in the series that had the unique variable region. White hatched shading indicate a
516 pseudogene.

517

518 **Figure 5. Recombination events within the *esaG* genes encoded in representative**
519 ***essC2*, *essC3* and *essC4* variant strains.** a-c. The genes downstream of *essC* are different
520 between each of the four *essC* variants. The regions spanning *essC* to the conserved gene
521 *SAOUHSC_00279* were aligned between RN6390, and a. ST398 (*essC2* variant), b.
522 MRSA252 (*essC3* variant) and c. HO 5096 0412 (*essC4* variant). Alignments were visualised
523 using genoPlotR. d. A maximum likelihood tree constructed with IQTREE and annotated in
524 iTOL for all *esaG* homologues found across the four representative *essC* variant strains
525 RN6390, ST398, MRSA252 and HO 5096 0412.
526
527 **Figure 6. Homologues of *tsaI* encoded at the *tspA* locus of RN6390.** a. Genetic
528 arrangement of *tsaI* genes in RN6390. *rclA* and *iolS* are a conserved gene found flanking the
529 TspA locus in *S. aureus* strains, encoding a pyridine nucleotide-disulfide oxidoreductase and
530 an aldo-keto reductase, respectively. b. An alignment of the encoded TsaI homologues. The
531 black boxes represent regions of high homology based in this alignment. c. Alignment of the
532 intergenic region downstream of each *tsaI* gene.
533
534 **Figure 7. Assessing recombination events within *tsaI* homologues.** a. A single region of
535 high homology across the RN6390 TsaI protein sequences (middle panel) and the
536 corresponding nucleotide sequences (bottom panel). The numbers which dictate the limits of
537 the conserved region are taken form the nucleotide sequence of *tsaI1*. b. RDP4 was used to
538 predict recombination events within the *tsaI* homologues encoded in RN6390. Each gene is
539 labelled in black, with regions of recombination labelled directly below in the colour of the gene
540 from which the recombinant section originated.
541
542 **Figure 8. *tsaI* genes are carried on Staphylococcal plasmids.** The *tsaI*-encoding regions
543 of unnamed plasmids from *S. simulans* MR1 (accession NZ_CP015643), *S. caprae* 26D
544 (accession NZ_CP031272), *S. warneri* SWO (accession NZ_CP033101), along with plasmids
545 pCAPBN21 (accession NZ_CP042342) and pSB1-57-a (accession CP070965) are shown.

546

547 **Figure 9. Homologues of *tipC* encoded at the *telC* locus of *Streptococcus mitis* BCC08.**
548 a. Genetic arrangement of *tipC* genes in *S. mitis* BCC08. B. An alignment of the encoded TipC
549 homologues. The blue boxes represent regions of high sequence homology, and the dashed
550 line at the N-terminus of the aligned sequences indicates a probable lipoprotein signal peptide.
551

552 **Figure 10. Recombination within the *S. mitis* BCC08 *tipC* homologues.** a. Regions of high
553 homology across the *S. mitis* BCC08 TipC protein sequences (middle panel) and the
554 corresponding nucleotide sequences (bottom panel). The positions of blocks of high homology
555 are shown in grey shading, with their relative positions along the gene sequence (top). The
556 first 18 amino acids of TipC form a predicted lipoprotein signal sequence which is indicated by
557 pale grey shading. The basepair positions that define the conserved regions are taken from
558 the nucleotide sequences of *D8786_RS05940*. b. RDP4 was used to predict recombination
559 events within the *tipC* homologues. The identity of each gene is given in black at the left, with
560 regions of recombination labelled directly below, in the colour of the gene from which the
561 recombinant section originated. c. A maximum likelihood tree was generated for *tipC*
562 homologues in IQTREE and visualised and annotated in iTOL. d. The conserved (cyan) and
563 variable (orange) regions of TipC were mapped to the crystal structure of *S. intermedius* TipC2
564 (pdb:6DHX; 53).
565

566 **Figure S1. Similarity plots for representative EsaG and TsaI amino acid sequences.** All
567 available amino acid sequences for EsaG and TsaI were obtained from RefSeq. Sequences
568 were aligned and a similarity plot produced using plotcon for a. EsaG and b. TsaI.
569

570 **Figure S2. Homology in the intergenic regions downstream of *esaG*genes in RN6390.** a.
571 The intergenic regions found directly downstream of each *esaG* gene were aligned and
572 visualised using boxshade. b. Plotcon analysis of RN6390 *esaG* genes and their 3' intergenic
573 regions. Alignment of *esaG* nucleotide sequences including the downstream IGR for *esaG*

574 genes from RN6390. e*saG1, esaG5 and esaG12* were excluded due to having a longer
575 intergenic region (as seen in Fig S2a). c. Nucleotide sequence alignment
576 for *esaG2* (NCTC8325_00242) and *esaG8 (NCTC8325_00250)* from NCTC8325-Sanger
577 with *SAOUHSC_00274* from NCTC8325-Oklahoma. d. Schematic representation of the
578 mosaic nature of *SAOUHSC_00274*.
579
580 **Figure S3. Alignment of the nucleotide sequences of *esD1-esaD12*.** The blocks of high
581 sequence homology corresponding to Fig 4a are outlined in blue.
582
583 **Figure S4. A recombination event in an epidemic lineage of USA300 results in loss of**
584 **part of an *esaG* cluster and generation of a novel *esaG* gene.** a. The *esaD* locus of
585 USA300 FPR3757 and USA300 BKV_2. The dashed lines represent the region that is missing
586 from the epidemic strain, USA300 BKV_2, when compared to the USA300 FPR3757 type
587 strain. b. Nucleotide sequence alignment for *SAUSA300_0295* and *SAUSA300_0299* from
588 USA300 FPR3757 and *esaG4* from USA300 BKV_2. Coloured blocks indicated homology
589 between *BKV_2 esaG4* and the genes with which it is aligned.
590
591 **Figure S5. Assessing recombination events in *esaG* homologues from a representative**
592 ***essC2*, *essC3* and *essC4* strain.** Alignments of *esaG* homologues from RN6390 with a.
593 ST398 (*essC2* variant), b. MRSA252 (*essC3* variant) and c. HO 5096 0412 (*essC4* variant)
594 were analysed using RDP4 to analyse recombination events. Each gene is labelled in black,
595 with regions of recombination labelled directly below in the colour of the gene from which the
596 recombinant section originated.
597
598 **Figure S6. Representation of the variability in the numbers of *tsaI* homologues encoded**
599 **by *S. aureus* strains.** Homologues of TsaI1 were obtained from RefSeq and genes
600 neighbouring the *tsaI* genes were identified - a selection of strains were used to visualise the
601 variability in number of *tsaI* genes encoded at this locus.

602

603 **Figure S7. Alignment of the nucleotide sequences of *tsal1-tsal11* from RN6390.** The
604 block of high sequence homology corresponding to Fig 7a is outlined in blue.

605

606 **Figure S8. Alignment of the nucleotide sequences of *tipC* homologues from *S. mitis***
607 **BCC08.** The blocks of high sequence homology corresponding to Fig 9a are outlined in blue.
608 The region encoding the predicted lipoprotein signal peptide is indicated with a dashed line.

609

610 **Table S1. SNP table for NCTC_8325-Oklahoma compared to NCTC8325-Sanger.** Large
611 deletions and insertions are included below.

612

613 **Table S2. SNP table for NCTC_8325-Oklahoma compared to RN6390.**

614

615 **Table S3. SNP table for NCTC_8325-Sanger compared to RN6390.**

616

617 **Table S4. RDP4 output for recombination within the *esaG* genes from RN6390.** Predicted
618 recombination events are highlighted in blue. Events that may have occurred due to an
619 evolutionary process other than recombination are highlighted in yellow.

620

621 **Table S5. RDP4 output for recombination within the *esaG* genes from RN6390 and**
622 **ST398.** Predicted recombination events are highlighted in blue.

623

624 **Table S6. RDP4 output for recombination within the *esaG* genes from RN6390 and**
625 **MRSA252.** Predicted recombination events are highlighted in blue. Events that may have
626 occurred due to an evolutionary process other than recombination are highlighted in yellow.

627

628 **Table S7. RDP4 output for recombination within the *esaG* genes from RN6390 and HO**
629 **5096 0412.** Predicted recombination events are highlighted in blue. Events that may have
630 occurred due to an evolutionary process other than recombination are highlighted in yellow.
631
632 **Table S8. RDP4 output for recombination within the *tsaI* genes from RN6390.** Events that
633 may have occurred due to an evolutionary process other than recombination are highlighted
634 in yellow.
635
636 **Table S9. RDP4 output for recombination within the *tipC* genes from *S. mitis* BCC08.**
637 Predicted recombination events are highlighted in blue. Events that may have occurred due
638 to an evolutionary process other than recombination are highlighted in yellow.

639 **References**

640 1. **Hsu T, Hingley-Wilson SM, Chen B, Chen M, Dai AZ *et al.*** The primary mechanism of
641 attenuation of bacillus Calmette-Guerin is a loss of secreted lytic function required for invasion
642 of lung interstitial tissue. *Proc Natl Acad Sci U S A* 2003;100:12420-12425.
643
644 2. **Pym AS, Brodin P, Majlessi L, Brosch R, Demangel C, *et al.*** Recombinant BCG
645 exporting ESAT-6 confers enhanced protection against tuberculosis. Nat Med 2003;9:533-
646 539.
647
648 3. **Stanley SA, Raghavan S, Hwang WW, Cox JS.** Acute infection and macrophage
649 subversion by *Mycobacterium tuberculosis* require a specialized secretion system. *Proc Natl*
650 *Acad Sci U S A* 2003;100:13001-13006.
651
652 4. **Akpe San Roman S, Facey PD, Fernandez-Martinez L, Rodriguez C, Vallin C *et al*.** A
653 heterodimer of EsxA and EsxB is involved in sporulation and is secreted by a type VII secretion
654 system in *Streptomyces coelicolor*. *Microbiology* 2010;156:1719-1729.
655
656 5. **Burts ML, Williams WA, DeBord K, Missiakas DM.** EsxA and EsxB are secreted by an
657 ESAT-6-like system that is required for the pathogenesis of *Staphylococcus aureus* infections.
658 *Proc Natl Acad Sci U S A* 2005;102:1169-1174.
659
660 6. **Bunduc CM, Fahrenkamp D, Wald J, Ummels R, Bitter W *et al*.** Structure and dynamics
661 of a mycobacterial type VII secretion system. *Nature* 2021;593:445-448.
662
663 7. **Beckham KSH, Ritter C, Chojnowski G, Ziemianowicz DS, Mullapudi E *et al*.** Structure
664 of the mycobacterial ESX-5 type VII secretion system pore complex. *Sci Adv*
665 2021;7:eabg9923.
666
667 8. **Abdallah AM, Gey van Pittius NC, Champion PA, Cox J, Luirink J, *et al.*** Type VII
668 secretion--mycobacteria show the way. *Nat Rev Microbiol* 2007;5:883-891.
669
670 9. **Zoltner M, Ng WM, Money JJ, Fyfe PK, Kneuper H, *et al.*** EssC: domain structures inform
671 on the elusive translocation channel in the Type VII secretion system. *Biochem J* 2016;
672 473:1941-1952.
673
674 10. **Bowman L, Palmer T.** The Type VII Secretion System of *Staphylococcus*. *Annu Rev*
675 *Microbiol* 2021;75:471-494.

676
677 11. **Tran HR, Grebenc DW, Klein TA, Whitney JC.** Bacterial type VII secretion: An important
678 player in host-microbe and microbe-microbe interactions. *Mol Microbiol* 2021;115:478-489.
679
680 12. **Whitney JC, Peterson SB, Kim J, Pazos M, Verster AJ, *et al.*** A broadly distributed toxin
681 family mediates contact-dependent antagonism between gram-positive bacteria. *Elife*. 2017;
682 6:e26938.
683
684 13. **Chatterjee A, Willett JLE, Dunny GM, Duerkop BA.** Phage infection and sub-lethal
685 antibiotic exposure mediate *Enterococcus faecalis* type VII secretion system dependent
686 inhibition of bystander bacteria. *PLoS genet* 2021;17:e1009204.
687
688 14. **Bowran K, Palmer T.** Extreme genetic diversity in the type VII secretion system of Listeria
689 monocytogenes suggests a role in bacterial antagonism. *Microbiology* 2021;167: doi:
690 10.1099/mic.0.001034
691
692 15. **Cao Z, Casabona MG, Kneuper H, Chalmers JD, Palmer T.** The type VII secretion
693 system of *Staphylococcus aureus* secretes a nuclease toxin that targets competitor bacteria.
694 *Nat Microbiol* 2016;2:16183.
695
696 16. **Warne B, Harkins CP, Harris SR, Vatsiou A, Stanley-Wall N, *et al.*** The Ess/Type VII
697 secretion system of *Staphylococcus aureus* shows unexpected genetic diversity. *BMC*
698 *Genomics* 2016;17:222.
699
700 17. **Baek KT, Frees D, Renzoni A, Barras C, Rodriguez N *et al*.** Genetic variation in the
701 *Staphylococcus aureus* 8325 strain lineage revealed by whole-genome sequencing. PLoS
702 ONE. 2013;8:e77122.
703
704 18. **Ohr RJ, Anderson M, Shi M, Schneewind O, Missiakas D.** EssD, a Nuclease Effector
705 of the *Staphylococcus aureus* ESS Pathway. *J Bacteriol* 2016;199:e00528-16.
706
707 19. **Ulhuq FR, Gomes MC, Duggan GM, Guo M, Mendonca C, *et al.*** A membrane-
708 depolarizing toxin substrate of the *Staphylococcus aureus* Type VII secretion system mediates
709 intra-species competition. *Proc Natl Acad Sci U S A* 2020;117:20836-20847.
710
711 20. **Ross BD, Verster AJ, Radey MC, Schmidtke DT, Pope CE *et al*.** Human gut bacteria
712 contain acquired interbacterial defence systems. *Nature* 2019;575:224-228.

713
714 21. **Novick RP.** Genetic systems in staphylococci. *Methods Enzymol* 1991;204:587-636.
715
716 22. **Herbert S, Ziebandt AK, Ohlsen K, Schafer T, Hecker M *et al*.** Repair of global
717 regulators in *Staphylococcus aureus* 8325 and comparative analysis with other clinical
718 isolates. *Infect Immun* 2010;78:2877-2889.
719
720 23. **Novick R.** Properties of a cryptic high-frequency transducing phage in *Staphylococcus*
721 *aureus*. Virology 1967;33:155-166.
722
723 24. **Peng HL, Novick RP, Kreiswirth B, Kornblum J, Schlievert P.** Cloning,
724 characterization, and sequencing of an accessory gene regulator (*agr*) in *Staphylococcus*
725 *aureus*. *J Bacteriol* 1988;170:4365-4372.
726
727 25. **Kneuper H, Cao ZP, Twomey KB, Zoltner M, Jäger F, *et al.*** Heterogeneity in ess
728 transcriptional organization and variable contribution of the Ess/Type VII protein secretion
729 system to virulence across closely related *Staphylocccus aureus* strains. *Mol Microbiol*
730 2014;93:928-943.
731
732 26. **Casabona MG, Kneuper H, Alferes de Lima D, Harkins CP, Zoltner M *et al*.** Haem-iron
733 plays a key role in the regulation of the Ess/type VII secretion system of *Staphylococcus*
734 *aureus* RN6390. *Microbiology* 2017;163:1839-1850.
735
736 27. **Gillaspy AF, Worrell V, Orvis J, Roe BA, Dyer DW, Iandolo JJ.** *Staphylococcus aureus*
737 NCTC8325 genome. In: Fischetti V, Novick R, Ferretti J, Portnoy D, Rood J, editors. Gram-
738 positive pathogens. Washington, D.C.: ASM Press; 2006. p. 381-412.
739
740 28. **Darling AE, Mau B, Perna NT.** progressiveMauve: multiple genome alignment with gene
741 gain, loss and rearrangement. *PLoS ONE* 2010;5:e11147.
742
743 29. **Seemann T.** snippy: fast bacterial variant calling from NGS reads
744 https://github.com/tseemann/snippy. 2015.
745
746 30. **Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S.** ViPTree: the viral
747 proteomic tree server. *Bioinformatics* 2017;33:2379-2380.
748

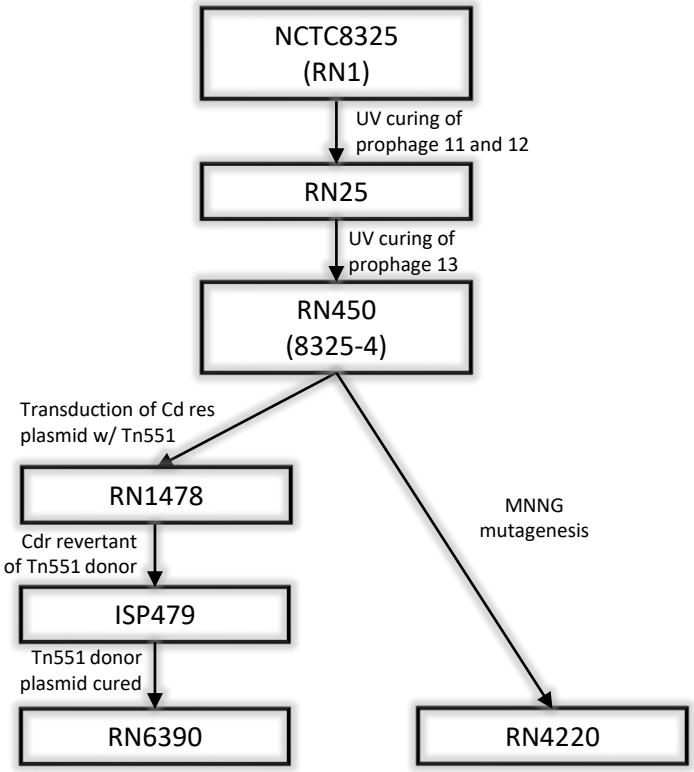
749 31. **Katoh K, Misawa K, Kuma K, Miyata T.** MAFFT: a novel method for rapid multiple
750 sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;30:3059-3066.
751
752 32. **Edgar RC.** MUSCLE: multiple sequence alignment with high accuracy and high
753 throughput. *Nucleic Acids Res* 2004;32:1792-1797.
754
755 33. **Martin DP, Murrell B, Golden M, Khoosal A, Muhire B.** RDP4: Detection and analysis
756 of recombination patterns in virus genomes. *Virus Evol* 2015;1:vev003.
757
758 34. **Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T.** trimAl: a tool for automated
759 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972-1973.
760
761 35. **Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD *et al*.** IQ-TREE 2:
762 New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol
763 Evol 2020;37:1530-1534.
764
765 36. **Letunic I, Bork P.** Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree
766 display and annotation. *Nucleic Acids Res* 2021;49:W293-W6.
767
768 37. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** Basic local alignment search
769 tool. *J Mol Biol* 1990;215:403-410.
770
771 38. **Guy L, Kultima JR, Andersson SG.** genoPlotR: comparative gene and genome
772 visualization in R. *Bioinformatics* 2010;26**:**2334-2335.
773
774 39. **RStudio Team.** RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL
775 http://www.rstudio.com/ 2020.
776
777 40. **Saha CK, Sanches Pires R, Brolin H, Delannoy M, Atkinson GC.** FlaGs and webFlaGs:
778 discovering novel biology through the analysis of gene neighbourhood conservation.
779 *Bioinformatics* 2021;37:1312-1314.
780
781 41. **Galata V, Fehlmann T, Backes C, Keller A.** PLSDB: a resource of complete bacterial
782 plasmids. *Nucleic Acids Res* 2019;47:D195-D202.
783
784 42. **Eddy SR.** Accelerated profile HMM searches. *PLoS Comput Biol* 2011;7:e1002195.
785

786 43. **Coordinators NR.** Database resources of the National Center for biotechnology
787 information. Nucleic Acids Res 2017;45:D12–17.
788
789 44. **Goerke C, Wirtz C, Fluckiger U, Wolz C.** Extensive phage dynamics in *Staphylococcus*
790 *aureus* contributes to adaptation to the human host during infection. *Mol Microbiol*
791 2006;61:1673-1685.
792
793 45**. Goerke C, Pantucek R, Holtfreter S, Schulte B, Zink M *et al*.** Diversity of prophages in
794 dominant *Staphylococcus aureus* clonal lineages. J Bacteriol 2009;191:3462-3468.
795
796 46. **Seybold U, Kourbatova EV, Johnson JG, Halvosa SJ, Wang YF *et al*.** Emergence of
797 community-associated methicillin-resistant *Staphylococcus aureus* USA300 genotype as a
798 major cause of health care-associated blood stream infections. *Clin Infect Dis* 2006;42:647-
799 656.
800
801 47. **Copin R, Sause WE, Fulmer Y, Balasubramanian D, Dyzenhaus S *et al*.** Sequential
802 evolution of virulence and resistance during clonal spread of community-acquired methicillin-
803 resistant Staphylococcus aureus. *Proc Natl Acad Sci U S A* 2019;116:1745-1754.
804
805 48. **Guy L, Kultima JR, Andersson SG.** genoPlotR: comparative gene and genome
806 visualization in R. *Bioinformatics* 2010;26:2334-2335.
807
808 49. **Arnold BJ, Huang IT, Hanage WP.** Horizontal gene transfer and adaptive evolution in
809 bacteria. *Nat Rev Microbiol* 2021 in press doi: 10.1038/s41579-021-00650-4.
810
811 50. **Stevenson C, Hall JP, Harrison E, Wood A, Brockhurst MA.** Gene mobility promotes
812 the spread of resistance in bacterial populations. *ISME J* 2017;11:1930-1932.
813
814 51. **Buckner MMC, Ciusa ML, Piddock LJV.** Strategies to combat antimicrobial resistance:
815 anti-plasmid and plasmid curing. *FEMS Microbiol Rev* 2018;42:781-804.
816
817 52. **Zhang D, de Souza RF, Anantharaman V, Iyer LM, Aravind L.** Polymorphic toxin
818 systems: Comprehensive characterization of trafficking modes, processing, mechanisms of
819 action, immunity and ecology using comparative genomics. *Biol Direct* 2012;7:18.
820

821 53. **Klein TA, Pazos M, Surette MG, Vollmer W, Whitney JC.** Molecular Basis for Immunity
822 Protein Recognition of a Type VII Secretion System Exported Antibacterial Toxin. *J Mol Biol*
823 2018;430:4344-4358.
824
825 54. **Michel B, Leach D.** Homologous Recombination-Enzymes and Pathways. *EcoSal Plus*
826 2012;5: doi: 10.1128/ecosalplus.7.2.7.
827
828 55. **Khasanov FK, Zvingila DJ, Zainullin AA, Prozorov AA, Bashkirov VI.** Homologous
829 recombination between plasmid and chromosomal DNA in *Bacillus subtilis* requires
830 approximately 70 bp of homology. *Mol Gen Genet* 1992;234:494-497.
831
832 56. **Shen P, Huang HV.** Homologous recombination in *Escherichia coli*: dependence on
833 substrate length and homology. *Genetics* 1986;112:441-457.
834
835 57. **Fujitani Y, Yamamoto K, Kobayashi I.** Dependence of frequency of homologous
836 recombination on the homology length. *Genetics* 1995;140:797-809.
837
838 58. **Tsuru T, Kawai M, Mizutani-Ui Y, Uchiyama I, Kobayashi I.** Evolution of paralogous
839 genes: Reconstruction of genome rearrangements through comparison of multiple genomes
840 within *Staphylococcus aureus*. *Mol Biol Evol* 2006;23:1269-1285.
841
842 59. **Tsuru T, Kobayashi I.** Multiple genome comparison within a bacterial species reveals a
843 unit of evolution spanning two adjacent genes in a tandem paralog cluster. *Mol Biol Evol*
844 2008;25:2457-2473.
845
846 60. **Ruhe ZC, Low DA, Hayes CS.** Polymorphic Toxins and Their Immunity Proteins:
847 Diversity, Evolution, and Mechanisms of Delivery. *Annu Rev Microbiol* 2020;74:497-520.
848
849 61. **Diep DB, Skaugen M, Salehian Z, Holo H, Nes IF.** Common mechanisms of target cell
850 recognition and immunity for class II bacteriocins. *Proc Natl Acad Sci U S A* 2007;104:2384-
851 2389.
852
853 62. **Kreiswirth BN, Lofdahl S, Betley MJ, O'Reilly M, Schlievert PM *et al*.** The toxic shock
854 syndrome exotoxin structural gene is not detectably transmitted by a prophage. *Nature*
855 1983;305:709-712.
856

857 63. **Novick RP.** Studies on plasmid replication. III. Isolation and characterization of replication-
858 defective mutants. *Mol Gen Genet* 1974;135:131-147.
859
860 64. **Pattee PA.** Distribution of Tn551 insertion sites responsible for auxotrophy on the
861 *Staphylococcus aureus* chromosome. *J Bacteriol* 1981;145:479-488.
862
863

a


NCTC8325
(RN1)
UV curing of
prophage 11 and 12
RN25
UV curing of
prophage 13
RN450
(8325-4)
Transduction of Cd res
plasmid w/ Tn551
RN1478
Cdr revertant
of Tn551 donor
ISP479
Tn551 donor
plasmid cured
RN6390
MNNG
mutagenesis
RN4220

b

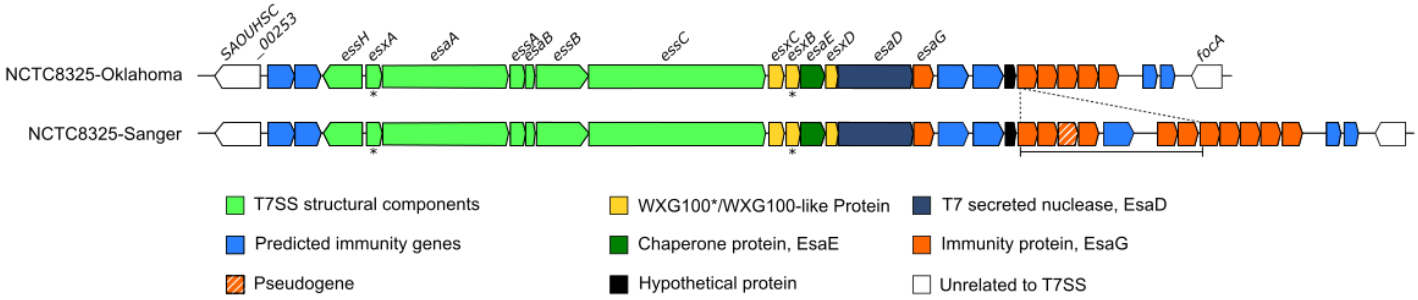

NCTC8325-Oklahoma
NCTC8325-Sanger
SAOUHSC_00253
essH
esxA
esaA
essA
esaB
essB
essC
esxC
esxB
esaE
esxD
esaD
esaG
focA
T7SS structural components
Predicted immunity genes
Pseudogene
WXG100*/WXG100-like Protein
Chaperone protein, EsaE
Hypothetical protein
T7 secreted nuclease, EsaD
Immunity protein, EsaG
Unrelated to T7SS

Fig 1

a

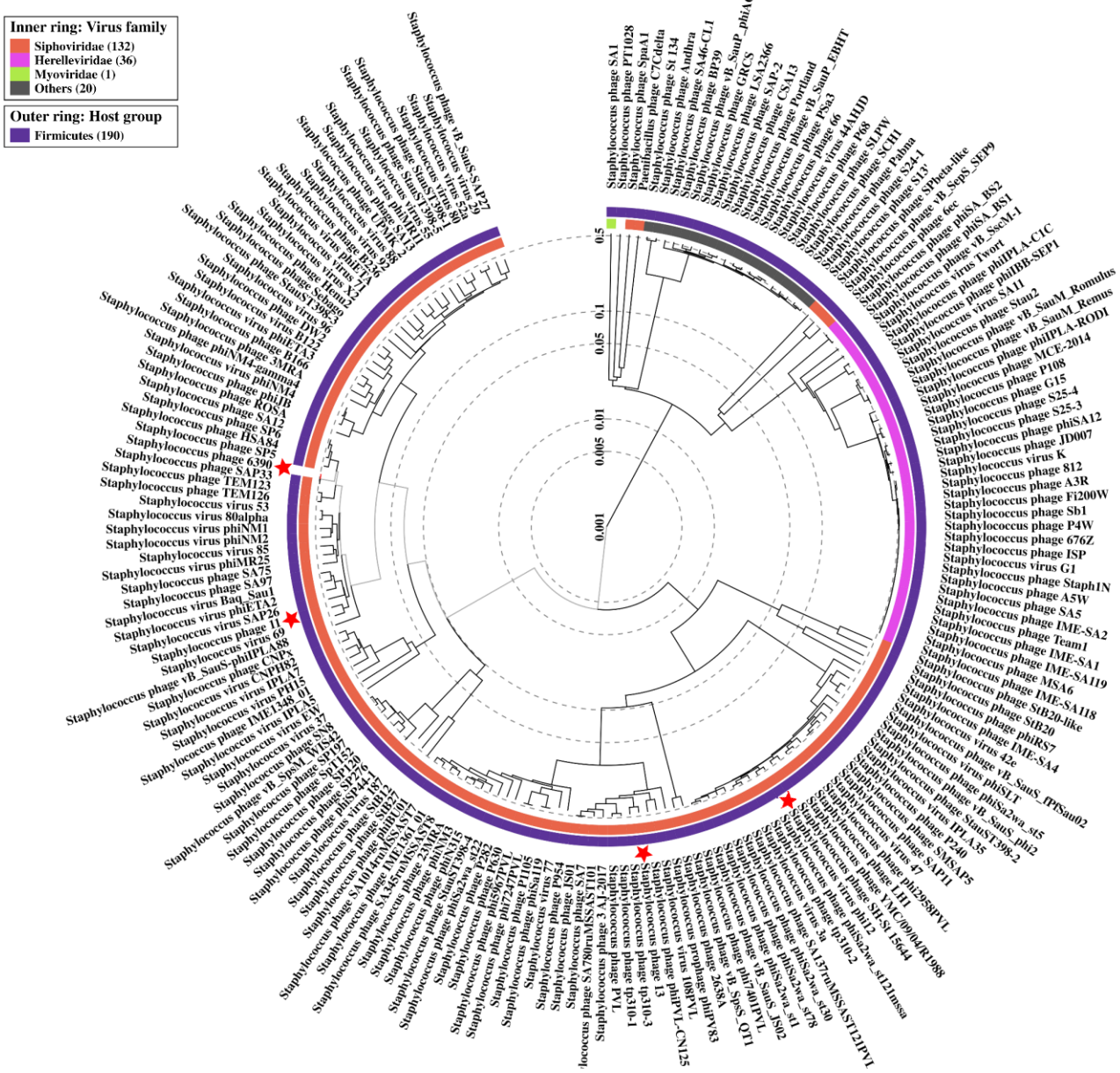
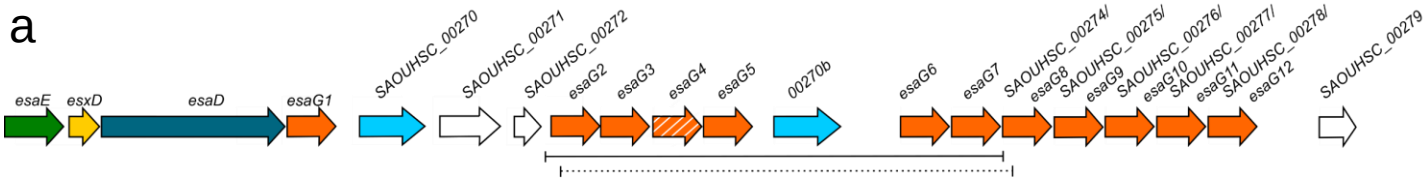

Inner ring: Virus family
Siphoviridae (132)
Herelleviridae (36)
Myoviridae (1)
Others (20)
Outer ring: Host group
Firmicutes (190)
0.5
0.1
0.05
0.01
0.005
0.001

Fig 2

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
```

Fig 3

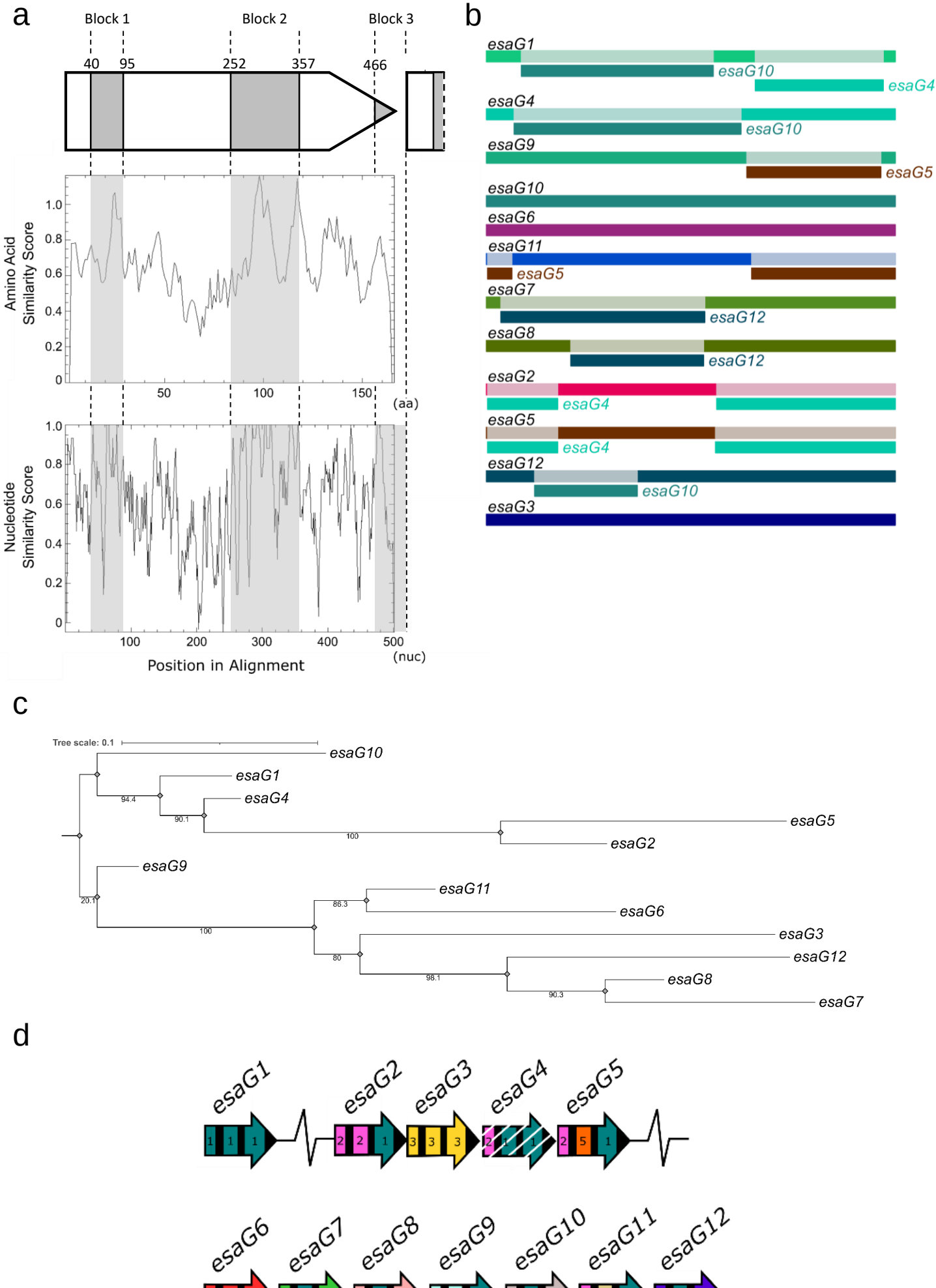

a
Block 1
Block 2
Block 3
40
95
252
357
466
Amino Acid Similarity Score
1.0
0.8
0.6
0.4
0.2
0
50
100
150
(aa)
Nucleotide Similarity Score
100
200
300
400
500
(nuc)
Position in Alignment
b
esaG1
esaG10
esaG4
esaG4
esaG10
esaG9
esaG5
esaG10
esaG6
esaG11
esaG5
esaG7
esaG12
esaG8
esaG12
esaG2
esaG4
esaG5
esaG4
esaG12
esaG10
esaG3
c
Tree scale: 0.1
esaG10
esaG1
esaG4
esaG5
esaG2
esaG9
esaG11
esaG6
esaG3
esaG12
esaG8
esaG7
94.4
90.1
100
20.1
86.3
100
80
98.1
90.3
d
esaG1
esaG2
esaG3
esaG4
esaG5
esaG6
esaG7
esaG8
esaG9
esaG10
esaG11
esaG12

Fig 4

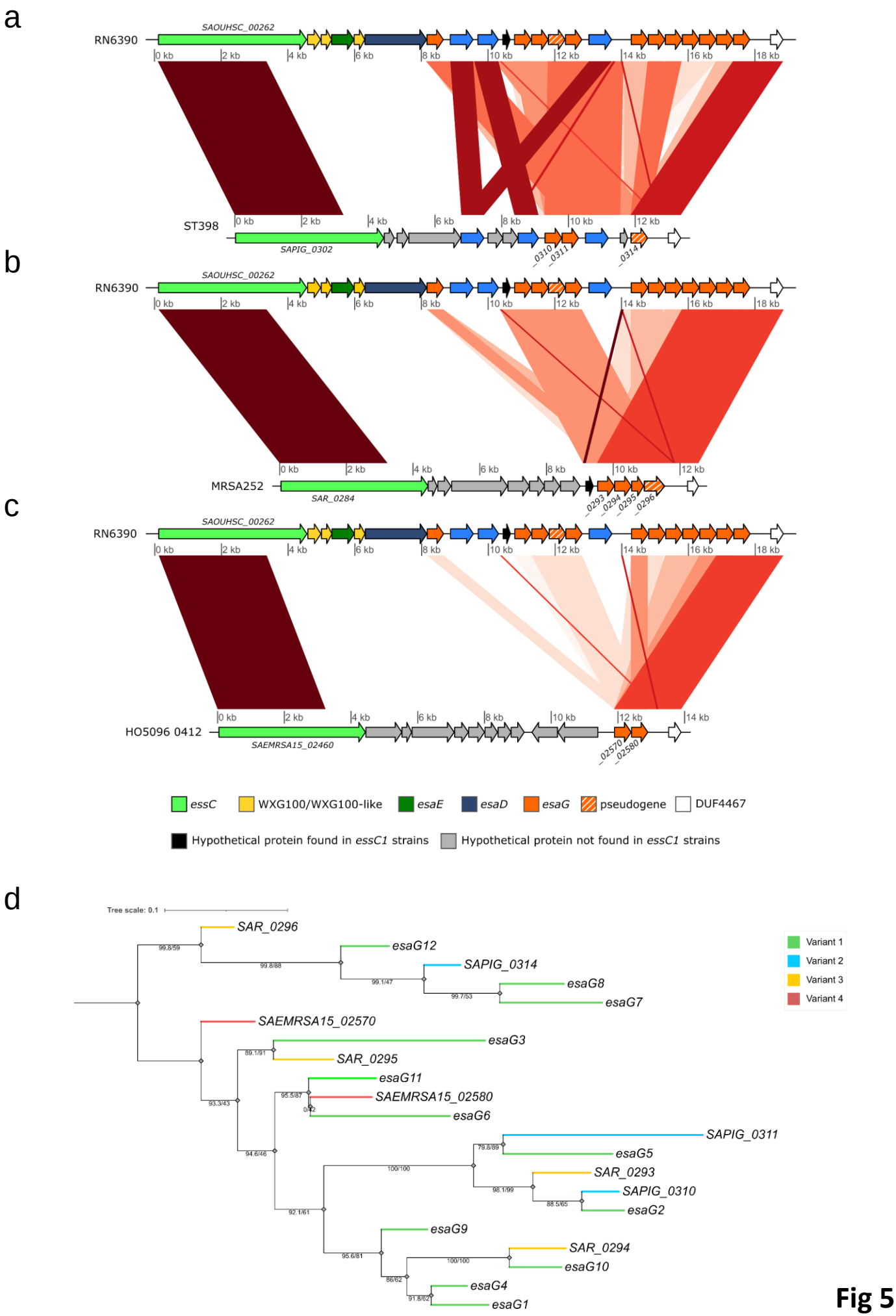
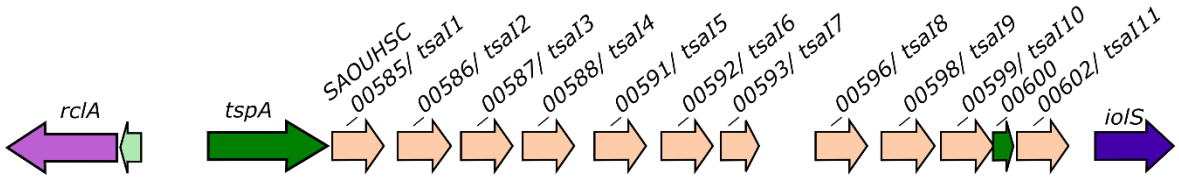

a
RN6390
SAOUHSC_00262
0 kb
2 kb
4 kb
6 kb
8 kb
10 kb
12 kb
14 kb
16 kb
18 kb
ST398
SAPIG_0302
_0310
_0311
_0314
b
RN6390
SAOUHSC_00262
MRSA252
SAR_0284
_0293
_0294
_0295
_0296
c
RN6390
SAOUHSC_00262
HO5096 0412
SAEMRSA15_02460
_02570
_02580
essC
WXG100/WXG100-like
esaE
esaD
esaG
pseudogene
DUF4467
Hypothetical protein found in essC1 strains
Hypothetical protein not found in essC1 strains
d
Tree scale: 0.1
SAR_0296
esaG12
SAPIG_0314
esaG8
esaG7
SAEMRSA15_02570
esaG3
SAR_0295
esaG11
SAEMRSA15_02580
esaG6
SAPIG_0311
esaG5
SAR_0293
SAPIG_0310
esaG2
esaG9
SAR_0294
esaG10
esaG4
esaG1
Variant 1
Variant 2
Variant 3
Variant 4

Fig 5

a



b

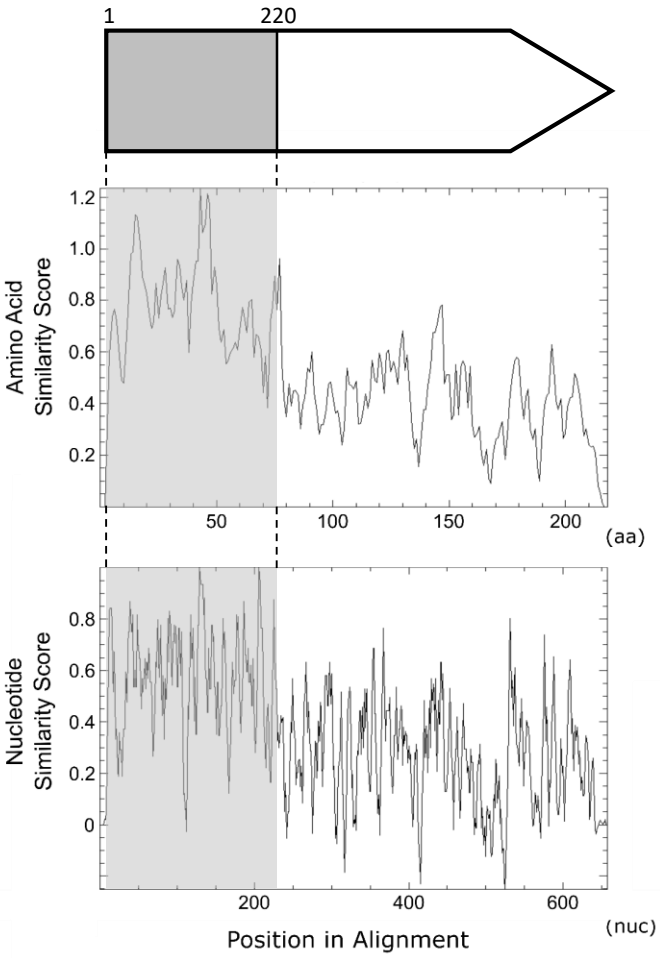
```
TsaI6      1 ---MLCESKVINKNPKYRIIKYDSEYLMIDLASNWIVFFFPFINWLIPKTYVKITKNDYE
TsaI5      1 ---MLCESKIINKNPKYRIIKYNDEYLMIDIISTWISLFFPFINWFIPKRYVKISREEFE
TsaI11     1 --MLLCESKIINKNPKYRIIKYNDEYLMVDIISTWISLFFPFINWFIPKEYVKISREEFE
TsaI3      1 ---MLCETEIINKNPKYRVIKYDDEYLMVDVIRTWLVYFFPFINWFIPKRCAKISREEYE
TsaI1      1 ---MLFNIKVINKNPRYKVVQYNDEYLLIDLVSTWLVYFFPFINWFIPKRYAKLSEKELE
TsaI9      1 ---MLCETENINKNPKYRIIKYKDEYLMIDLVSTWLALFFPMINWLIPKKYVKISEKDFE
TsaI7      1 ---MLCESRQIYKNPKYRVIRYNNEYFMVDLVSTWITYFFPMINWFLPKKYAKISENEFE
TsaI8      1 --MLLCDVRVIYKNPKYKVIQHNGEYLLVDLVSTWFVYFFPFINWFIPKKYAIISEEEFE
TsaI2      1 MEILLCEVRVVNKNPRYRIIKYKNDYLMIDLVSTWLVLFFPFINWLIPKKYVQISREDFD
TsaI4      1 ---MLCESRVINQNPKYRIIKYNNEYFMVDLVSTWIAYFLPMINWFIPKKYAKISREEFE
TsaI10     1 ---MLCESKVINKNPKYRVIKYGDEYLMIDLVSTWLTLFLPMINWLIPKKYVKISKKEFD

TsaI6     58 KLNIVKPVKNKSIGWTIFAGIVLLGGTVRRNTYLFDFQLEELIVWSSCFIGFLEIIFFYC
TsaI5     58 NLNIVKPAKKNVF-WPVAGISTLFAVTLRKYTHLLDTQLDRKLVIAICCITFIGILTFYV
TsaI11    59 NLNIVKPAKKNVF-WPVAGSSALLGVALRKYTHLLDIQLDKKLVIAICCITFIGILIFYV
TsaI3     58 KLNTVKPVKNKNF-WPVVGGTILLGATSRKYIHLLNIQLEKRSVIFICFVVFLCILIFFV
TsaI1     58 NLNVDKQNKNNIF-WPVTGSSFLFVVILRKYVHTFEVQLDNKILISLCFIGFIGIAAFYI
TsaI9     58 TLNIVKTAKINSF-WPVAGSTVLFGVMLRRYSHLFIVKYEYSIVILICCIIILGIFLFFL
TsaI7     58 RLNIVEPVKNNVF-WPVAGSSVLFGIILRKYGNFFNVQFEKQLAITVFFIMLIGMLIFYF
TsaI8     59 NLNVVKPNKNNVF-WSVIGSSVLFGVTLRKYIHVFDVQLDKLVVMILCALALICVIVFYF
TsaI2     61 NLNIVKPVKNKAL-WPAIGSILLFGTMFRDKIYIPDSHLEKNCVIIICSVLLLSILVFYI
TsaI4     58 SLNIVKPAKNNTF-WPVAGFAVLLTTLTRKYIYLLNIHLEKEIVILTCCMILLGVFALFI
TsaI10    58 DLNIVKPVKNKAF-WPVAGSTILFGVTFRKYIPSLNIQLEKNMVIVICCAIFLGVLILFL

TsaI6    118 YLNKKLTLNIYNESKNNELKLRLLPSFKNICFTIFYYLFTGFMSYGAFYLLVFENVQNLI
TsaI5    117 RLIKKSSLNIYN-TKNKRSKIILIPTLKNFCLTLFRYAFFILWTVIFSYALLSMSYQNII
TsaI11   118 RLIKKSSLNIYN-TKNKRSKIFLIPTLKNVCFTLFGYILFGGLTMLFLDALLSMSYQNII
TsaI3    117 LLNRKLKLKVFD-NKKEEQKIILVPTLKNAVLILYGYLLIGGMSILALSMLLTLENQNLI
TsaI1    117 YLNKKLKLKIYDDNLDNENRVILVPTFKDGSFIVFTYLLLGGCSILFLIWLMTIKPQNLL
TsaI9    117 YLNQKLKLQIYNENKNKSNKIIIFPTLKSLCLSIVLYIYLGGGSFFTIYMLLTIEVQNII
TsaI7    117 YLNKKLTLKIFNTNVVNKNRVVLIPTFKQGLLIVFAYFF----------------------
TsaI8    118 NLNRKLKLKVFDTNIEKNKRVILIPTFKLGCFLVFGYIFAGSFSIFSLIALMTIEPQNII
TsaI2    120 YLNQKVKLSIYN-DRSSNGKIMIFPSFKNLCFVLFSYFFCGGLSIMFLDVLISLSIQNII
TsaI4    117 YINTKLKLHIFDKNKSNNEKIILIPTFKNICLSLFAYILFGGLSTMALSMLVTSSPQNII
TsaI10   117 FLNRKLRLEIYN-NNSSKGKIILFPSLKNFCFTIFYYFLFGGLSIMALSMLLTLNPQNII

TsaI6    178 LYVSWLFMTMLFMFMNMHSIIDKKVHIF-LKSNK----
TsaI5    176 VYFAWITAIMGFFLVNIALIIDKNIHVI-LKN------
TsaI11   177 VYFVWIAVIMGFFLVNIALIIDKNIHVI-LKNQ-----
TsaI3    176 TFIAWGMGLMLFFLMNITLIVNKTVKVI-KR-------
TsaI1    177 VFIMWIIITIFFFLISMGSISNKKVYAK-LKKQ-----
TsaI9    177 LFITLFVIFLFFLFLNMCSLYDNKVHVL-FKSNGIEKF
TsaI7        ----------------------------------------
TsaI8    178 IFIYWIMMTMLFFLLNMTSIGNEKVRVI-MKNN-----
TsaI2    179 VFIAWVIMTMLFFFINMSSIIDKKIHVIYLRSYKY---
TsaI4    177 EFLALIGMTACFFLLNMSSVLDKKIHVI-LKTNK----
TsaI10   176 GFIGWLVMTAGFFLLNMSSIIDKKIYVL-SKTNTVEK-
```

c

```
tsaI1      1 -------AAATTACTAAACTTAGATTGTAGTTCGTAAGTTA--------------------
tsaI5      1 ----GTGAAAGTACTAAATTCAGATTAAAAATATGAAAATATCAG----------------
tsaI8      1 --------------------TTACATTTAAAATATTCTAAATGTTG---------------
tsaI11     1 -------AAATTACCAAATTAAATTTGCAATGGCTTTAATATTGTCGTTCTTAAATGTTT
tsaI12     1 -------------------TTTTGATACAAAAGGGCACAAGTGTTT---------------
tsaI9      1 ------------------------------------------------------------
tsaI3      1 ------------------------------------------------------------
tsaI4     20 TAGTATCGGATACTTAAATGTTGGTTCATAAAAAGCAATGATTTT---------------
tsaI7    181 TAGTATCGCATATTTAAACGGTGGTTCAAAAAAATATAATCATAT---------------
tsaI6      1 ------------------------------------------------------------

tsaI1     35 ---------------------------------------------AGTAATAAACAGGAAA
tsaI5     42 ------------------------------------------TTAATAAAACCTTTGATG
tsaI8     27 ------------------------------------------TCGACACAATCCTT----
tsaI11   110 TATGCATATCAATTTAAGCAATACTTATTTTAACT---GAGTTTTATATAACGTTT--TC
tsaI12    28 ------------------------------------------------------------
tsaI9      1 ------------------------------------------------------------
tsaI3      1 ---------------------------------------------ATAAATAACTA-ATT
tsaI4    113 ------------------------------------AAAAAGTGAATTAATAACTA-ATT
tsaI7    279 ATTGTG-ATTGATAAAGGAAAAAACTGTTTTAATTTAAAAATGAAATATAGAGCGT-GTT
tsaI6      1 ------------------------------------------------------------
```

Fig 6

a


1
220
Amino Acid Similarity Score
1.2
1.0
0.8
0.6
0.4
0.2
50
100
150
200
(aa)
Nucleotide Similarity Score
0.8
0.6
0.4
0.2
0
200
400
600
(nuc)
Position in Alignment

b

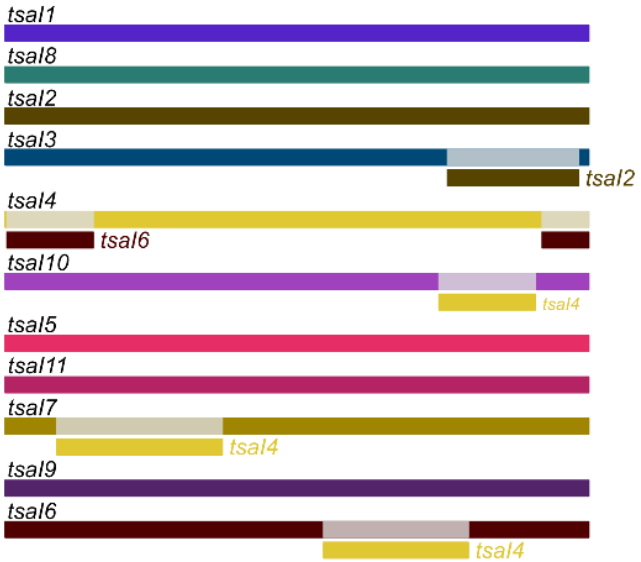

tsal1
tsal8
tsal2
tsal3
tsal2
tsal4
tsal6
tsal10
tsal4
tsal5
tsal11
tsal7
tsal4
tsal9
tsal6
tsal4

Fig 7

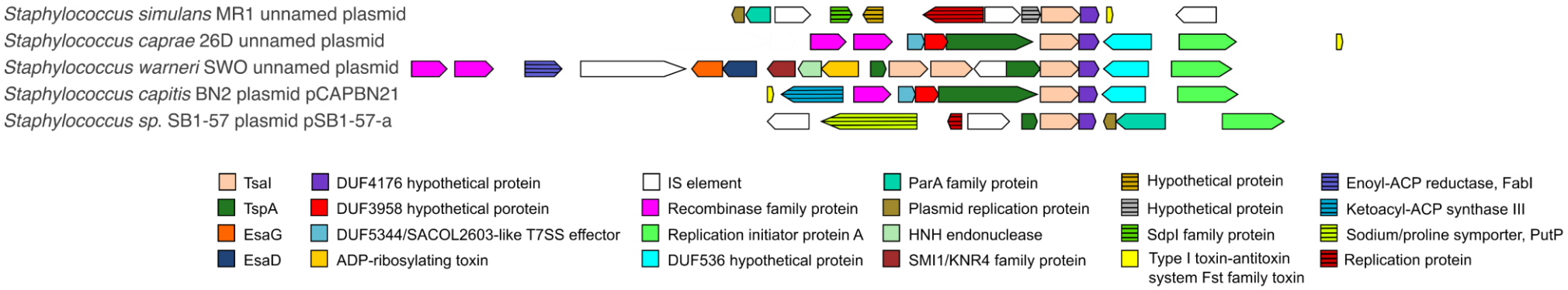
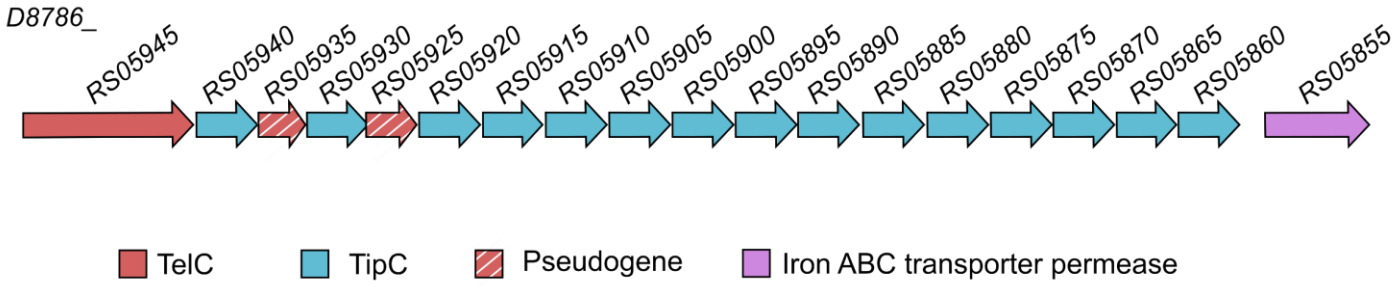

Staphylococcus simulans MR1 unnamed plasmid
Staphylococcus caprae 26D unnamed plasmid
Staphylococcus warneri SWO unnamed plasmid
Staphylococcus capitis BN2 plasmid pCAPBN21
Staphylococcus sp. SB1-57 plasmid pSB1-57-a
TsaI
TspA
EsaG
EsaD
DUF4176 hypothetical protein
DUF3958 hypothetical porotein
DUF5344/SACOL2603-like T7SS effector
ADP-ribosylating toxin
IS element
Recombinase family protein
Replication initiator protein A
DUF536 hypothetical protein
ParA family protein
Plasmid replication protein
HNH endonuclease
SMI1/KNR4 family protein
Hypothetical protein
Hypothetical protein
SdpI family protein
Type I toxin-antitoxin system Fst family toxin
Enoyl-ACP reductase, FabI
Ketoacyl-ACP synthase III
Sodium/proline symporter, PutP
Replication protein

Fig 8

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  MKKTLSILAIVIIIISSCFYFFSRQPKNIFDEIYQETEKTYRTNNILRKIDGFEISPGWP
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
```

Fig 9

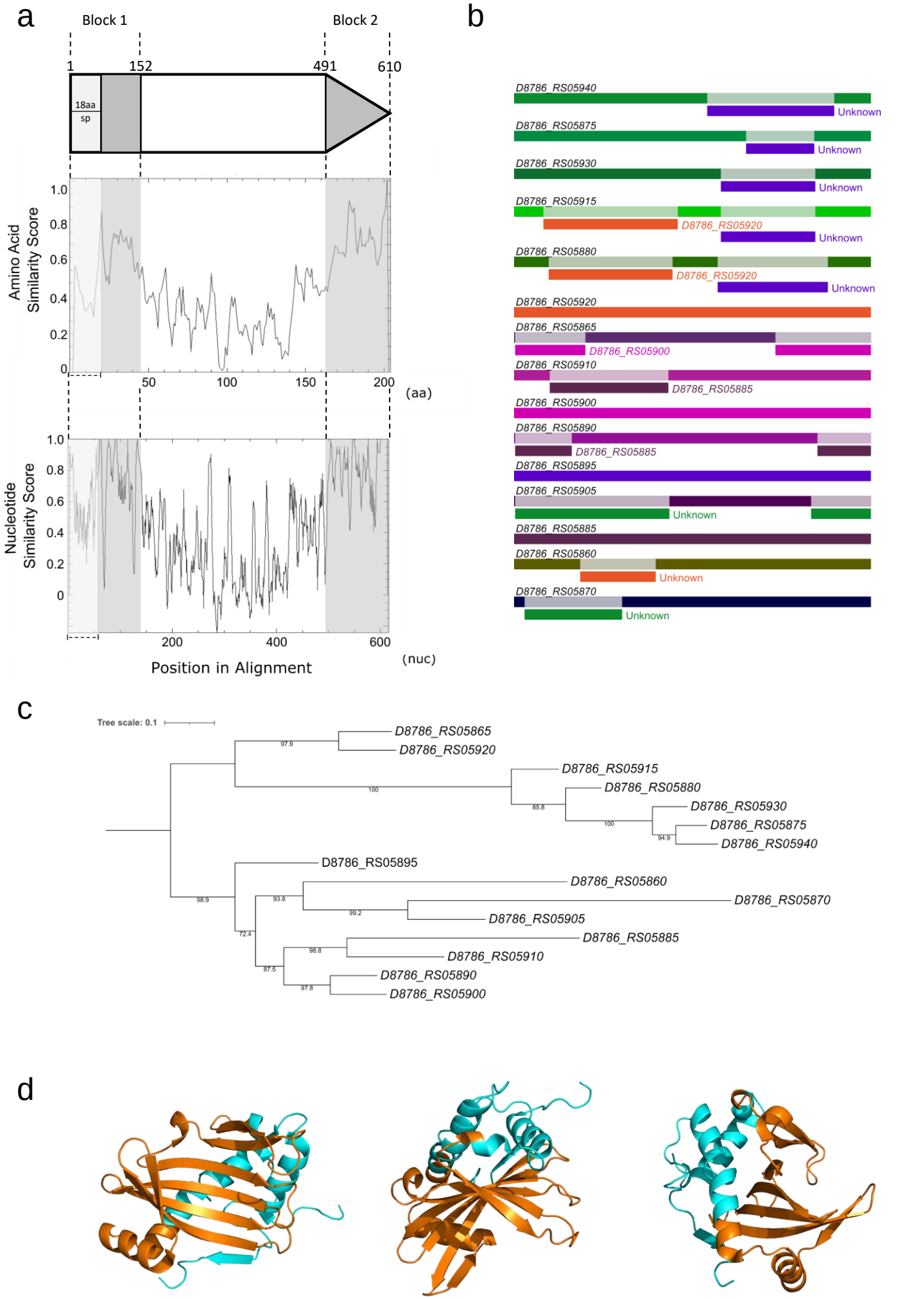
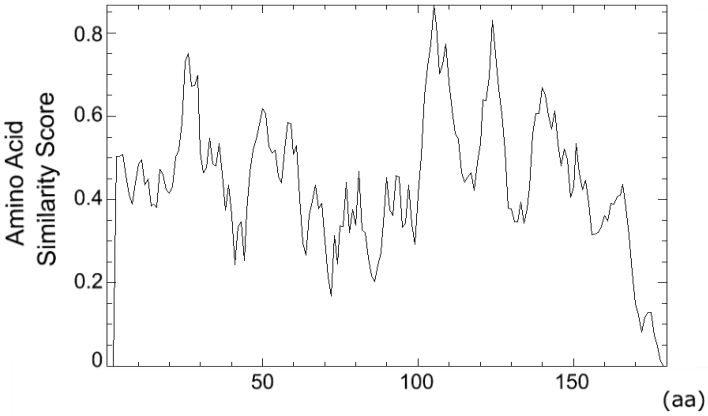
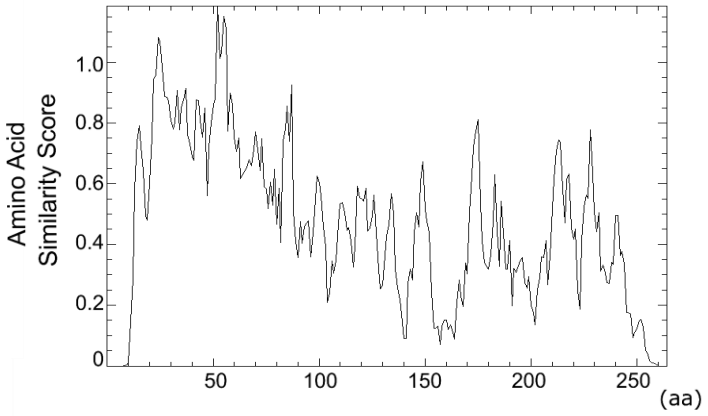

a
Block 1
Block 2
1
152
491
610
18aa
sp
Amino Acid Similarity Score
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
(aa)
Nucleotide Similarity Score
1.0
0.8
0.6
0.4
0.2
0
200
400
600
(nuc)
Position in Alignment
b
D8786_RS05940
Unknown
D8786_RS05875
Unknown
D8786_RS05930
Unknown
D8786_RS05915
D8786_RS05920
Unknown
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

Fig 10

a


Amino Acid
Similarity Score
0.8
0.6
0.4
0.2
0
50
100
150
(aa)

b

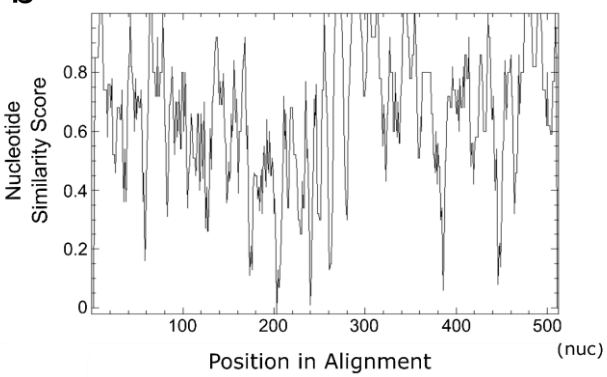

Amino Acid
Similarity Score
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
(aa)

Sup Fig 1

a

```
esaG1     1 ------------------------------ATGCTATTTAAAAAATAATGTATAGGAGAT--------------------AGAT
esaG5     1 ------------------------------ATGCTATTACAAAATCAATGTATAGGAGATAGG----------------TAAAT
esaG12    1 AAACATTGTTCAAACATCACAATGATAAAGCATATTATCAGTATTGTAGTGTGTGGAAAATGACAGCCATCTAAGGAGAAAAATG
esaG4i    1 ------------------------------------------------------GGAGTATAACATTTCT---------------
esaG2     1 ------------------------------------------------------AGGCGATAAC----------------------
esaG3     1 ------------------------------------------------------AGGCGATAAC----------------------
esaG8     1 ------------------------------------------------------AGGTGATAAC----------------------
esaG4ii   1 ------------------------------------------------------GGGCGATAAT----------------------
esaG10    1 ------------------------------------------------------GGGCGATAAT----------------------
esaG11    1 ------------------------------------------------------GGGCGATAAC----------------------
esaG7     1 ------------------------------------------------------GGGAGATAAT----------------------
esaG9     1 ------------------------------------------------------GGGAGATAAC----------------------
esaG6     1 ----------------------------------------------------GGGGAGATAAC----------------------
```

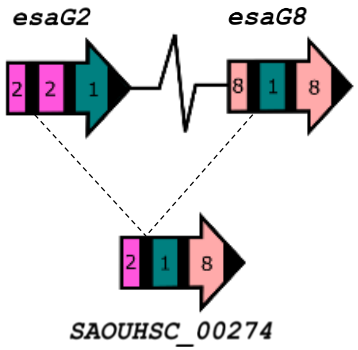
b



c

```
esaG2      1 ATGACTTTCGAAGAAAAATTAAGTGAAATGTATAGCGAGATTGCGAATAAGATTAGCAGC
SAOUHSC_00274 1 ATGACTTTCGAAGAAAAATTAAGTGAAATGTATAGCGAGATTGCGAATAAGATTAGCAGC
esaG8      1 ATGACTTTCGAAGGAAAACTAAGTCAAATGTACAACGAGATTGCGAATGAGATTAGTGGG

esaG2     61 ATGATACCGGTAGAGTGGGAAAAGGTATATGCAATGGCATATATAGATGACGGAGGAGGA
SAOUHSC_00274 61 ATGATACCGGTAGAGTGGGAGCAAGTATATGCAATGGCATATGTAACTGATCAAGCTGGA
esaG8     61 ATGATACCGGTAGAATGGGAGCAAGTATTTACAATAGCTTATGTAACTGATCAAGCTGGA

esaG2    121 GAAGTGTTCTACTATTACACAGAACCTGGAAGTAATGAATTATACTACTATACTAGTGTA
SAOUHSC_00274 121 GAAGTCATCTTTAATTATACTAAACCAGATAGTGATGAATTAAATTATTATTCAGACATA
esaG8    121 GAAGTCATCTTTAATTATACTAAACCAGATAGTGATGAATTAAATTATTATTCAGACATA

esaG2    181 TTAAATAAATATGATATATCGGAATCGGAATTTATGGATTCAGTGTATGAGTTGTATAAA
SAOUHSC_00274 181 CCTAAAGATTGCAATGTCTCAAAAGATATTTTTAAGAATTCATGGTTTAAAGTTTATCGA
esaG8    181 CCTAAAGATTGCAATGTCTCAAAAGATATTTTTAAGAATTCATGGTTTAAAGTTTATCGA

esaG2    241 CAATTTCAAAATTTAAGGAATTTATTTAAAGAAGAAGGACATGAACCATGGACATCATGC
SAOUHSC_00274 241 ATGTTTGATGAGTTAAGAGAAACTTTTAAAGAAGAAGGGCTTGAACCATGGACATCATGC
esaG8    241 ATGTTTGATGAGTTAAGAGAAACTTTTAAAGAAGAAGGGCTTGAACCATGGACATCATGC

esaG2    301 GAATTTGACTTTACAAGAGAAGGTGAATTAAAAGTTTCATTTGATTATATTGATTGGATA
SAOUHSC_00274 301 GAATTTGACTTTACAAGAGATGGCAAATTGAATGTATCTTTTGATTATATAGATTGGATA
esaG8    301 GAATTTGACTTTACAAGAGATGGCAAATTGAATGTATCTTTTGATTATATAGATTGGATA

esaG2    361 AATTCAGAATTTGGTCAAGTAGGTCGACAAAATTACTATAAGTATAGAAAATTTGGAATT
SAOUHSC_00274 361 AATACAGAGTTTGATCAATTGGGCCGTCAAAATTATTATATGTACAAAAATTTGGGGTT
esaG8    361 AATACAGAGTTTGATCAATTGGGCCGTCAAAATTATTATATGTACAAAAATTTGGGGTT

esaG2    421 TTACCAGAAACGGAATATGAAATTAATAAAGTTAAAGAAATCGAGCAATATGTTAAAGAG
SAOUHSC_00274 421 ATACCAGAAATGGAATATGAAATGGAAGAAGTTAAAGAAATCGAACAATATATTAAAGAG
esaG8    421 ATACCAGAAATGGAATATGAAATGGAAGAAGTTAAAGAAATCGAACAATATATTAAAGAG

esaG2    481 CAAGAAGAAGCTGAACAATAG
SAOUHSC_00274 481 CAAGAAGAAGCTGAACAATAG
esaG8    481 CAAGAAGAAGCTGAACAATAG
```

d

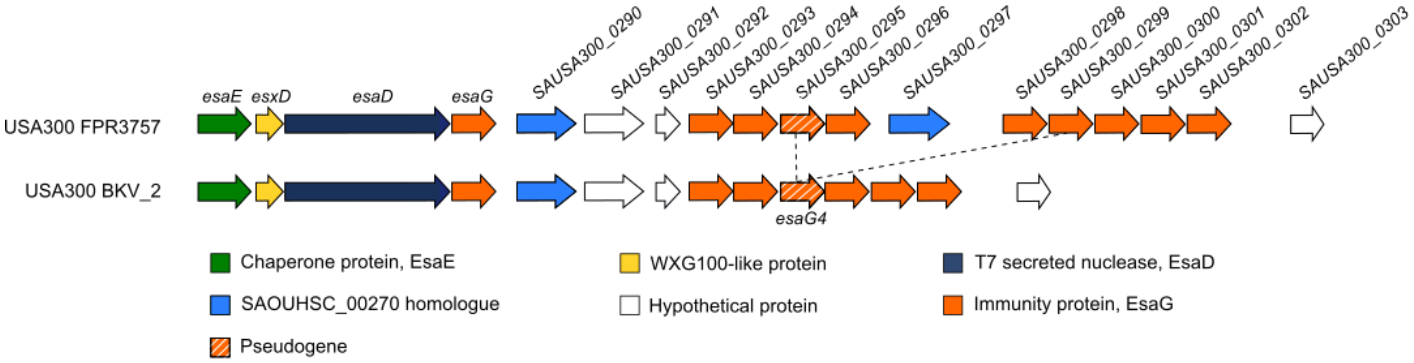


Sup Fig 2

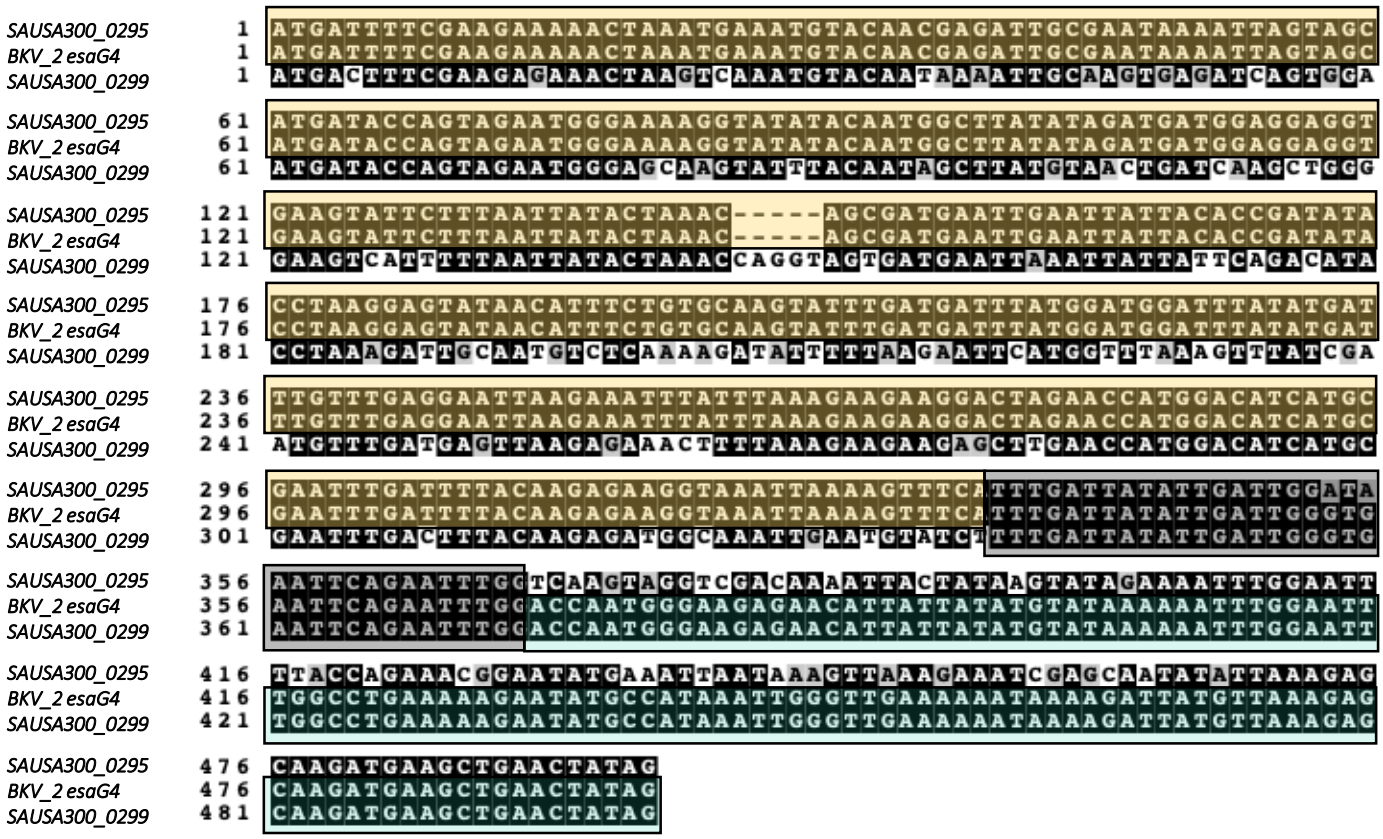
```
esaG1     1  ATGACATTTGAAGAGAAGCTTAGCAAAATATACAATGAAATTGCGAATGAGATTAGCAGT
esaG4     1  ATGATTTTCGAAGAAAAACTAAATGAAATGTACAACGAGATTGCGAATAAAATTAGTAGC
esaG9     1  ATGACTTTCGAAGAGAAAATAAGCAAATTATATAATGAGATTGCGAATGAGATTAGCAGT
esaG10    1  ATGACTTTCGAAGAGAAAATAAGCAAATTATATAATGAGATTGCGAATGAGATTAGCAGT
esaG6     1  ATGACTTTCGAAGAAAAACTAAGTGAAATATACAACAAGATTGCGAATGAGATTAGTGGG
esaG11    1  ATGACTTTCGAAGAAAAACTAAGTCAAATGTACAATGAAATTGCAAATGAAATCAGTGGA
esaG7     1  ATGACTTTCGAAGAGAAACTAAGTCAAATGTACAATAAAATTGCAAGTGAGATCAGTGGA
esaG8     1  ATGACTTTCGAAGGAAAACTAAGTCAAATGTACAACGAGATTGCGAATGAGATTAGTGGG
esaG2     1  ATGACTTTCGAAGAAAAATTAAGTGAAATGTATAGCGAGATTGCGAATAAGATTAGCAGC
esaG5     1  ATGACTTTCGAAGAAAAATTAAGTGAAATGTATAGCGAGATTGCGAATAAGATTATTAGC
esaG12    1  ATGACTTTCGAAGAAAAGCTAAGTCAAATGTACAATGAAATTGCAAATGAAATCAGTGGA
esaG3     1  ATGAATTTCGAAGAAAAATTAAGTGAAATGTACAATAAAATTGCAAATGAGATCAATGAA

esaG1    61  ATGATACCGGTAGAGTGGGAAAAAGTATATACAATGGCTTATATAGATGATGGAGGAGGT
esaG4    61  ATGATACCAGTAGAATGGGAAAAGGTATATACAATGGCTTATATAGATGATGGAGGAGGT
esaG9    61  ATGATACCGGTAGAGTGGGAAAAAGTATATACAATGGCTTATATAGATGATGGAGGAGGT
esaG10   61  ATGATACCGGTAGAGTGGGAAAAAGTATATACAATGGCTTATATAGATGATGGAGGAGGT
esaG6    61  ATGATACCAGTAGAATGGGACCAAGTATTTACGATAGCTTATGTAAATGATAGAGGTGGA
esaG11   61  ATGATACCAGTTGAATGGGAAAATATATATACAATTGCCTATGTAACTGATCAAGGTGGA
esaG7    61  ATGATACCAGTAGAATGGGAGCAAGTATTTACAATAGCTTATGTAACTGATCAAGCTGGG
esaG8    61  ATGATACCGGTAGAATGGGAGCAAGTATTTACAATAGCTTATGTAACTGATCAAGCTGGA
esaG2    61  ATGATACCGGTAGAGTGGGAAAAGGTATATGCAATGGCATATATAGATGACGGAGGAGGA
esaG5    61  ATGATACCGGTAGAGTGGGAAAAAGTATATGCAATGGCATATATCGATGAACAATGTGGA
esaG12   61  ATGATACCAGTAGAATGGGAAAAAAGTATATACAATTGCCTACGTAGATGATGAAGGTGGA
esaG3    61  ATGATACCAGTAGAGTGGGAAAAGGTATATGTAATTGCATATGTGGACGATGGAGGTGGG

esaG1   121  GAAGTATTCTTTAATTATACTAAACCAGGTAGTGATGACTTGAATTATTACACCAATATA
esaG4   121  GAAGTATTCTTTAATTATACTAAAC-----AGCGATGAATTGAATTATTACACCGATATA
esaG9   121  GAAGTATTCTTTAATTATACTAAACCAGGTAGTGATGACTTGAATTATTACACCGATATA
esaG10  121  GAAGTATTCTTTAATTATACTAAACCAGGAAGTGAAGATTTGAATTATTATACCGATATA
esaG6   121  GAGATCGTTTTAATTATACTAAACCAGGTAGCGATGAATTGAATTATTACACATATATC
esaG11  121  GAGGTCATTTTTAATTATACTAAACCAGGTAGCGATGAATTGAATTATTACACATATATC
esaG7   121  GAAGTCATTTTTAATTATACTAAACCAGGTAGTGATGAATTAAATTATTATTCAGACATA
esaG8   121  GAAGTCATCTTTAATTATACTAAACCAGATAGTGATGAATTAAATTATTATTCAGACATA
esaG2   121  GAAGTGTTCTACTATTACACAGAACCTGGAAGTAATGAATTATACTACTATACTAGTGTA
esaG5   121  GAAGTGTTTTACAATTATACTGAACCAAGCAGTGATGAATTATTTACTATACAAGTGTG
esaG12  121  GAGGTTGTTTTTAATTATACTAAACCAGGAAGTGAAGATTTGAATTATTATTCAGATATT
esaG3   121  CAAGTTATATTTTATTATACCAAGCCCAGAAATGATGAGCTGTATTATTATTCGAGTATC

esaG1   181  CCTAAGGAGTATAACATTTCTGTGCAAGTATTTGATGATTTATGGATGGATTTATATGAT
esaG4   176  CCTAAGGAGTATAACATTTCTGTGCAAGTATTTGATGATTTATGGATGGATTTATATGAT
esaG9   181  CCTAAGGAGTATAACATCTCTGTGCAAGTATTTGATGATTTATGGATGGATTTATATGAT
esaG10  181  CCTAAGGAGTATAATGTTTCTGTGCAAGTATTTGATGATTTATGGATGGATTTATATGAT
esaG6   181  CCTAGAGAGTATAATGTCTCTGAAAAAGTATTTTATGATTTGTGGACGGATTTATATAGA
esaG11  181  CCTAGAGAGTATAATGTCTCTGAAAAAGTATTTTATGATTTGTGGACGGATTTATATAGA
esaG7   181  CCTAAAGATTGCAATGTCTCAAAAGATATTTTTAAGAATTCATGGTTTAAAGTTTATCGA
esaG8   181  CCTAAAGATTGCAATGTCTCAAAAGATATTTTTAAGAATTCATGCTTTAAAGTTTATCGA
esaG2   181  TTAAATAAATATGATATATCGGAATCGGAATTTATGGATTCAGTGTATGAGTTGTATAAA
esaG5   181  ATAAAGAAGTATAATTTATTGAAATCAAGCTTTATGGATTCAGTATATGAGTTGCATGAT
esaG12  181  CCTAAAGATTGCAATGTCTCAAAAGATATTTTTAAGAATTCATGGTTTAAAGTTTATCGA
esaG3   181  GTTGAAGATTATAATGTTTTAGAAGAAATCTTTGATGATTTATGGATGGAACTTTATAGA

esaG1   241  TTATTTGAGGAGTTAAGAGATTTATTTAAAGAAGAAGATTTAGAACCATGGACATCATGC
esaG4   236  TTGTTTGAGGAATTAAGAAATTTATTTAAAGAAGAAGGACTAGAACCATGGACATCATGC
esaG9   241  TTGTTTGAGGAATTAAGAGATTTATTTAAAGAAGAAGGACTTGAACCATGGACATCATGT
esaG10  241  TTGTTTAAGAATTTAAGAAATTTATTTAAAGAAGAAGGACTTGAACCATGGACATCATGT
esaG6   241  TTGTTTAAGAAGTTAAGAAACGCATTTAAAGAAGAAGATTTAGAACCATGGACATCATGT
esaG11  241  TTGTTTAAGAAGTTAAGAGAAACTTTTAAAGAAGAAGGGCTTGAACCATGGACATCAAGT
esaG7   241  ATGTTTGATGAGTTAAGAGAAACTTTTAAAGAAGAAGAGCTTGAACCATGGACATCATGC
esaG8   241  ATGTTTGATGAGTTAAGAGAAACTTTTAAAGAAGAAGGGCTTGAACCATGGACATCATGC
esaG2   241  CAATTTCAAAATTTAAGGAATTTATTTAAAGAAGAAGGACATGAACCATGGACATCATGC
esaG5   241  CAATTTGAGGAACTAAGAGAAGTATTTATAGAAGAAGGTCTTGAACCATGGACATCATGC
esaG12  241  ATGTTTGATGAGTTAAGAGAAACTTTTAAAAAAGAAGATTTAGAACCGTGGACATCATGT
esaG3   241  TCATTTAAAAATTAAGAAATATATTTAAAGAAGAAAGTCTTGAACCATGGACATCATGT

esaG1   301  GAATTTGATTTTACAAGAGAAGGTGAATTAAAAGTTTCATTTGATTATATTGATTGGATA
esaG4   296  GAATTTGATTTTACAAGAGAAGGTAAATTAAAAGTTTCATTTGATTATATTGATTGGATA
esaG9   301  GAATTTGACTTTACAAGCGAAGGTAAATTAAAAGTTTCATTTGATTATATAGATTGGATA
esaG10  301  GAATTTGACTTTACAAGAGACGGCAAATTGAATGTTTCATTTGATTATATTGATTGGGCG
esaG6   301  GAATTTGACTTTACAAGAGATGGTAAATTGAATGTTGTATTTGATTATGTTGATTGGATG
esaG11  301  GAATTTGACTTTACAAGCGAAGGTAAATTAAAAGTTTCATTTGATTATATTGATTGGATA
esaG7   301  GAATTTGACTTTACAAGAGATGGCAAATTGAATGTATCTTTTGATTATATTGATTGGGTG
esaG8   301  GAATTTGACTTTACAAGAGATGGCAAATTGAATGTATCTTTTGATTATATTAGATTGGATA
esaG2   301  GAATTTGACTTTACAAGAGAAGGTGAATTAAAAGTTTCATTTGATTATATTGATTGGATA
esaG5   301  GAATTTGATTTTACAAAAGAAGATAAATTAAAAGTTTCATTTGATTATATTGATTGGATA
esaG12  301  GAATTTGACTTTACAAGAAAGGGAAATTTAAAAGTATCATTTGATTATATTAGATTGGATT
esaG3   301  GAATTTGACTTTACAAAAGAAGATAAATTAAAAGTTTCATTTGATTATATTGATTGGAAG

esaG1   361  AATTCAGAATTTGGTCAAATAGGTCGACAAAATTACTATAAGTATAGAAAATTTGGAATT
esaG4   356  AATTCAGAATTTGGTCAAGTAGGTCGACAAAATTACTATAAGTATAGAAAATTTGGAATT
esaG9   361  AATACAGAGTTTGATCAATTAGGCCGTGAAAATTATATATGTATAAAAAATTTGGGGTT
esaG10  361  AATTCAGAGTTTGGACAAATGGGAAGAGAACATTATTACATGTATAAAAAATTTGGAATT
esaG6   361  AATTCAGAATTTGGTCCAATAGCAAAAGAAAATTATTATATGTACAAAAAGTTTGGTGTT
esaG11  361  AATACAGAGTTTGATCAATTAGGCCGTGAAAACTATTATATGTATAAAAAGTTTGGTGTT
esaG7   361  AATTCAGAATTTGGACCAATGGGAAGAGAACATTATTATATGTATAAAAAATTTGGAATT
esaG8   361  AATACAGAGTTTGATCAATTGGGCCGTCAAAATTATTATATGTACAAAAAATTTGGGGTT
esaG2   361  AATTCAGAATTTGGTCAAGTAGGTCGACAAAATTACTATAAGTATAGAAAATTTGGAATT
esaG5   361  AATACAGAGTTTGATCAATTAGGCCGTGAAAATTATTATATGTACAAAAAATTTGGTGTT
esaG12  361  AAATTAGGTTTTGGCCCATCAGGAAAGGAAAACTACTATATGTACAAAAAATTTGGTATT
esaG3   361  AATACAGAATTTGATCAATTAGGCCAAGAAAACTACTATAACTATAGAAAATTCGGAATT

esaG1   421  TTACCAGAAACGGAATATGAAATTAATAAAGTTAAAGAAATCGAGCAATATATTAAAGAG
esaG4   416  TTACCAGAAACGGAATATGAAATTAATAAAGTTAAAGAAATCGAGCAATATATTAAAGAG
esaG9   421  TTACCAGAAATGGAATATGAAATGGAAGAAATTAAAGAAATCGATCAATATATTAAAGAG
esaG10  421  TGGCCTGAAAAAGAATATGCCATAAATTGGGTAAAAAAAATAAAAGATTATGTTAAAGAG
esaG6   421  TTACCAGAAACGGAATATGAAATTAATAAAGTTAAAGAAATCGATCAATATGTTAAAGAG
esaG11  421  TTACCAGAAATGGAATACGAAATGGAAGAAGTTAAAGAAATCGAGCAATATATTAAAGAG
esaG7   421  TGGCCTGAAAAAGAATATGCCATAAATTGGGTTGAAAAAATAAAAGATTATGTTAAAGAG
esaG8   421  ATACCAGAAATGGAATATGAAATGGAAGAAGTTAAAGAAATCGAACAATATATTAAAGAG
esaG2   421  TTACCAGAAACGGAATATGAAATTAATAAAGTTAAAGAAATCGAGCAATATGTTAAAGAG
esaG5   421  TTACCAGAAACGGAATATGAAATGGAGGAGGTTAAAGAAATCGAGCAAATATATTAAAGAG
esaG12  421  TTACCAGATATGGAATATGAAATGGAAGAAATTCGAGCAGTAGAGAAGTATGTTAAAGAG
esaG3   421  ATACCAGAAATGGAATATGAAATGGAAGAGGTTAAACAAATCGAGCAATATATTAAAGAG

esaG1   481  CTAGAAGAA---------TAA
esaG4   476  CAAGATGAAGCTGAACTATAG
esaG9   481  CAAGATGAAGCTGAAATATAG
esaG10  481  CAAGATGAAGCTGAACTATAG
esaG6   481  CAAGATGAAGCTGAACTATAG
esaG11  481  CAAGATGAAGCTGAACTATAG
esaG7   481  CAAGATGAAGCTGAACTATAG
esaG8   481  CAAGAAGAAGCTGAACAATAG
esaG2   481  CAAGAAGAAGCTGAACAATAG
esaG5   481  CAAGAAGAA---------TAG
esaG12  481  CAAGAG------------TAG
esaG3   481  CAAGAAGAAGCTGAACAATAG
```

Sup Fig 3

a

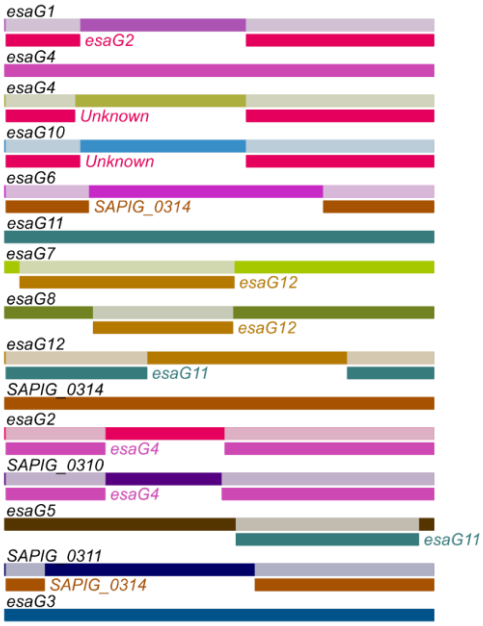

esaE
esxD
esaD
esaG
SAUSA300_0290
SAUSA300_0291
SAUSA300_0292
SAUSA300_0293
SAUSA300_0294
SAUSA300_0295
SAUSA300_0296
SAUSA300_0297
SAUSA300_0298
SAUSA300_0299
SAUSA300_0300
SAUSA300_0301
SAUSA300_0302
SAUSA300_0303
USA300 FPR3757
USA300 BKV_2
esaG4
Chaperone protein, EsaE
WXG100-like protein
T7 secreted nuclease, EsaD
SAOUHSC_00270 homologue
Hypothetical protein
Immunity protein, EsaG
Pseudogene

b

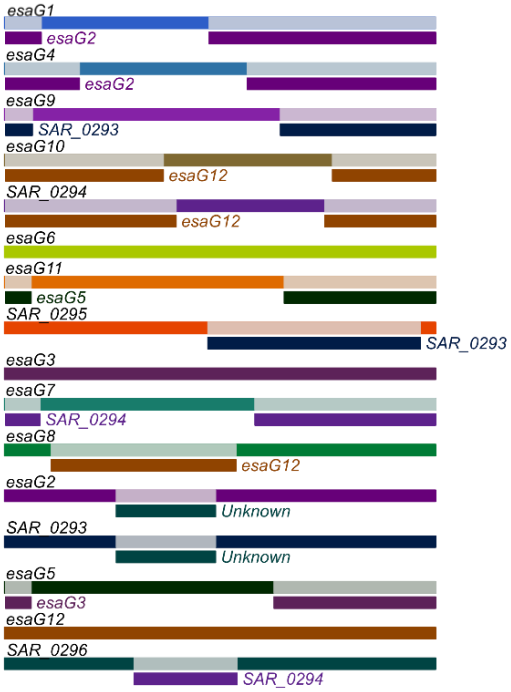

SAUSA300_0295 1 ATGATTTTCGAAGAAAAACTAAATGAAATGTACAACGAGATTGCGAATAAAATTAGTAGC
BKV_2 esaG4 1 ATGATTTTCGAAGAAAAACTAAATGAAATGTACAACGAGATTGCGAATAAAATTAGTAGC
SAUSA300_0299 1 ATGACTTTCGAAGAGAAACTAAGTCAAATGTACAATAAAATTGCAAGTGAGATCAGTGGA
SAUSA300_0295 61 ATGATACCAGTAGAATGGGAAAAGGTATATACAATGGCTTATATAGATGATGGAGGAGGT
BKV_2 esaG4 61 ATGATACCAGTAGAATGGGAAAAGGTATATACAATGGCTTATATAGATGATGGAGGAGGT
SAUSA300_0299 61 ATGATACCAGTAGAATGGGAGCAAGTATTTACAATAGCTTATGTAACTGATCAAGCTGGG
SAUSA300_0295 121 GAAGTATTCTTTAATTATACTAAAC-----AGCGATGAATTGAATTATTACACCGATATA
BKV_2 esaG4 121 GAAGTATTCTTTAATTATACTAAAC-----AGCGATGAATTGAATTATTACACCGATATA
SAUSA300_0299 121 GAAGTCATTTTTAATTATACTAAACCAGGTAGTGATGAATTAAATTATTATTCAGACATA
SAUSA300_0295 176 CCTAAGGAGTATAACATTTCTGTGCAAGTATTTGATGATTTATGGATGGATTTATATGAT
BKV_2 esaG4 176 CCTAAGGAGTATAACATTTCTGTGCAAGTATTTGATGATTTATGGATGGATTTATATGAT
SAUSA300_0299 181 CCTAAAGATTGCAATGTCTCAAAAGATATTTTTAAGAATTCATGGTTTAAAGTTTATCGA
SAUSA300_0295 236 TTGTTTGAGGAATTAAGAAATTTATTTAAAGAAGAAGGACTAGAACCATGGACATCATGC
BKV_2 esaG4 236 TTGTTTGAGGAATTAAGAAATTTATTTAAAGAAGAAGGACTAGAACCATGGACATCATGC
SAUSA300_0299 241 ATGTTTGATGAGTTAAGAGAAACTTTTAAAGAAGAAGAGCTTGAACCATGGACATCATGC
SAUSA300_0295 296 GAATTTGATTTTACAAGAGAAGGTAAATTAAAAGTTTCATTTGATTATATTGATTGGATA
BKV_2 esaG4 296 GAATTTGATTTTACAAGAGAAGGTAAATTAAAAGTTTCATTTGATTATATTGATTGGGTG
SAUSA300_0299 301 GAATTTGACTTTACAAGAGATGGCAAATTGAATGTATCTTTTGATTATATTGATTGGGTG
SAUSA300_0295 356 AATTCAGAATTTGGTCAAGTAGGTCGACAAAATTACTATAAGTATAGAAAATTTGGAATT
BKV_2 esaG4 356 AATTCAGAATTTGGACCAATGGGAAGAGAACATTATTATATGTATAAAAAATTTGGAATT
SAUSA300_0299 361 AATTCAGAATTTGGACCAATGGGAAGAGAACATTATTATATGTATAAAAAATTTGGAATT
SAUSA300_0295 416 TTACCAGAAACGGAATATGAAATTAATAAAGTTAAAGAAATCGAGCAATATATTAAAGAG
BKV_2 esaG4 416 TGGCCTGAAAAAGAATATGCCATAAATTGGGTTGAAAAAATAAAAGATTATGTTAAAGAG
SAUSA300_0299 421 TGGCCTGAAAAAGAATATGCCATAAATTGGGTTGAAAAAATAAAAGATTATGTTAAAGAG
SAUSA300_0295 476 CAAGATGAAGCTGAACTATAG
BKV_2 esaG4 476 CAAGATGAAGCTGAACTATAG
SAUSA300_0299 481 CAAGATGAAGCTGAACTATAG

Sup Fig 4

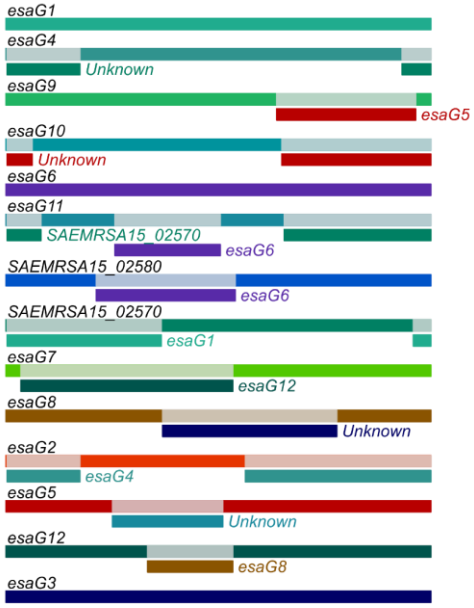
a


esaG1
esaG2
esaG4
esaG4
Unknown
esaG10
Unknown
esaG6
SAPIG_0314
esaG11
esaG7
esaG12
esaG8
esaG12
esaG12
esaG11
SAPIG_0314
esaG2
esaG4
SAPIG_0310
esaG4
esaG5
esaG11
SAPIG_0311
SAPIG_0314
esaG3

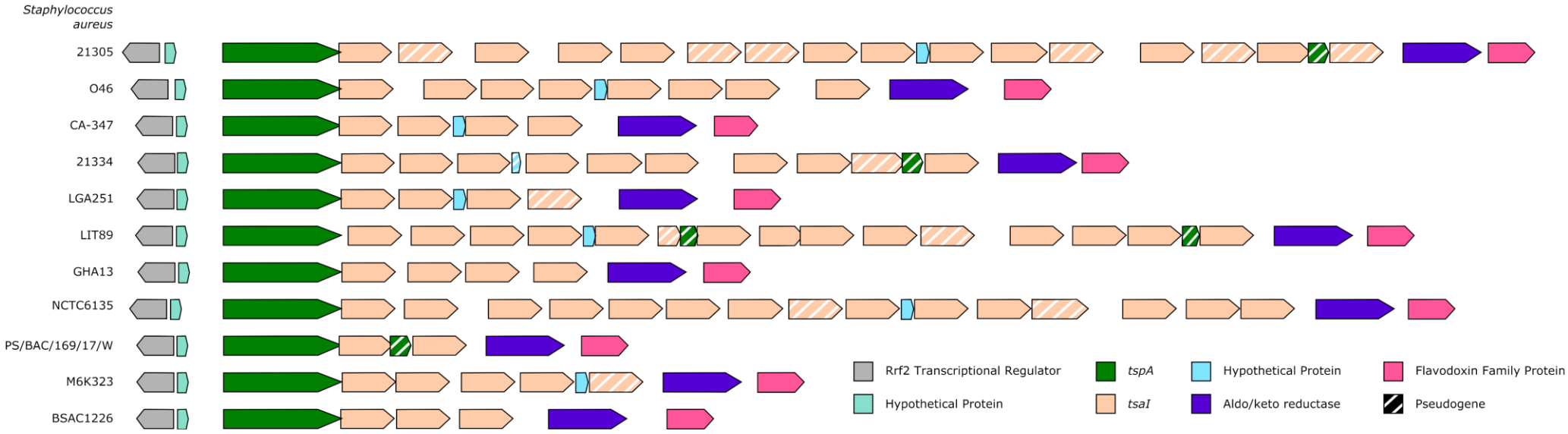
b


esaG1
esaG2
esaG4
esaG2
esaG9
SAR_0293
esaG10
esaG12
SAR_0294
esaG12
esaG6
esaG11
esaG5
SAR_0295
SAR_0293
esaG3
esaG7
SAR_0294
esaG8
esaG12
esaG2
Unknown
SAR_0293
Unknown
esaG5
esaG3
esaG12
SAR_0296
SAR_0294

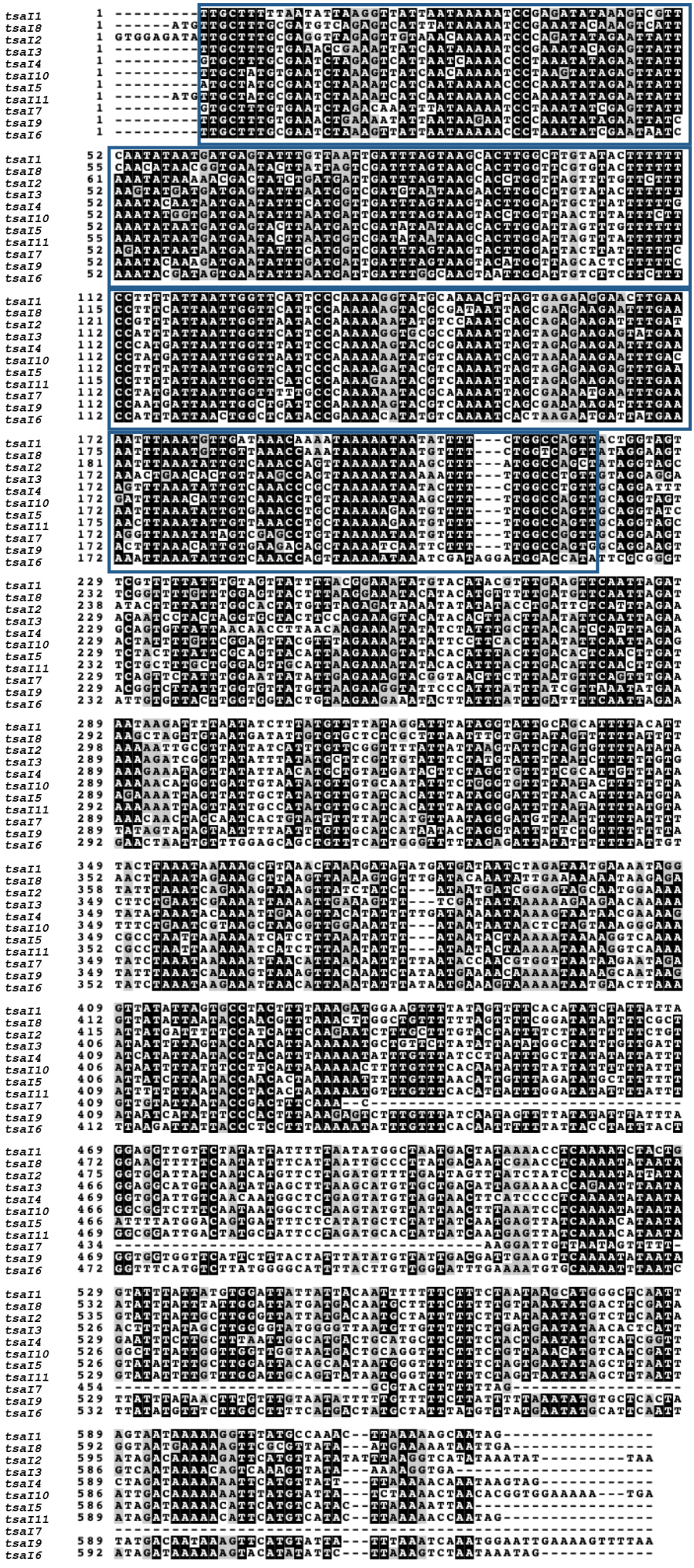
c


esaG1
esaG4
Unknown
esaG9
esaG5
esaG10
Unknown
esaG6
esaG11
SAEMRSA15_02570
esaG6
SAEMRSA15_02580
esaG6
SAEMRSA15_02570
esaG1
esaG7
esaG12
esaG8
Unknown
esaG2
esaG4
esaG5
Unknown
esaG12
esaG8
esaG3

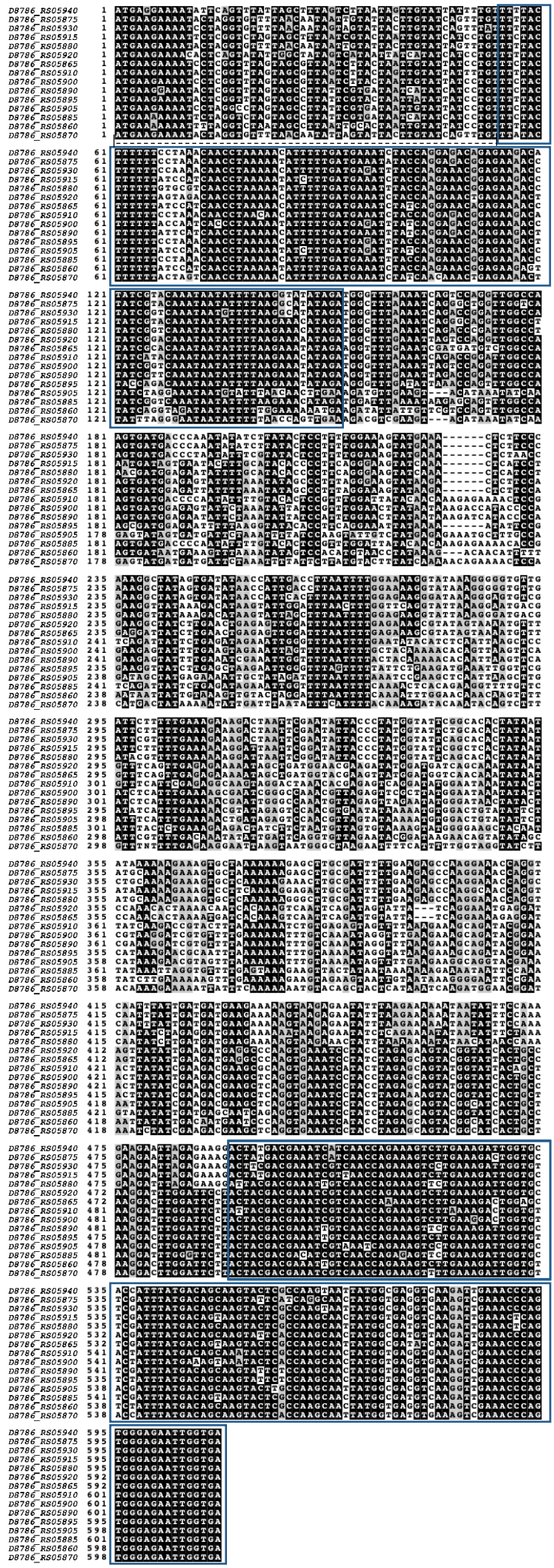
Sup Fig 5


Staphylococcus aureus
21305
O46
CA-347
21334
LGA251
LIT89
GHA13
NCTC6135
PS/BAC/169/17/W
M6K323
BSAC1226
Rrf2 Transcriptional Regulator
Hypothetical Protein
tspA
tsaI
Hypothetical Protein
Aldo/keto reductase
Flavodoxin Family Protein
Pseudogene

Sup Fig 6


tsaI1
tsaI8
tsaI2
tsaI3
tsaI4
tsaI10
tsaI5
tsaI11
tsaI7
tsaI9
tsaI6

Sup Fig 7


D8786_RS05940
D8786_RS05875
D8786_RS05930
D8786_RS05915
D8786_RS05880
D8786_RS05920
D8786_RS05865
D8786_RS05910
D8786_RS05900
D8786_RS05890
D8786_RS05895
D8786_RS05905
D8786_RS05885
D8786_RS05860
D8786_RS05870

Sup Fig 8