Massive genome decay and expansion of insertion sequences drive the evolution of a novel host-restricted bacterial pathogen

Gonzalo Yebra[1,a], Andreas F Haag[2,a], Maan M Neamah[2,3,4], Bryan A Wee[1], Emily J Richardson[1], Pilar Horcajo[5], Sander Granneman[6], María Ángeles Tormo-Más[7,8], Ricardo de la Fuente[5], J Ross Fitzgerald[1,*], José R Penadés[2,7,9,*]

[1]The Roslin Institute, University of Edinburgh, Edinburgh, United Kingdom; [2]Institute of Infection, Immunity & Inflammation, University of Glasgow, Glasgow, United Kingdom; [3]Faculty of Veterinary Medicine, University of Kufa, Kufa, Iraq; [4]Middle Euphrates Centre for Cancer and Genetic Research, University of Kufa, Kufa, Iraq; [5]Facultad de Veterinaria, Universidad Complutense de Madrid, Madrid, Spain; [6]Centre for Synthetic and Systems Biology, University of Edinburgh, Edinburgh, United Kingdom; [7]Departamento de Ciencias Biomédicas, Facultad de Ciencias de la Salud, Universidad CEU Cardenal Herrera, Moncada, Valencia, Spain; [8]Severe Infection Group, Health Research Institute Hospital La Fe, Valencia, Spain; [9]MRC Centre for Molecular Bacteriology and Infection, Imperial College London, London, United Kingdom.

[a] These authors contributed equally.

* Corresponding authors:
JRF (ross.fitzgerald@roslin.ed.ac.uk)
JRP (j.penades@imperial.ac.uk)

Keywords:

Bacterial evolution; host adaptation; host restriction; genome reduction; gene decay; insertion elements; pseudogenes; structural variation; gene expression regulation; phylodynamics.

Abstract

Background. The emergence of new bacterial pathogens represents a major threat to public and veterinary health. *Staphylococcus aureus* is a multi-host bacterial species comprising pathogenic clones with distinct tropisms for human and livestock species. A *S. aureus* microaerophilic subspecies, *Staphylococcus aureus* subsp. *anaerobius*, is responsible for outbreaks of a specific lymphadenitis pathology (Morel's disease) exclusively found in small ruminants. However, the evolutionary history of *S. aureus* subsp. *anaerobius* and its genetic relatedness to *S. aureus* are unknown.

Results. Evolutionary genomic analyses of clinical *S. aureus* subsp. *anaerobius* isolates sampled across 3 decades revealed this clone emerged from a *S. aureus* progenitor about 1000 years ago (95%CI: 716-1184), before differentiating into two distinct lineages representing African (emerged in 1930 [1907-1951)) and European (1777 [1716-1832]) isolates. *S. aureus* subsp. *anaerobius* has undergone limited clonal expansion, with a restricted population size, and an evolutionary rate 10-fold slower than *S. aureus*. The transition to a highly niche-specific pathogen of small ruminant lymph nodes involved acquisition of a pathogenicity island encoding an effector with ruminant host-specificity, large genomic rearrangements, and the accumulation of at least 205 pseudogenes resulting in a highly fastidious metabolism underpinning its restricted ecological niche. Importantly, acquisition and expansion of ~87 insertion sequences located in conserved intergenic regions provided distinct mechanisms for the control of expression of flanking genes, representing a novel concept of transcriptional regulon.

Conclusions. Our findings provide a remarkable example of the evolutionary trajectory of a host-restricted bacterial pathogen that resulted from extensive remodelling of the *S. aureus* genome.

Background

Bacteria are particularly able to adapt to new environmental niches, a characteristic that may promote their capacity to become highly successful pathogens. Many pathogens can infect multiple hosts [1], but occasionally strains of free-living bacteria become restricted to one particular host. The evolutionary process of host restriction, observed across distantly-related bacterial groups, typically involves one or more processes such as genome reduction, acquisition of mobile genetic elements and insertion sequences, and chromosomal rearrangements [2-4]. The magnitude of these changes may be most manifest after long-term associations between bacterium and the host, with obligate intracellular pathogens the most extreme of these cases exhibiting the smallest and most compact genomes [5].

Staphylococcus aureus is an important and versatile bacterial pathogen, capable of infecting a broad array of animals [6], and thus representing a threat for public and livestock health. Understanding the evolutionary events underpinning successful host switches may assist the design of strategies to contain the emergence of new human and animal pathogenic clones. In some cases, the genetic determinants involved in host adaptation have been identified. While some virulence factors encoded in certain mobile genetic elements play an important role in the adaptation of *S. aureus* to ruminants and poultry [7, 8], a single mutation is sufficient to allow the *S. aureus* human-to-rabbit host jump [9, 10]. These diverse strategies highlight the versatility that *S. aureus* has in adapting and infecting new hosts.

A close relative of *S. aureus*, *S. aureus* subsp. *anaerobius*, responsible for Morel's disease [11], is highly restricted to small ruminants and its clinical manifestation is confined to abscesses located near major superficial lymph nodes. Moreover, and contrary to *S. aureus* subsp. *aureus*, *S. aureus* subsp. *anaerobius* is microaerophilic and catalase-negative [12]. Importantly, outbreaks of this disease have been detected in Europe, Africa and the Middle East, causing serious economic losses [13].

Here, we investigate the evolutionary history of *S. aureus* subsp. *anaerobius*, revealing that it descended from an ancestor resembling contemporary *S. aureus* and adapted to its current

ecological niche by extensive genome diversification through a number of processes including massive genome decay, gene acquisition, and genome rearrangements. Of note, the emergence of an IS regulon, involving regulatory control of almost 90 chromosomal genes by an ISSau8-like insertion sequence, represents a novel mode of transcriptional remodelling that accompanied the transition to a new niche. Overall, our results reveal a new strategy used by *S. aureus* to adapt to a new host, highlighting the extraordinary versatility of this important pathogen.

Results

S. aureus subsp. *anaerobius* exhibits a highly conserved genome with limited strain-dependent variation

The genomic architecture of *S. aureus* subsp. *anaerobius* remains largely unknown since only one fragmented draft genome was available until now, corresponding to the strain ST1464 isolated in Sudan [14]. In order to provide a fully assembled genome sequence, we applied PacBio sequencing to the type strain MVF7, isolated in Central Spain in 1981-1982 [12]. We obtained a closed genome which comprised a circular chromosome of 2,755,024bp with 32.74% GC content (**Fig. 1A**) and no plasmids. This genome contained 2,888 coding DNA sequences (CDS), 56 tRNAs and 5 complete rRNA copies.

In order to explore the population structure and the variability in gene content of *S. aureus* subsp. *anaerobius*, we performed whole genome sequencing using Illumina on a collection of 39 additional strains from different countries [13]. Of them, 30 were sampled in Spain during different outbreaks spanning 3 decades (1981-2012). The rest of the isolates, sampled between 1996 and 2011, were from Sudan (n=3), Italy (n=3), Poland (n=2) and Denmark (n=1). These newly generated assemblies were analysed together with that of the ST1464 isolate. The 39 draft assemblies presented an average length of 2,683,726bp (range: 2,612,583-2,749,895bp). The pan-genome analysis showed that the core genome (i.e. genes shared by all isolates) included 2,479 (92.7%) of the total 2,675 genes found. Of the remaining,

strain-specific genes, 66 (2.5%) were “soft core” genome (present in 95-99% of sequences), and the rest (26, 4.8%) were mostly truncated genes in assemblies with low quality. Interestingly, 15 genes were present in European isolates but absent in Sudanese ones (including 2 clusters of 4 genes each) and 5 genes vice versa.

Phylogenetic analyses reveal two distinct sub-lineages of *S. aureus* subsp. *anaerobius*

To explore the population structure of *S. aureus* subsp. *anaerobius*, we generated a core genome alignment including all our samples and the only publicly available draft assembly to make a total of 41 genomes. In this alignment we detected 5 putative recombinant regions (size range: 54-2,950bp) that were excluded from the core genome alignment. The definitive alignment was 2,324,010bp long and included 3,443 variable sites, and it was used for further phylogenetic analysis.

The phylogenetic analyses of the core genome alignment revealed two distinct lineages within *S. aureus* subsp. *anaerobius* – one including the isolates sampled in Sudan and the other one including the rest of isolates, all sampled in Europe (**Fig. 2**). The core alignment showed that the average number of SNPs between pairs of *S. aureus* subsp. *anaerobius* genomes was 494 SNPs (range: 0-1,746). This value was 231 (0-450) within the European clade and 97 (2-142) within the Sudanese clade. The average distance between these two clades was 1,697 SNPs (1,663-1,746). For reference, the average number of SNPs between the *S. aureus* subsp. *anaerobius* isolates and the *S. aureus* RF122 isolate was 27,611 (27,582-27,650).

In contrast to previous results that ascribed all *S. aureus* subsp. *anaerobius* isolates to a single, homogenous sequence type (ST1464) [13, 15], all Sudanese sequences belonged to the same ST but we found 3 other STs within European sequences (see **Additional Information** in Additional File 1 for more detail).

Linear regression analysis provided evidence of a molecular clock-like evolution in the ML phylogeny. The subsequent time-scaled Bayesian phylogenetic analysis showed that the model combination that best fitted the data was a strict molecular clock paired with a constant

127 population size. This analysis dated the most recent common ancestor (MRCA) of *S. aureus*
128 subsp. *anaerobius* in the year 1069 AD (95% Bayesian confidence interval (BCI): 828-1296)
129 (**Fig. 2**). The European and Sudanese lineages diversified in the years 1777 (95% BCI: 1716-
130 1832) and 1930 (95% BCI: 1907-1951), respectively. The estimated evolutionary rate was 4.4
131 (95% BCI: 3.3-5.5) × 10^{-7} substitutions/site/year (s/s/y), which approximates to 1.2 SNPs per
132 year. This is roughly one order of magnitude slower than *S. aureus* subsp. *aureus*. The rest of
133 models tested in BEAST provided very similar results, with mean MRCA estimates between
134 1111 and 1333 AD and mean evolutionary rates between 4.2 and 5.4 × 10^{-7} s/s/y. This agrees
135 with the slow *in vitro* growth observed for *S. aureus* subsp. *anaerobius* [11], as evolutionary
136 rates are proportional to doubling times [16].

137 To ascertain the phylogenetic relationship between *S. aureus* subsp. *anaerobius* and
138 *S. aureus* subsp. *aureus,* we constructed a core genome SNP tree with 807 whole-genome
139 sequences representative of the *Staphylococcus* diversity [17]. In this tree (**Fig. 3**), *S. aureus*
140 subsp. *anaerobius* formed a well-defined cluster within *S. aureus* subsp. *aureus* but without
141 being genetically close to any clonal complex. This phylogenetic evidence reinforces the idea
142 that this subspecies evolved and diversified from an unknown *S. aureus* lineage after a host
143 adaptation and restriction process that started at least a millennium ago. Although there has
144 not been any study specifically designed to estimate the *S. aureus* subsp. *aureus* emergence
145 time, a recent article [17] which analysed a globally diverse dataset found a root date of 7,000
146 BC – approximately 10 times older.

147 **Massive gene decay characterises the *S. aureus* subsp. *anaerobius* genome**

148 The complete genome of *S. aureus* subsp. *anaerobius* strain MVF7 contained 205
149 pseudogenes, with an average of 204.9 pseudogenes per genome across all *S. aureus*
150 subsp. *anaerobius* strains (range: 201-210). On average, 66.6% (range: 64.7-68.8%) of the
151 pseudogenes were split CDSs and 33.4% (range: 31.2-35.3%) were truncated CDSs. These
152 pseudogenes were evenly distributed across the genome (**Fig. 1A**), and originated through
153 point mutations causing frameshifts, premature stop codons or alternative downstream start

154 codons. Pseudogenes represented an average of 9.5% (range: 9.2-10.1%) of the genome
155 length across all *S. aureus* subsp. *anaerobius* genomes.

156 A total of 164 pseudogenes (approximately 80% of those found in any given isolate) were
157 shared among all *S. aureus* subsp. *anaerobius* genomes. Of those pseudogenes, 92.1% were
158 caused by the same mutation in all isolates suggesting they were an early event in its
159 evolutionary history. The proportion of shared pseudogenes was higher within each of the two
160 main phylogenetic clades described above: Sudanese isolates shared 183 pseudogenes
161 (89%) and European isolates shared 179 (87%).

162 The functions of the 205 pseudogenes present in MVF7 were classified into clusters of
163 orthologous groups (COGs) of proteins (for a more detailed description, see **Additional**
164 **Information** in Additional File 1 and **Table S1** in Additional File 2). Whereas 43 (26.2%) had
165 unknown or unavailable functions, half of these proteins (n=82, 50%) corresponded to
166 metabolic functions, 29 (17.7%) to cellular processes and signalling, and 10 (6.1%) to
167 information storage and processing. Within the pseudogenes affecting metabolic functions,
168 most were associated with amino acids (n=29, 32.6%), carbohydrates (n=14, 15.7%),
169 inorganic ions (n=9, 10.1%) and functions affecting energy production (n=8, 9.0%). The
170 enrichment analyses of both KEGG pathways and GO terms (see **Table S2** in Additional File
171 2 for a full list) revealed a statistically significant pseudogenisation of genes involved in
172 biosynthesis of amino acids and metabolic pathways.

173 Although most pseudogenes were involved in numerous metabolic pathways, others were
174 defence mechanisms (such as catalase and other oxidoreductases, transmembrane pumps
175 involved in antibiotic resistance and fluoride toxicity, and type I restriction enzymes), and
176 virulence factors (such as proteins involved in capsule biosynthesis, adherence factors and
177 exoproteins). In short, genes affected by pseudogenisation could explain the limited
178 metabolism and pathogenicity of this bacterium.

179

180 ***S. aureus* subsp. *anaerobius* contains novel mobile genetic elements**

181 All *S. aureus subsp. anaerobius* sequences carried a 43.2 kb-long, novel prophage (ΦSaa1)
182 belonging to the *Siphoviridae* family (**Fig. 1B**), which also contained many pseudogenes when
183 compared to the closest available prophage (Φ2958PVL, accession number NC_011344).
184 The prophage presented a novel integrase type (see **Fig. S1** in Additional File 1) and a
185 pseudogenised variant of the virulence-associated protein E, *virE.*

186 We also identified one novel 13 kb Staphylococcal Pathogenicity Island (SaPI) present in all
187 isolates, which encoded additional copies of the Staphylococcal complement inhibitor (SCIN)
188 and the von Willebrand factor-binding protein (vWbp) (**Fig. 1B**). vWbp has previously been
189 demonstrated to show host-specific activity, being able to coagulate ruminant plasma [7, 8]. A
190 phylogenetic network using SplitsTree v4.14.6 and the NeighborNet algorithm [18] of
191 SaaPIMVF7 and previously described SaPIs (**Fig. S2** in Additional File 1) showed that
192 SaaPIMVF7 is closest to SaPIov2 (93.1% nucleotide identity), firstly described in the *S. aureus*
193 subsp. *aureus* clonal complex CC133 (responsible for infections of small ruminants including
194 sheep and goats) which also carries vWbp and SCIN (**Fig. 1B**). SaaPIMVF7 contained
195 pseudogenes including the integrase and primase genes, suggesting that it is no longer mobile
196 but stable in the chromosome. Of note, *S. aureus* subsp. *anaerobius* also has chromosomal
197 copies of *vwb* and *scn* of which the former is pseudogenised. SaaPIMVF7-encoded *vwb* and
198 *scn* were transcribed (**Fig. 1C**) and apparently functional in SaaPIMVF7 suggesting a key role
199 in the pathogenicity of *S. aureus* subsp. *anaerobius.* The SaaPIMVF7-encoded *vwb* was
200 cloned into the expression vector pCN51, and the ability of the expressed protein to coagulate
201 ruminant plasma confirmed (**Fig. S3** in Additional File 1), verifying its role in virulence.

202 **Insertion sequences drive *S. aureus* subsp. *anaerobius* evolution**

203 Yet another unique feature of the *S. aureus* subsp. *anaerobius* genomes was the presence of
204 87 insertion sequences in MVF7, all identified as ISSau8 (IS family ISL3 described in the strain
205 RF122) by ISsaga [19]. The average nucleotide similarity between all was 99.3%, and that to
206 ISSau8 was 97.0%. Most of them had internal stop codons that split the transposase-coding

207 region in two or more fragments whereas only 10 ISs contained full-length transposase genes.
208 ISs were distributed evenly throughout the genome (**Fig. 1A**).

209 Since *S. aureus* subsp. *anaerobius* has undergone massive gene decay, we initially
210 hypothesised that ISs had caused pseudogenes, inactivating genes not required in the new
211 niche. However, all the ISs were located in intergenic regions—except only one inserted within
212 a gene (*msb*A), resulting into a split CDS (i.e. pseudogene). Since only ~16% of the *S. aureus*
213 subsp. *anaerobius* genome corresponds to non-coding intergenic regions, the non-random
214 distribution of the ISs through the genome suggested an additional role for these elements in
215 the evolutionary adaptation of this species. We hypothesised that *S. aureus*
216 subsp. *anaerobius* uses ISs to control the expression of the genes located downstream of the
217 IS elements. These would represent a novel concept of regulon, in which a subset of genes
218 would be controlled not by a transcriptional regulator but by an IS.

219 To test this idea, we studied the conservation of IS locations across all the *S. aureus*
220 subsp. *anaerobius* strains. Because their high number and genetic similarity (>99% nucleotide
221 identity) caused frequent contig breaks in the draft genomes assembled from short-reads, ISs
222 were only reliably reconstructed in the MVF7 close genome sequence. We tested the
223 presence of each MVF7 IS in the rest of isolates by mapping their respective Illumina reads
224 against the border between each IS and their flanking gene, and identifying reads overlapping
225 this boundary. This analysis revealed that 68 out of the 87 MVF7 ISs (78.2%) were inserted
226 in the same locations (i.e. we detected reads mapping against both IS flanks) in all *S. aureus*
227 subsp. *anaerobius* samples. We observed differences between the European and Sudanese
228 clades: 75 (86.2%) ISs shared genomic location among all the European isolates, whereas 7
229 were missing in all Sudanese isolates. These data suggest that IS have been important not
230 only for the initial host jump, but also for the origin and divergence observed between clades
231 (**Fig. 2**). Interestingly, one of the ISs detected in all European isolates but absent in all
232 Sudanese ones was flanking one of the translocations (number V) described below, which
233 agrees with the fact that such translocation was not detected in the Sudanese strains.

234 Our prediction of promoter and transcriptional terminator (TT) sequences showed that all IS
235 had a functional promoter but not a TT between the *tnp* gene and the downstream gene,
236 supporting the idea that the *tnp* transcript could influence expression of the gene located 3' of
237 the IS gene. This also reported the existence of two additional putative TTs in the IS genes:
238 one bidirectional TT (TT1) located in the 5' region of the IS gene, and a unidirectional one
239 (TT2, antisense orientation) located at the end (3' region) of the IS (see **Fig. 4**). The
240 bidirectional TT1 would putatively prevent the potential interference between transcripts
241 initiated either from the IS or from the gene located upstream to the IS. By contrast, TT2 would
242 prevent the impact that an antisense transcript originated from a gene located downstream of
243 the ISs could have in the expression of the IS transcript (**Fig. 4**). This organisation suggests
244 that the ISs have established a strategy that allows their transcripts to interfere with the genes
245 located in position 3' to the IS, while blocking any external interference that the flanking genes
246 could have in the expression of the regulatory IS transcript.

247 To validate the functionality and directionality of the TTs, we used an expression construct
248 based on the plasmid pCN42 [20], in which the *bla*Z reporter is under the control of the P_{cad}
249 promoter. The conserved 5' and 3' region of the ISs containing the bi- and unidirectional TTs,
250 respectively, were cloned, in both orientations, between the P_{cad} promoter and the *bla*Z
251 reporter (see **Fig. 4**), and the expression of the reporter gene tested with and without induction
252 of the P_{cad} promotor. The results confirm the existence, functionality and directionality of both
253 TTs, suggesting that the role of the ISs is to control the expression of the chromosomal genes
254 located 3' to the ISs.

255 **IS affect *S. aureus* subsp. *anaerobius* gene regulation by interference with promoter**
256 **region and operon structure**

257 ISs are inserted into the *S. aureus* subsp. *anaerobius* either upstream or downstream of
258 targeted genes. To evaluate their impact on the expression of neighbouring genes, we initially
259 analysed whether the presence of an IS controls the expression of the downstream gene
260 located in the same orientation as the IS. Since *S. aureus* subsp. *anaerobius* is not genetically

261 tractable, we designed reporter constructs mimicking the genetic organisation found in
262 *S. aureus* subsp. *anaerobius* (**Fig. S4** in Additional File 1) and tested expression levels of the
263 selected genes or gene products in *S. aureus* subsp. *aureus*. Specifically, we selected four
264 different genes (*rps*P, *dna*D, *met*C and *met*N) that contained an upstream IS at different
265 distance to each gene's start codon and constructed *bla*Z transcriptional reporter fusion using
266 pCN41 [20] (**Fig. 5A-D**). Then we tested the expression of these genes in presence or
267 absence of the ISs. Importantly, and in support of the hypothesis, deletion of the ISs affected
268 expression of the neighbouring genes, confirming their regulatory role, although this effect was
269 different depending on the genes. Thus, while removal of the ISs increased transcription of
270 *dna*D, *met*C or *met*N, it decreased *rps*P expression (**Fig. 5A-D**). Based on these results, it is
271 tempting to speculate that the impact of the IS on downstream gene expression depends on
272 the position of the IS relative to the gene's start codon and the nature of its integration
273 (conservative or deleterious, see **Fig. S4** in Additional File 1). In some cases, the presence of
274 the IS will reduce expression by altering either the promoter region or the DNA binding region
275 of a putative inducer of the expression of the targeted gene, while in other cases it may
276 increase transcription by eliminating the binding site of putative repressors. These ideas are
277 now under study.

278 **ISs affect *S. aureus* subsp. *anaerobius* gene expression through a novel mechanism of**
279 **anti-antisense RNA interference**

280 Most ISs were located downstream and in antisense orientation to putative target genes. Since
281 the previously identified mechanism cannot be involved in the IS-mediated control of the
282 expression of the flanking genes, we reasoned that additional strategies of gene control
283 involving ISs should occur. To test this, we selected two genes located in an antisense
284 orientation to the ISs and tested whether expression from the IS-contained promoter would
285 actively influence their expression (**Figs. S4** and **S5** in Additional File 1). We selected the
286 global regulator-encoding gene *mgr*A, which is a major effector of *S. aureus* gene regulation
287 [21-23], and *adh*A encoding a zinc-dependent alcohol dehydrogenase, which is induced under

288 low oxygen environmental conditions in a RexA-dependent manner [24]. Since the expression
289 of these flanking genes cannot be monitored using transcriptional fusions, they were cloned
290 so they expressed 3xFLAG versions of the encoded proteins, which were used to monitor the
291 expression of these proteins by western blot. To test the impact of the ISs, some of the
292 plasmids contained the flanking IS (wild-type structure), while we generated as controls
293 plasmids without the ISs (see scheme in **Fig. S5** in Additional File 1). Surprisingly, no
294 differences in the expression levels of MgrA or AdhA were observed in any of the constructs
295 (**Figs. S5A** and **S5B**, respectively, in Additional File 1) suggesting that the presence of ISs did
296 not actively alter expression levels of the downstream genes by a classical anti-sense
297 mechanism.

298 However, the analysis of the genetic context of *mgr*A revealed another interesting possibility
299 involving ISs in controlling the expression of this gene. In the ancestor *S. aureus* subsp.
300 *aureus*, the gene downstream of *mgr*A has an antisense orientation (**Fig. S4** in Additional File
301 1) and the transcripts from this gene could interfere with the expression of *mgr*A. This is an
302 interesting hypothesis since overlapping antisense transcripts for *mgr*A and its downstream
303 gene have been identified in several *S. aureus* subsp. *aureus* strains in a previous study [25].
304 Since as previously demonstrated the ISs have different functional transcriptional terminators
305 (**Fig. 4**), we hypothesized that the presence of the ISs between the two contiguous genes
306 could block this interference (antisense) mechanism. To assess whether this was the case,
307 we constructed another set of reporters (see scheme in **Fig. 5E**) in which we used the inducible
308 expression plasmid pCN51 to mimic expression from the gene adjacent to *mgr*A. Since the
309 gene located downstream of *mgr*A belongs to a large operon which was unfeasible to clone
310 completely in a plasmid, we used the cadmium-inducible promoter present in plasmid pCN51
311 as the origin of the anti-sense mRNA for *mgr*A, mimicking the natural genetic context. The
312 different constructs were introduced into the *S. aureus* subsp. *aureus* strain RN4220 Δ*spa* and
313 expression from the cadmium-inducible pCN51 promoter induced. Absence of the ISs
314 significantly reduced MgrA expression levels (**Fig. 5E**), compared to the reporter construct

315 where these were present. This suggests that the IS prevents transcripts from the upstream
316 gene through the presence of a strong transcriptional terminator and thus protects *mgr*A
317 mRNA from antisense mRNA interference. In support of this idea, insertion of a TT between
318 *mgr*A and the plasmid-encoded cadmium-inducible promoter produced the same effect as the
319 presence of the complete IS (**Fig. 5E**).

320 To test whether this organisation affected the regulation of an IS targeted gene in its native
321 context containing a downstream gene and its promoter in antisense orientation with known
322 overlapping transcription [25], we selected *sle1* (aka *aaa*) encoding an N-acetylmuramyl-L-
323 alanine amidase involved in staphylococcal cell separation [26] and essential for β-lactam
324 resistance of *S. aureus* [27]. The gene downstream of *sle1* is expressed as a single transcript
325 from its own promoter allowing us to construct a set of reporters similar to the constructs
326 prepared for *mgr*A in the plasmid pCN47 [20]. Note that in this scenario all genes are
327 expressed using they native promoter regions. Since the potential expression patterns of the
328 gene downstream of *sle1* were unknown, we assessed the impact of the IS on the expression
329 levels of Sle1 at different stages during bacterial growth (**Fig. 5F**). Similar to the results
330 observed for MgrA, absence of the IS significantly reduced Sle1 expression levels albeit in a
331 growth phase-dependent manner. Thus, certain ISs in *S. aureus* subsp. *anaerobius* uncouple
332 genes from ancestral antisense regulation by acting as anti-antisense elements, a hitherto
333 unprecedented regulatory role.

334 ***S. aureus* subsp. *anaerobius* has undergone large genomic rearrangements**

335 The widespread presence of ISs in *S. aureus* subsp. *anaerobius* caused promote genomic
336 rearrangements by homologous recombination, a phenomenon observed in different bacterial
337 species [28]. When compared to *S. aureus* subsp. *aureus*, the genome of *S. aureus*
338 subsp. *anaerobius* contained at least six large genomic translocations, flanked by ISs, whose
339 size ranged between 70 and 346 kbp (**Fig. 6** and **Fig. S6** in Additional File 1). Two pairs of
340 translocations (I and VI, and II and V) switched their original genomic locations whereas the
341 rest (III and IV) were produced by excision and insertion. These translocations interchanged

342 regions between the leading and lagging strands, but because all also involved inversion,
343 genes in the translocated areas conserved their original orientation and the GC skew remained
344 unaffected. Translocations did not produce any noticeable loss of genomic regions.

345 We designed a series of primer pairs to qualitatively test whether the same breakpoints
346 between translocations were present in all the *S. aureus* subsp. *anaerobius* strains. The PCR
347 results, summarised in **Table S3** in Additional File 2, showed that the genomic rearrangements
348 observed in the MVF7 full genome were also present in the rest of European isolates, whereas
349 the isolates from Sudan lacked the translocations numbers II and V. This agrees with the fact
350 that one of the ISs delimiting the translocation number II was one of those present in all
351 European isolates but absent in all Sudanese ones. It is therefore likely that this particular
352 translocation happened after these two phylogenetic lineages diverged.

353 Although there were some cases of PCR pairs failing in certain European isolates, which might
354 be due to the DNA integrity, in general this shows how some of the genomic rearrangements
355 might have occurred after the European and African lineages separated (**Table S3** in
356 Additional File 2).

357 In summary, our results demonstrate an extraordinary role of the ISs in *S. aureus* subsp.
358 *aureus,* not just controlling the expression of many chromosomal genes, but also shaping the
359 structure of the *S. aureus* subsp. *aureus* chromosome.

360 Discussion

361 Here, we demonstrate that *Staphylococcus aureus* subsp. *anaerobius* represents a
362 remarkable example of the drastic genomic changes affecting a bacterial pathogen during the
363 evolutionary process of host restriction. This process involved an outstanding variety of
364 simultaneous evolutionary strategies such as mobile genetic element acquisition, pseudogene
365 accumulation, large genomic rearrangements and, importantly, a novel transcriptional
366 regulatory system orchestrated by widespread insertion elements.

This bacterium was described as a respiratory deficient variant of *S. aureus* due to its lack of aerobic growth and catalase activity [12]. While both are very closely related, their pathogenic abilities are radically different: *S. aureus* is a major pathogen responsible for acute and chronic infections both in humans and animals, while *S. aureus* subsp. *anaerobius* has only been occasionally identified when causing a specific lymphadenitis of sheep and goats. In order to reconstruct the evolutionary trajectory that transformed *S. aureus* subsp. *anaerobius* into a host- and niche-specific pathogen, we performed the first population genomics study by analysing 40 isolates sampled in different countries over the last 30 years. Our results suggest that *S. aureus* subsp. *anaerobius* evolved from a ruminant-adapted *S. aureus* strain (containing specialised mobile genetic elements) through a process of host restriction that modified drastically its genome. Such modifications accumulated gradually during restriction to a novel host of a once free-living bacterium [29], are recurring across the bacterial kingdom. They have been described in diverse lineages such as *Mycobacterium* [30], *Shigella* [31], *Salmonella* [32] and *Burkholderia* [33] (see also review in [2]).

Host shifts represent drastic, sudden changes in the environmental conditions for a bacterial population, associated with genetic bottlenecks which cause a reduction in population size and a weakened purifying (negative) selection [2, 29]. This scenario would explain the fact that one tenth of the *S. aureus* subsp. *anaerobius* genome is made up of pseudogenes, due to an incapacity to clear deleterious mutations that cause gene truncation and frameshifts, which are fixed then by genetic drift. The conservation of pseudogenes could be explained by a weakened purifying selection (which prevents purging these truncated genes no longer needed) as a consequence of a reduced population size and/or by the fact that the extreme adaptation of *S. aureus* subsp. *anaerobius* is a recent event in the evolutionary scale and that selection is limited by a reduced mutation rate. Both explanations agree with our phylodynamic analyses, which shows a relatively recent date of emergence (a thousand years ago), a restricted population size growth, and a reduced evolutionary rate.

393 The pseudogenisation of genes involved in many metabolic pathways, defence mechanisms,
394 and virulence factors explains the atypical phenotype of *S. aureus* subsp. *anaerobius* when
395 compared to *S. aureus*, including its limited pathogenicity, reduced metabolism and fastidious
396 growth *in vitro*. In addition, its microaerophilic metabolism, caused by a lack of defence against
397 radicals (by truncation of the catalase and others) rather than disruption of oxidative
398 respiration, suggests that this bacterium has become confined to a nutrient-rich, oxygen-
399 limited niche. This explanation agrees with the fact that it can only be cultured from the benign
400 lymph node abscesses it causes [12].

401 An ineffective purifying selection also allows insertion sequences (IS), normally present in
402 bacterial genomes but in low numbers (<10), to proliferate and spread across the genome in
403 the early stages of host-restriction only to be purged later on –therefore recently adapted
404 bacteria will harbour more ISs than more ancient symbionts [2, 4]. ISs will normally cause
405 gene inactivation (through insertion within coding regions) and chromosomal rearrangements
406 of genetic segments flanked by ISs.

407 The latter is true for *S. aureus* subsp. *anaerobius*, with five large rearrangements originated
408 in the common ancestor of all isolates and an additional one exclusive to the European clade.
409 Effects of genomic rearrangements are difficult to establish although some of the ones
410 described include reduction in growth rate (again, a characteristic phenotype in *S. aureus*
411 subsp. *anaerobius)* and colony variation [34]. In addition, large rearrangements can change
412 the distance of a gene from the origin of chromosome replication leading to altered gene copy
413 number and thus affecting its expression [34, 35]. During the initial analysis of the *S. aureus*
414 subsp. *anaerobius*, and based in the high number of pseudogenes and ISs present in this
415 species, we hypothesised that most of the pseudogenes would be consequence of the
416 presence of an IS. A more detailed analysis revealed that this idea was wrong, and only one
417 However, only one of the ISs in *S. aureus* subsp. *anaerobius* caused inactivation of one gene,
418 while the rest (n=86) were inserted in intergenic regions, a non-random distribution that we
419 hypothesised was a key feature in the evolutionary origin of *S. aureus* subsp. *anaerobius*. It

420 has been extensively reported that individual ISs can activate expression of the neighbouring
421 genes, either by an extended transcription from an internal promoter or by the generation of a
422 hybrid promoter [28, 36]. Consequently, domestication of ISs by the bacterium to control
423 specific functions is an evolutionary process that has been unexpectedly common. An
424 example of this is *Bordetella pertussis*, with more than 230 copies of the IS481 in each genome
425 [37], most of which also influence the expression of their flanking genes [38]. In this case,
426 however, the transcripts produced by the ISs are strain-specific due to the dynamic variation
427 in IS copy number and genomic localisation [38]. By contrast, the localisation of most of the
428 ISs present in *S. aureus* subsp. *anaerobius* is conserved in all the strains analysed, indicating
429 they control a set of genes important in the adaptation to the new host. Thus, the insertion of
430 most of these ISs occurred early in the *S. aureus* subsp. *anaerobius* evolution, being
431 subsequently inherited and inactivated (note that most of the *tnp* genes were pseudogenes).
432 This represents a new concept of IS regulon, were a set of genes is transcriptionally controlled
433 by the presence of an ISs. Interestingly, the IS-mediated regulation involves at least two
434 different mechanisms of control, depending on whether the genes flanking the IS are in sense
435 or antisense orientation in relation to the IS. While upstream IS oriented in sense to the flanking
436 gene (~25% of ISs) act mainly through modulation of promoter and operon structure, the
437 majority of ISs (~75%) is located downstream and in antisense orientation to their flanking
438 genes. In the ancestral *S. aureus* subsp. *aureus* strain, the expression of many of these genes
439 is controlled through antisense mechanisms caused by transcripts of a downstream located
440 gene in antisense orientation. This ensures that expression levels of both genes can be
441 modulated under different environmental conditions, allowing *S. aureus* subsp. *aureus* to
442 adapt to changing environments. By contrast, in *S. aureus* subsp. *anaerobius*, the integration
443 of an IS in antisense orientation to flanking genes can effectively act as anti-antisense
444 regulatory elements and uncouple the targeted genes from their native interdependent
445 expression control. This new regulatory mechanism might provide a selective advantage to
446 *S. aureus* subsp. *anaerobius* in its more restricted and consequently more stable host niche

447 enabling the simultaneous expression of both genes. This strongly suggests that expression
448 of both proteins is required at the same time in this specialised pathogen.

449 Additionally, the presence of some ISs whose location was strain-dependent suggests that
450 some ISs remain functional in order to generate more variation and/or regulation. Indeed,
451 during its differentiation into two lineages in Africa and Europe a small proportion of ISs
452 integrated in different genomic locations in these lineages. The absence of a few ISs in certain
453 non-phylogenetically related isolates could be due to the limitations of the detection approach,
454 or alternatively imply that ISs could be occasionally lost in some isolates.

455 Finally, our comparative genomics analysis based on newly generated whole genome
456 sequences unveiled the previously unknown population structure of *S. aureus*
457 subsp. *anaerobius* and revealed the existence of two distinct phylogenetic clades containing
458 the African and European lineages, respectively, which diverged a few centuries ago.
459 However, most of the evolutionary events that shaped the genome of this bacterium (including
460 MGE acquisition, IS insertion, widespread pseudogenisation, and genomic rearrangements)
461 took place before this diversification and were very likely key processes in the initial adaptation
462 of *S. aureus* subsp. *anaerobius* to its current niche in the sheep host.

463 In conclusion, *S. aureus* subsp. *anaerobius* represents a remarkable example of drastic
464 genomic changes during a relatively recent host and niche adaptation from a versatile and
465 generalist ancestor –an evolutionary journey notably marked by accumulation of pseudogenes
466 and the development of a new concept of regulon orchestrated by insertion elements found in
467 conserved genome localisations.

468 Methods

469 Bacterial strains and growth conditions

470 The bacterial strains used in this study are detailed in **Table S4** in Additional File 2. *S. aureus*
471 strains were grown in Tryptic soy broth (TSB) or on Tryptic soy agar plates and *Escherichia*
472 *coli* strains were grown in Luria-Bertani broth (LB) or on LB agar plates. Antibiotic selection

473 was used where appropriate (erythromycin 10 µg ml^{-1} for *S. aureus* strains and ampicillin
474 100 µg ml^{-1} for *E. coli* strains).

475 **Sample sequencing and gene content analysis**

476 Forty *S. aureus* subsp. *anaerobius* ovine isolates previously reported [13] were analysed, of
477 which 31 were sampled in Spain during different outbreaks spanning 3 decades (1981-2012).
478 The rest of the isolates, sampled between 1996 and 2011, were from Sudan (n=3), Italy (n=3),
479 Poland (n=2) and Denmark (n=1).

480 They were sequenced using a MiSeq platform ((Illumina, San Diego, CA, USA); the paired-end
481 short reads were trimmed using Trimmomatic v0.36 [39] and *de novo* assembled into contigs
482 using SPAdes v3.10.0 [40] and Velvet v1.2.10 [41]. Illumina read data is available in the
483 European Nucleotide Archive under the study accession number PRJEB30965.

484 One of the Spanish isolates (MVF7, the type strain) was sequenced on the RSII platform
485 (Pacific Biosciences, Menlo Park, CA, USA) using SMRT technology in the Centre for
486 Genomic Research (University of Liverpool, UK). The long reads were de novo assembled
487 into one, full-length contig using the Hierarchical Genome Assembly Process (HGAP) method.

488 The only *S. aureus* subsp. *anaerobius* draft genome publicly available, isolate ST1464
489 (assembly GCA_000588835.1) [14] was included in the analyses.

490 Genes of all 41 genomes and draft assemblies were annotated using Prokka v1.12 [42]. The
491 pan-genome was calculated using Roary v3.8.2 [43] applying a 95% identity cut-off. *In silico*
492 multilocus sequence typing (MLST) of the assemblies was performed using the MLST tool
493 v2.8 (https://github.com/tseemann/mlst). Phage sequences and genomic islands were
494 identified using PHASTER [44] and Island Viewer v4 [45], respectively, and manually
495 inspected for reliability. Antimicrobial resistance genes were detected using ResFinder v3.0
496 [46].

497

498

499 **Pseudogene detection**

500 Pseudogenes are gene copies that have lost their protein-coding ability. Here, pseudogenes
501 were defined as protein-coding sequences that had homologous genes in an *S. aureus* subsp.
502 *aureus* reference and that were split or truncated (i.e. <80% of the reference length) with
503 respect to that reference. For this purpose, the isolate RF122 (assembly GCA_000009005.1)
504 was used as reference and a custom python script was developed to identify pseudogenes
505 from the output of a Roary analysis that compared the genetic content of each *S. aureus*
506 subsp. *anaerobius* to the RF122 genome. The analyses were repeated using different
507 *S. aureus* subsp. *aureus* genome references, but this did not alter significantly the
508 pseudogenes detected (data not shown).

509 Functions of pseudogenes were investigated by assigning them to clusters of orthologous
510 groups (COGs) using eggNOG v4.5.1 [47]. Enrichment analyses of GO (Gene Ontology) terms
511 and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways assigned to the
512 pseudogenes by eggNOG were performed using the R packages topGO v2.34.0 and
513 ClusterProfiler v3.10.0 [48]. Additionally, the Integrated Microbial Genomes (IGM) annotation
514 tool [49] was used to infer phenotypic metabolic characteristics from the presence/absence of
515 protein pathways.

516 **Insertion sequence detection and transcription analyses**

517 Transposase coding sequences were identified in the MVF7 full genome using the Prokka
518 output and the ISsaga web tool hosted in the ISfinder platform [19].

519 For the rest of the isolates, *de novo* assembly from Illumina short reads failed to reconstruct
520 the highly conserved and repeated transposase genes with enough specificity, causing contig
521 breaks in those locations. Consequently, few or no transposases could be inferred from their
522 draft assemblies. This is a common issue when trying to resolve repetitive regions using short
523 read data, which causes ambiguity in selecting the best location in the genome to which align
524 such reads [50].

525 Therefore, the presence of ISs was tested by mapping the Illumina reads against them using
526 BWA-MEM [51] to fragments of the MVF7 genome containing each transposase edge together
527 with part of their flanking genes. This was done to detect Illumina reads that overlapped the
528 border between the IS and the flanking gene as indicator of the presence of that IS in that
529 particular genomic location.

530 **Phylogenetic analyses**

531 A core genome alignment was created by aligning the Illumina short reads and the ST1464
532 assembly to the MVF7 whole genome using Snippy v3.1
533 (https://github.com/tseemann/snippy). Sites containing any gap character ('-') or unknown
534 nucleotide ('N') were discarded. Gubbins v2.2.0 [52] was used detect recombinant regions in
535 this alignment, which were discarded.

536 A maximum likelihood (ML) tree was constructed using IQ-TREE [53], applying the GTR
537 nucleotide substitution model together with a gamma-distributed rate heterogeneity across
538 sites and 1,000 ultrafast bootstrap replicates. Temporal signal was investigated using
539 TempEst v1.5 [54] by means of the correlation between root-to-tip distances and sampling
540 dates.

541 Bayesian phylogenetic analysis was performed using BEAST v1.9.0 [55]. The HKY model for
542 nucleotide substitution was used, whereas different models were tested for the molecular clock
543 (strict and uncorrelated lognormal relaxed) and demographic (constant, exponential and
544 Bayesian skygrid) models. Each of these model combinations were run for 100 million
545 generations, with sampling every 10,000 and discarding the initial 10% as burn-in. Runs were
546 compared via a marginal likelihood estimation (MLE) using path sampling and stepping stone
547 sampling methods implemented in BEAST. The posterior distribution of trees was summarised
548 into a maximum clade credibility tree.

549 In order to explore the relationship between *S. aureus* subsp. *anaerobius* and *S. aureus*,
550 another core genome SNP tree was created, also using Snippy v3.1 and IQ-TREE (same

551 settings than above). It included a sample of 787 *S. aureus* genomes (corresponding to 43
552 different animal host species and 77 clonal complexes (CCs), isolated in 50 different countries)
553 along with 17 isolates of the most closely-related staphylococcal species (*S. schweitzeri* and
554 *S. argenteus*) [17].

555 Bioinformatics analyses were carried out using the Cloud Infrastructure for Microbial
556 Bioinformatics (CLIMB) facility [56].

557 **DNA methods**

558 General DNA manipulations were performed using standard procedures. Cloning and
559 assembly of reporter constructs was performed using restriction digest and ligation or Gibson
560 assembly (NEBuilder HiFi DNA Assembly, NEB) according to the manufacturer's instructions.
561 The plasmids and oligonucleotides used in this study are listed in **Tables S5** and **S6** in
562 Additional File 2, respectively.

563 **Western blot**

564 *S. aureus* cultures harbouring the respective pCN47 or pCN51 derivative plasmids were
565 diluted 1:50 from overnight cultures and grown in TSB at 37°C and 120 rpm until sample
566 collection. For pCN51 constructs induced with $CdCl_2$, cultures were split at early exponential
567 phase (OD_{540}~0.15), 5 µM $CdCl_2$ was added to half cultures and incubation continued for
568 another 2 h. *S. aureus* strains carrying pCN47 reporter constructs were grown to early
569 exponential (OD_{540} ~0.15), mid-exponential (OD_{540} ~0.8) or stationary phase (OD_{540} ~2 after
570 overnight culture). A sample amount corresponding to and OD_{540} of 0.5 in 1 ml was pelleted
571 and stored at -20°C. The sample pellets were re-suspended in 100 µl digestion/lysis buffer
572 (50 mM Tris-HCl, 20 mM $MgCl_2$, 30% (w/v) raffinose) plus 1 µl of lysostaphin (12.5 µg ml^{-1})
573 and incubated at 37°C for 1 h. Samples for SDS-PAGE were prepared by adding 4X NuPAGE
574 LDS Sample Buffer (Invitrogen) and proteins denatured at 95°C for 10 min. Samples were
575 separated by SDS-PAGE (NuPAGE™ 4-12% Bis-Tris Protein Gels, Invitrogen and then
576 transferred to a PVDF transfer membrane (Thermo Scientific, 0.2 µM) following standard
577 procedures [57, 58]. FLAG-tagged proteins were detected using mouse anti-FLAG-HRP

578 antibody (Monoclonal ANTI-FLAG® M2-Peroxidase (HRP), Sigma-Aldrich) following the
579 manufacturer's protocol.

580 **Enzyme assay for the quantification of β-lactamase activity in transcriptional fusion**
581 **plasmids**

582 Samples for pCN41 based reporters were grown to either early exponential phase
583 (OD_{540}~0.15) or stationary phase for promoters with low activity (*met*C) (OD_{540} ~2 after
584 overnight culture) and 1 ml of culture snap frozen. β-Lactamase assays, using nitrocefin (BD
585 Diagnostic systems) as substrate, were performed as previously described [59] using an
586 ELx808 microplate reader (BioTek) measuring absorbance at 490 nm. Promoter activity was
587 calculated using the following equation:

$$Promoter\ activity = \frac{dA_{490}}{dt(h)} \frac{1}{OD_{540} * DF * V}$$

588

589 ## Declarations

590 **Ethics approval and consent to participate**

591 Not applicable.

592 **Consent for publication**

593 Not applicable.

594 **Availability of data and materials**

595 Illumina short-read sequence data was deposited at the European Nucleotide Archive with
596 study accession number PRJEB30965. The MVF7 full genome assembly was deposited at
597 NCBI with GenBank accession number CP062279.

598 **Competing interests**

599 The authors declare that they have no competing interests.

600

601

602 **Funding**

603 The study was supported by a project grant (BB/K00638X/1) and institute strategic grant
604 funding ISP2: BB/P013740/1 from the Biotechnology and Biological Sciences Research
605 Council (United Kingdom) to J.R.F.; grant MR/N02995X/1 from the Medical Research Council
606 (United Kingdom) to J.R.F; Wellcome Trust collaborative award 201531/Z/16/Z to J.R.F. and
607 J.R.P.

608 **Authors' contributions**

609 JRF and JRP lead the project. GY, BAW and EJR performed the genomics analyses. AFH
610 and MMN performed the *in vitro* analyses. SG advised on the transcriptional analyses. PH,
611 MAT-M and R de la F provided the isolates and their metadata. GY, AFH, JRP and JRF wrote
612 the first draft of the manuscript. All authors read and approved the final manuscript.

613 **Acknowledgments**

614 We wish to thank Dr. Prerna Vohra and Dr. Cosmin Chintoan-Uta (The Roslin Institute,
615 University of Edinburgh) for their help culturing the *S. aureus* subsp. *anaerobius* isolates under
616 microaerophilic conditions.

617 ## References

618 1. Woolhouse ME, Taylor LH, Haydon DT. Population biology of multihost pathogens. Science. 2001;
619 292(5519):1109-1112.
620 2. Moran NA, Plague GR. Genomic changes following host restriction in bacteria. Curr Opin Genet
621 Dev. 2004; 14(6):627-633.
622 3. Baumler A, Fang FC. Host specificity of bacterial pathogens. Cold Spring Harb Perspect Med. 2013;
623 3(12):a010041.
624 4. Bobay LM, Ochman H. The evolution of bacterial genome architecture. Front Genet. 2017; 872.
625 5. Weinert LA, Welch JJ. Why might bacterial pathogens have small genomes? Trends Ecol Evol.
626 2017; 32(12):936-947.
627 6. Fitzgerald JR, Holden MT. Genomics of natural populations of *Staphylococcus aureus*. Annu Rev
628 Microbiol. 2016; 70459-478.
629 7. Guinane CM, Ben Zakour NL, Tormo-Mas MA, Weinert LA, Lowder BV, Cartwright RA, et al.
630 Evolutionary genomics of *Staphylococcus aureus* reveals insights into the origin and molecular basis
631 of ruminant host adaptation. Genome Biol Evol. 2010; 2454-466.

632 8. Viana D, Blanco J, Tormo-Mas MA, Selva L, Guinane CM, Baselga R, et al. Adaptation of
633 *Staphylococcus aureus* to ruminant and equine hosts involves SaPI-carried variants of von
634 Willebrand factor-binding protein. Mol Microbiol. 2010; 77(6):1583-1594.
635 9. Viana D, Comos M, McAdam PR, Ward MJ, Selva L, Guinane CM, et al. A single natural nucleotide
636 mutation alters bacterial pathogen host tropism. Nat Genet. 2015; 47(4):361-366.
637 10. Wilson GJ, Seo KS, Cartwright RA, Connelley T, Chuang-Smith ON, Merriman JA, et al. A novel
638 core genome-encoded superantigen contributes to lethality of community-associated MRSA
639 necrotizing pneumonia. PLoS Pathog. 2011; 7(10):e1002271.
640 11. de la Fuente R, Suarez G. Respiratory deficient *Staphylococcus aureus* as the aetiological agent of
641 "abscess disease". Zentralbl Veterinarmed B. 1985; 32(6):397-406.
642 12. De la Fuente R, Suarez G, Schleifer KH. Staphylococcus aureus subsp. anaerobius subsp. nov.,
643 the causal agent of abscess disease of sheep. Int J Syst Bacteriol. 1985; 35(1):99-102.
644 13. de la Fuente R, Ballesteros C, Bautista V, Medina A, Orden JA, Domínguez-Bernal G, et al.
645 *Staphylococcus aureus* subsp. *anaerobius* isolates from different countries are clonal in nature. Vet
646 Microbiol. 2011; 150(1-2):198-202.
647 14. Elbir H, Robert C, Nguyen TT, Gimenez G, El Sanousi SM, Flock JI, et al. *Staphylococcus aureus*
648 subsp. *anaerobius* strain ST1464 genome sequence. Stand Genomic Sci. 2013; 9(1):1-13.
649 15. Elbir H, Feil EJ, Drancourt M, Roux V, El Sanousi SM, Eshag M, et al. Ovine clone ST1464: a
650 predominant genotype of *Staphylococcus aureus* subsp. *anaerobius* isolated from sheep in Sudan.
651 J Infect Dev Ctries. 2010; 4(4):235-238.
652 16. Gibson B, Wilson DJ, Feil E, Eyre-Walker A. The distribution of bacterial doubling times in the wild.
653 Proc Biol Sci. 2018; 285(1880).
654 17. Richardson EJ, Bacigalupe R, Harrison EM, Weinert LA, Lycett S, Vrieling M, et al. Gene exchange
655 drives the ecological success of a multi-host bacterial pathogen. Nat Ecol Evol. 2018.
656 18. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol.
657 2006; 23(2):254-267.
658 19. Varani AM, Siguier P, Gourbeyre E, Charneau V, Chandler M. ISsaga is an ensemble of web-based
659 methods for high throughput identification and semi-automatic annotation of insertion sequences in
660 prokaryotic genomes. Genome Biol. 2011; 12(3):R30.
661 20. Charpentier E, Anton AI, Barry P, Alfonso B, Fang Y, Novick RP. Novel cassette-based shuttle vector
662 system for gram-positive bacteria. Appl Environ Microbiol. 2004; 70(10):6076-6085.
663 21. Luong TT, Lee CY. The arl locus positively regulates *Staphylococcus aureus* type 5 capsule via an
664 mgrA-dependent pathway. Microbiology (Reading). 2006; 152(Pt 10):3123-3131.
665 22. Crosby HA, Schlievert PM, Merriman JA, King JM, Salgado-Pabon W, Horswill AR. The
666 *Staphylococcus aureu*s Global Regulator MgrA Modulates Clumping and Virulence by Controlling
667 Surface Protein Expression. PLoS Pathog. 2016; 12(5):e1005604.
668 23. Crosby HA, Tiwari N, Kwiecinski JM, Xu Z, Dykstra A, Jenul C, et al. The *Staphylococcus aureus*
669 ArlRS two-component system regulates virulence factor expression through MgrA. Mol Microbiol.
670 2020; 113(1):103-122.

671 24. Pagels M, Fuchs S, Pane-Farre J, Kohler C, Menschner L, Hecker M, et al. Redox sensing by a
672 Rex-family repressor is involved in the regulation of anaerobic gene expression in *Staphylococcus*
673 *aureus*. Mol Microbiol. 2010; 76(5):1142-1161.
674 25. Lasa I, Toledo-Arana A, Dobin A, Villanueva M, de los Mozos IR, Vergara-Irigaray M, et al. Genome-
675 wide antisense transcription drives mRNA processing in bacteria. Proc Natl Acad Sci U S A. 2011;
676 108(50):20172-20177.
677 26. Kajimura J, Fujiwara T, Yamada S, Suzawa Y, Nishida T, Oyamada Y, et al. Identification and
678 molecular characterization of an N-acetylmuramyl-L-alanine amidase Sle1 involved in cell
679 separation of *Staphylococcus aureus*. Mol Microbiol. 2005; 58(4):1087-1101.
680 27. Thalsø-Madsen I, Torrubia FR, Xu L, Petersen A, Jensen C, Frees D. The Sle1 cell wall amidase is
681 essential for β-lactam resistance in community-acquired methicillin-resistant *Staphylococcus aureus*
682 USA300. Antimicrob Agents Chemother. 2019; 64(1).
683 28. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and
684 diversity. FEMS Microbiol Rev. 2014; 38(5):865-891.
685 29. Toft C, Andersson SG. Evolutionary microbial genomics: insights into bacterial host adaptation. Nat
686 Rev Genet. 2010; 11(7):465-475.
687 30. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, et al. Massive gene decay
688 in the leprosy bacillus. Nature. 2001; 409(6823):1007-1011.
689 31. Feng Y, Chen Z, Liu SL. Gene decay in *Shigella* as an incipient stage of host-adaptation. PLoS One.
690 2011; 6(11):e27754.
691 32. Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, et al. Patterns of genome
692 evolution that have accompanied host adaptation in *Salmonella*. Proc Natl Acad Sci USA. 2015;
693 112(3):863-868.
694 33. Song H, Hwang J, Yi H, Ulrich RL, Yu Y, Nierman WC, et al. The early stage of bacterial genome-
695 reductive evolution in the host. PLoS Pathog. 2010; 6(5):e1000922.
696 34. Periwal V, Scaria V. Insights into structural variations and genome rearrangements in prokaryotic
697 genomes. Bioinformatics. 2015; 31(1):1-9.
698 35. Block DH, Hussein R, Liang LW, Lim HN. Regulatory consequences of gene translocation in
699 bacteria. Nucleic Acids Res. 2012; 40(18):8979-8992.
700 36. Prentki P, Teter B, Chandler M, Galas DJ. Functional promoters created by the insertion of
701 transposable element IS1. J Mol Biol. 1986; 191(3):383-393.
702 37. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, et al. Comparative analysis
703 of the genome sequences of *Bordetella pertussis, Bordetella parapertussis* and *Bordetella*
704 *bronchiseptica*. Nat Genet. 2003; 35(1):32-40.
705 38. Amman F, D'Halluin A, Antoine R, Huot L, Bibova I, Keidel K, et al. Primary transcriptome analysis
706 reveals importance of IS elements for the shaping of the transcriptional landscape of Bordetella
707 pertussis. RNA Biol. 2018; 15(7):967-975.
708 39. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.
709 Bioinformatics. 2014; 30(15):2114-2120.

710 40. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome
711 assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012; 19(5):455-
712 477.
713 41. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs.
714 Genome Res. 2008; 18(5):821-829.
715 42. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014; 30(14):2068-2069.
716 43. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale
717 prokaryote pan genome analysis. Bioinformatics. 2015; 31(22):3691-3693.
718 44. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of
719 the PHAST phage search tool. Nucleic Acids Res. 2016; 44(W1):W16-21.
720 45. Bertelli C, Laird MR, Williams KP, Simon Fraser University Research Computing G, Lau BY, Hoad
721 G, et al. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic
722 Acids Res. 2017.
723 46. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of
724 acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012; 67(11):2640-2644.
725 47. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a
726 hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic
727 and viral sequences. Nucleic Acids Res. 2016; 44(D1):D286-293.
728 48. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes
729 among gene clusters. OMICS. 2012; 16(5):284-287.
730 49. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, et al. IMG: the Integrated
731 Microbial Genomes database and comparative analysis system. Nucleic Acids Res. 2012;
732 40(Database issue):D115-122.
733 50. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational
734 challenges and solutions. Nat Rev Genet. 2011; 13(1):36-46.
735 51. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
736 Bioinformatics. 2009; 25(14):1754-1760.
737 52. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic
738 analysis of large samples of recombinant bacterial whole genome sequences using Gubbins.
739 Nucleic Acids Res. 2015; 43(3):e15.
740 53. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic
741 algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32(1):268-274.
742 54. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of
743 heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016; 2(1):vew007.
744 55. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the
745 BEAST 1.7. Mol Biol Evol. 2012; 29(8):1969-1973.
746 56. Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, et al. CLIMB (the Cloud
747 Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology
748 community. Microb Genom. 2016; 2(9):e000086.

57. Ausubel F, Brent R, Kingston R, Moore D, Seidman J, Smith J, et al: Current Protocols in Molecular Biology. New York, NY: John Wiley & Sons; 1990.
58. Sambrook J, Fritsch EF, Maniatis T: Molecular cloning: a laboratory manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1989.
59. Ji G, Beavis R, Novick RP. Bacterial interference caused by autoinducing peptide variants. Science. 1997; 276(5321):2027-2030.

List of Figures and Tables

Main Figures:

1. Distribution of pseudogenes, insertion sequences and mobile genetic elements in the *Staphylococcus aureus* subsp. *anaerobius* isolate MVF7.
2. Bayesian maximum clade credibility tree of the *Staphylococcus aureus* subsp. *anaerobius*
3. The position of *Staphylococcus aureus* subsp. *anaerobius* within the *S. aureus* diversity.
4. Characterisation of the transcriptional terminators (TTs) present in the ISs.
5. Presence of IS impacts on the expression levels of downstream genes and genes product.
6. Map of the genomic translocations across the *Staphylococcus aureus* subsp. *anaerobius* isolate MVF7 chromosome.

Supplementary Figures (in Additional File 1)**:**

S1. Phylogenetic tree of the integrase gene from representative examples of *S. aureus* prophages.
S2. Phylogenetic network of representative examples of *S. aureus* Pathogenicity Islands.
S3. Expression of SaaPIMVF7-encoded *vwb.*
S4. Schematic representation of IS loci selected for analysis
S5. Presence of the IS does not affect expression of the downstream gene product through active transcription.
S6. Pairwise genome alignment of *S. aureus* subsp. *anaerobius versus S. aureus* subsp. *aureus*.

Supplementary Tables (in Additional File 2):

S1. List of pseudogenes found in the S. aureus subsp. anaerobius strain MVF7.
S2. GO terms and KEGG pathways enriched in pseudogenes vs intact genes.
S3. PCR results to test for rearrangements.
S4. Strains used in this study.
S5. Plasmids used in this study.
S6. Oligonucleotides used in this study.

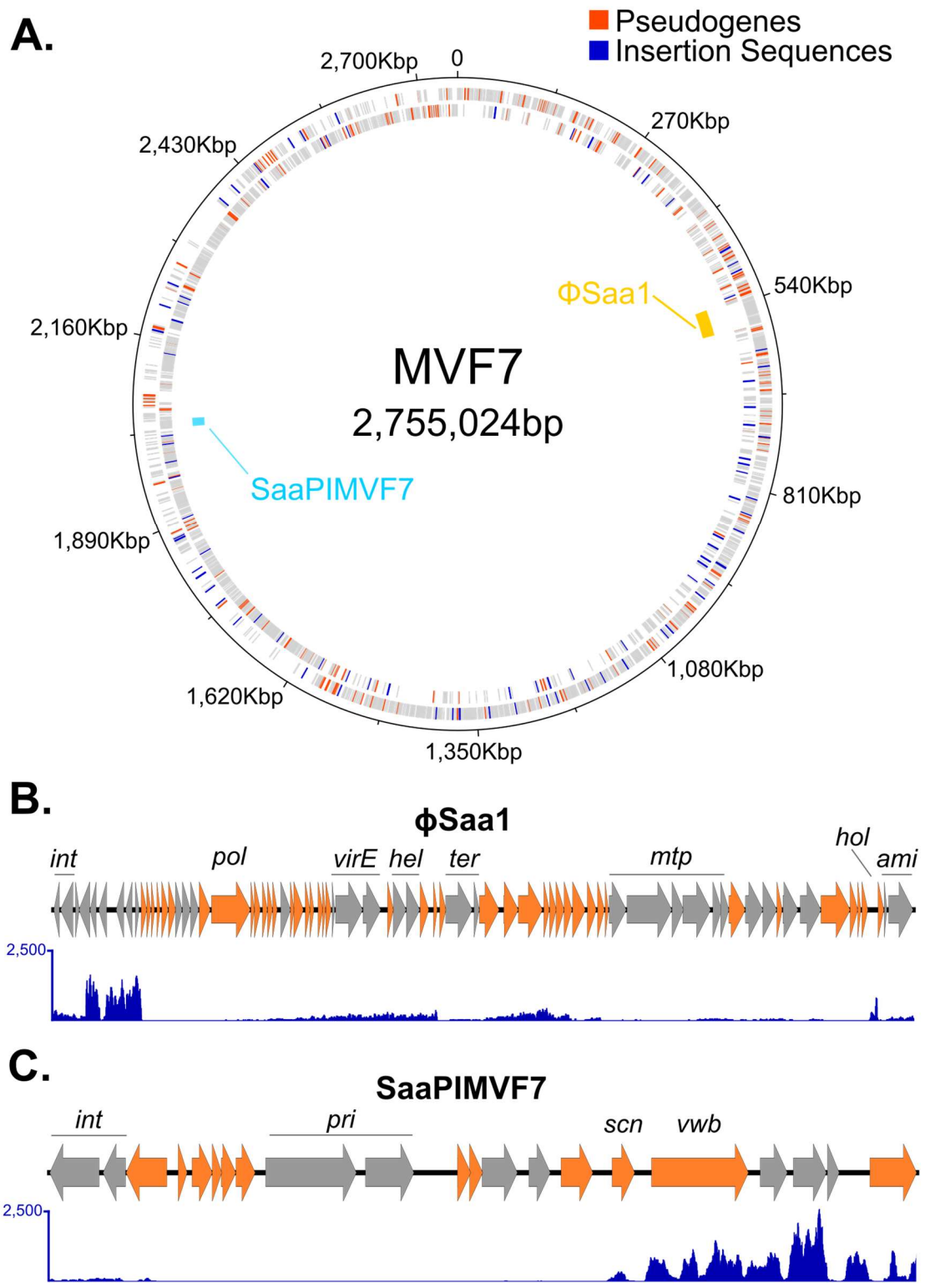

Figure 1. Distribution of pseudogenes, insertion sequences and mobile genetic elements in the *Staphylococcus aureus* subsp. *anaerobius* isolate MVF7. (A): Circular map of the chromosome. Rings show, from outside to inside: annotated genes in the positive strand, those in the negative strand (with the 205 pseudogenes and 87 insertion sequences shown in red and blue, respectively), and the mobile genetic elements found across all isolates (gold: prophage ΦSaa1; cyan: pathogenicity island SaaPIMVF7). (B) and (C): genes found in ΦSaa1 and SaaPIMVF7, respectively. Genes in grey are pseudogenes and genes in orange are intact (according to homology against their most similar phage (Φ2958PVL) and SaPI (SaPIov2) relatives. The histograms in blue represent the genes' expression levels (inferred from RNA-seq read coverage).

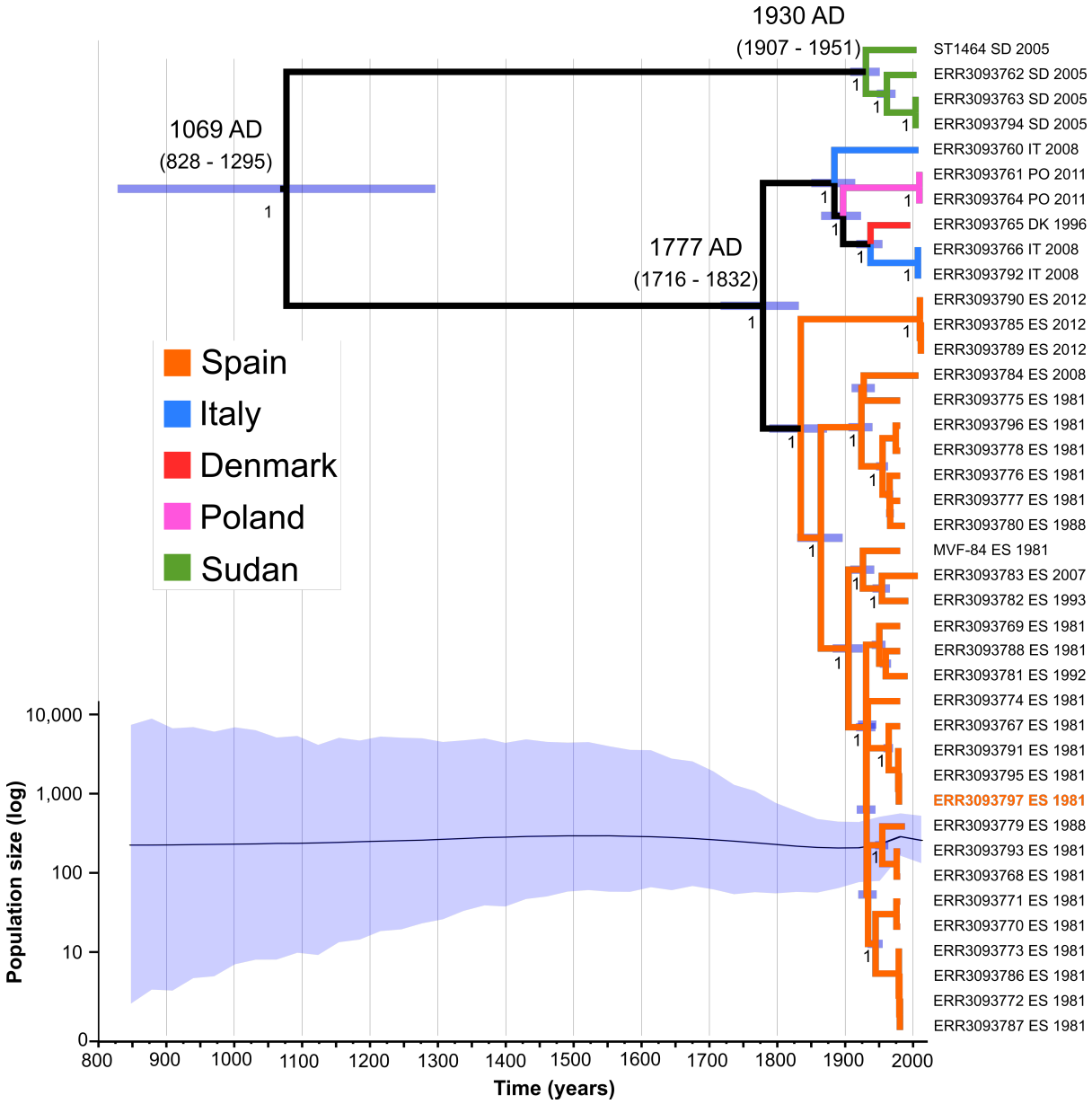

Figure 2. Bayesian maximum clade credibility tree of the *Staphylococcus aureus* subsp. *anaerobius* isolates. The x-axis is expressed in calendar years. Branch colours denote sampling country (see key). Purple bars at each node show its most recent common ancestor (MRCA) confidence interval. Numbers in nodes indicate the posterior probability. The estimated dates of the MRCAs of the main lineages are indicated. The isolate that was whole genome sequenced using PacBio technology (MVF7) is highlighted in orange. The bottom plot represents the changes in the effective population size over time, with the shadowed area representing the 95% credible interval, following the same timescale as the tree.

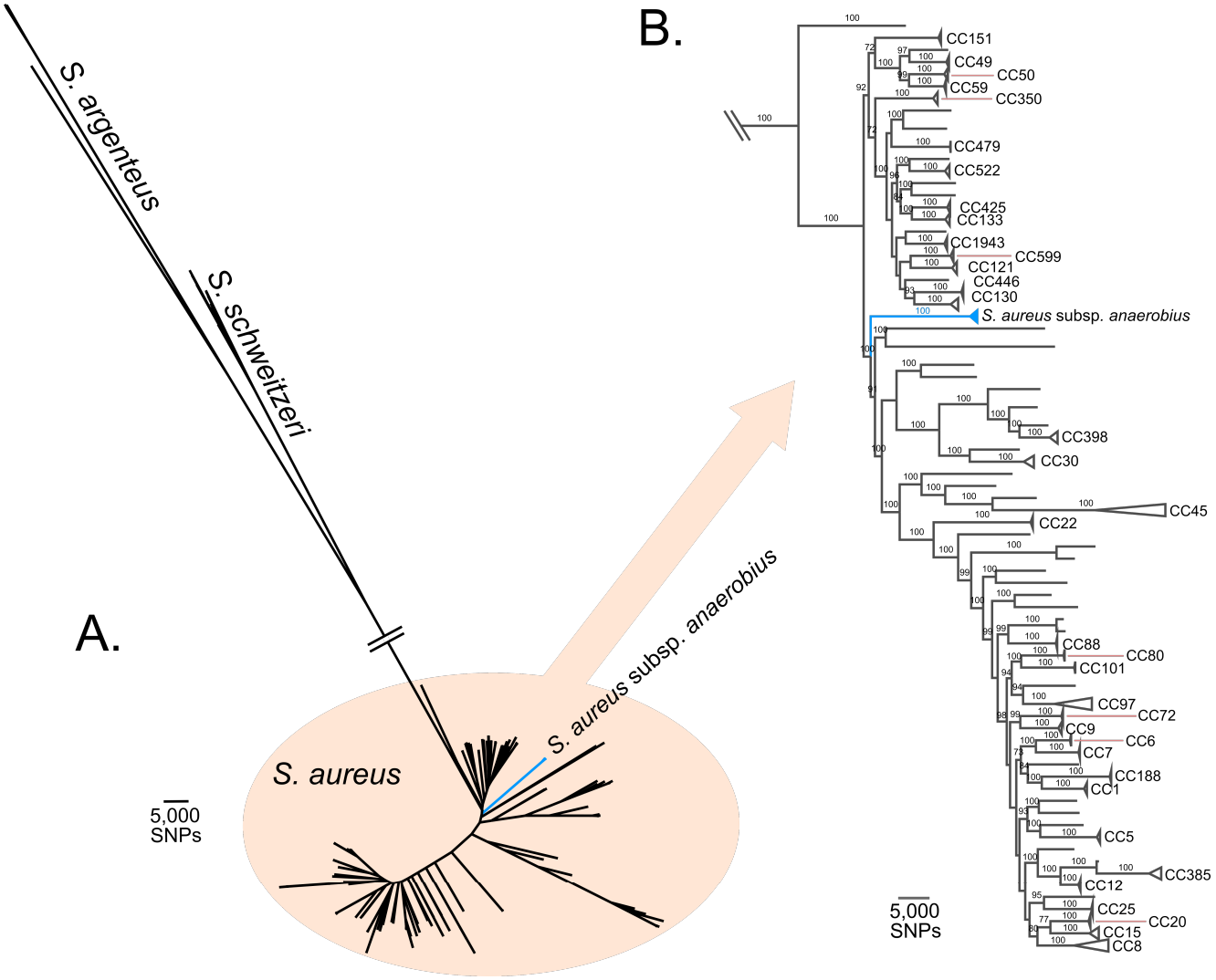

Figure 3. The position of *Staphylococcus aureus* subsp. *anaerobius* within the *S. aureus* diversity. Maximum Likelihood tree constructed from a SNP alignment of the studied *S. aureus* subsp. *anaerobius* sequences (clades in blue) and 787 sequences of different *S. aureus* subsp. *aureus* (in black). (A): unrooted tree showing the divergence of 17 sequences of other *Staphylococcus* (*S. schweitzeri* and *S. argenteus*) whereas *S. aureus* subsp. *aureus* is embedded in the *S. aureus* diversity. (B): a subtree showing the position of the *S. aureus* subsp. *anaerobius* clade (collapsed in blue) with respect to the other *S. aureus* subsp. *aureus* clonal complexes (CC).

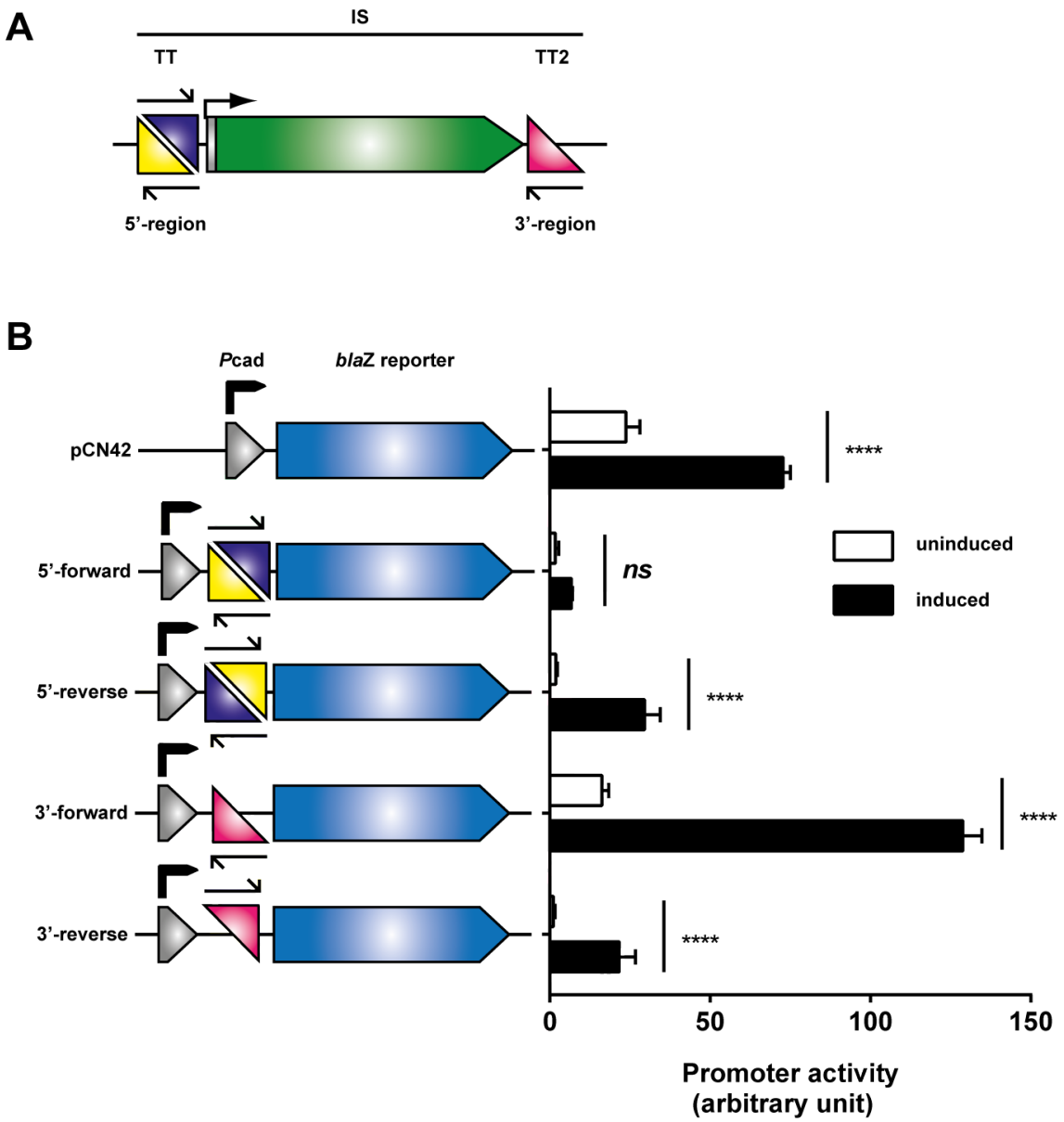

Figure 4. Characterisation of the transcriptional terminators (TTs) present in the ISs. (A) Schematic representation of IS-associated TTs. (B) Schematic representation of and functional assessment of the bi-directional TT1 localised in the 5' region of the IS. The bi-directional TT1 is composed of two sub-TTs that share the same sequence but are inverted repeats, thus generating a master bi-directional TT1. The expression of the *bla*Z reporter in plasmid pCN42 is controlled by a cadmium-inducible promoter. The identified transcriptional terminators were cloned between the promoter and the *bla*Z reporter gene. β-lactamase expression was monitored either in uninduced (black bars) or Cd-induced (open bars) cultures 180 min after induction. Data show the means of three independent biological replicates and error bars show the standard deviation from the mean. Statistical analysis was performed using two-way ANOVA followed by Dunnett's post-test. Multiplicity adjusted p-values are shown for uninduced to induced comparison and all terminator containing reporters showed significant differences when compared to their respective empty plasmid control. **** $p<0.0001$, ns not significant.

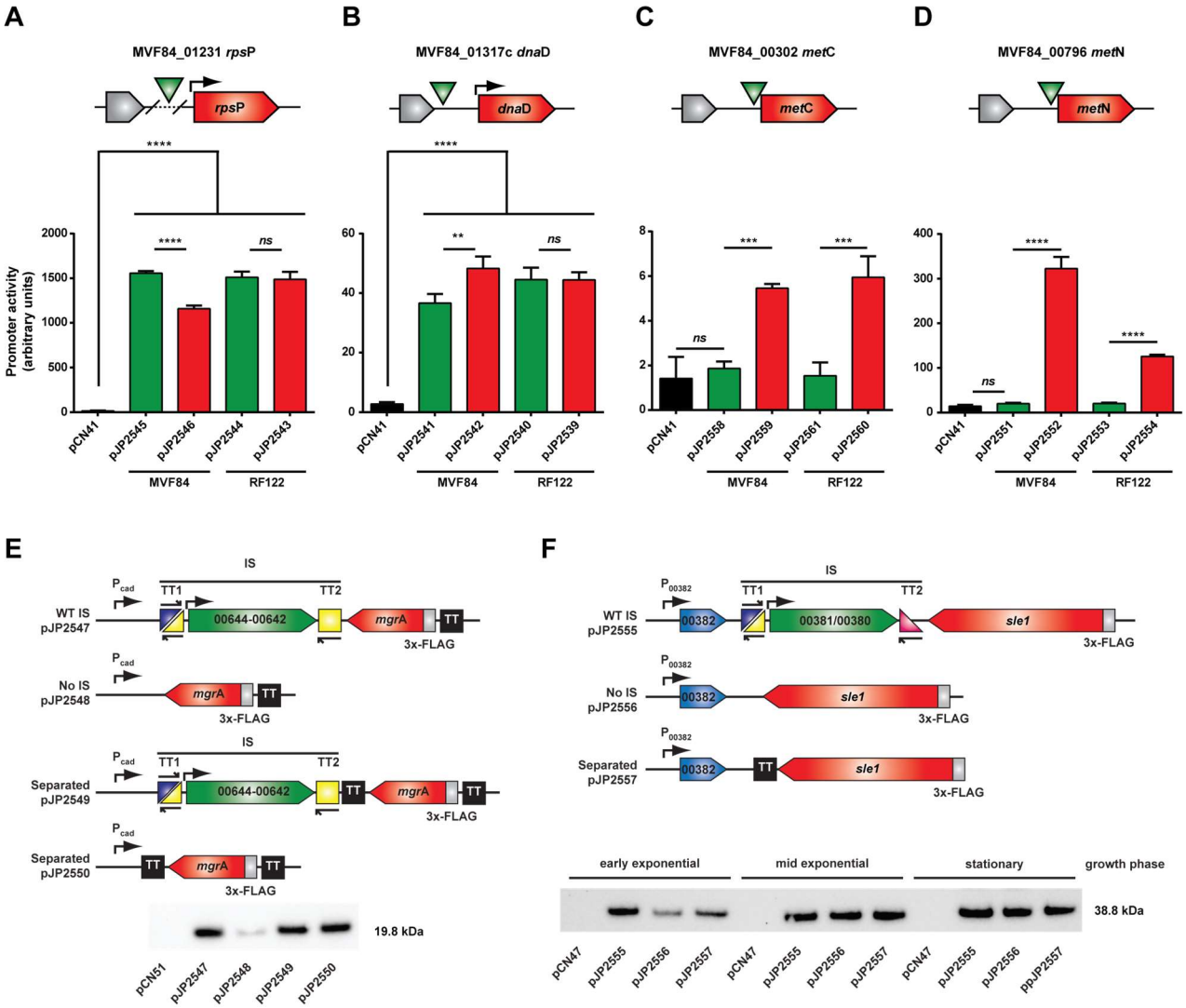

Figure 5. Presence of IS impacts on the expression levels of downstream genes and genes product. (A-D) Reporter constructs for presence of the IS upstream and in sense orientation of the target gene were designed in plasmid pCN41 containing a β-lactamase reporter gene. The position of the IS relative to the target gene's start codon is indicated by green triangle. All reporters were constructed either by removing the IS gene from strain *S. aureus* subsp. *anaerobius* strain MVF84 or by introducing the IS into *S. aureus* subsp. *aureus* strain RF122 to confirm the causative role of the IS on any changes to gene expression. Note that in (A) the removal of the IS results in a 207 bp deletion to the promoter region not present in strain RF122. Reporter plasmids were introduced into strain RN4220 and samples taken and processed as described in Materials and Methods. Data show the mean of three independent biological replicates and error bars represent the standard deviation from the mean. Statistical analysis was performed using One-Way ANOVA followed by Tukey's multiple comparisons test. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001, *ns* not significant. (E&F) Western blot analysis of (E) *mgr*A and (F) *sle1* expression constructs to address whether presence of the IS impacts on MgrA and Sle1 expression levels by interrupting expression of an antisense transcript. (E) The *mgr*A gene encoding a 3xFLAG tag was cloned in antisense orientation to a cadmium-inducible promoter into plasmid pCN51 to mimic the expression of an antisense transcript to *mgr*A. Constructs either contained or lacked the *S. aureus* subsp. *anaerobius* IS, or contained a transcriptional terminator (TT) separating the IS transcript or expression from the plasmid-encoded cadmium-inducible promoter from the *mgr*A transcript. Expression of the cadmium-inducible promoter was induced by the addition of 5 µM $CdCl_2$. (F) To assess the impact of IS in the context of a natural antisense transcript, 3xFLAG-encoding expression constructs of *sle1* including the gene downstream of *sle1* in the ancestral genome were cloned into the plasmid pCN47. Samples were taken during different growth phases as indicated. Plasmids were introduced into the *S. aureus* strain RN4220 Δ*spa* and samples for expression analysis taken and processed as described in Materials and Methods. Western blots shown are representative of results obtained in at least two independent biological replicates. TT transcriptional terminator (E-F).

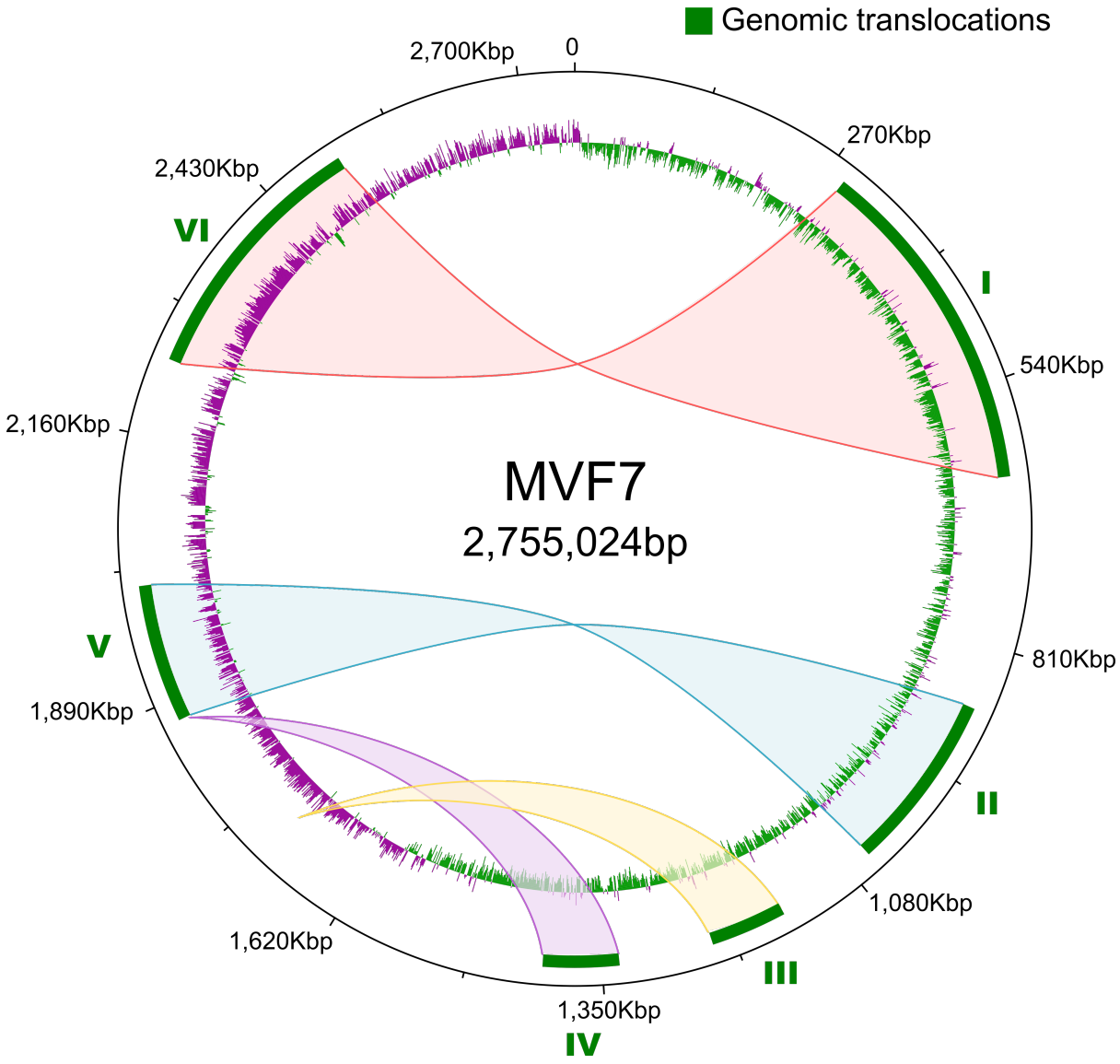

Figure 6. Map of the genomic translocations across the *Staphylococcus aureus* subsp. *anaerobius* isolate MVF7 chromosome. The green blocks represent the 6 large translocations detected in MVF7 when compared to the *Staphylococcus aureus* subsp. *aureus* RF122 strain (see also **Fig. S6** in the Additional File 1). The shaded areas indicate the original location of the translocated portions in the putative ancestral genome. The inner ring shows the GC skew (green for positive skew, purple for negative), which is unaffected by the translocations.