═══REVIEW COMMONS MANUSCRIPT═══

IMPORTANT:

- Manuscripts submitted to Review Commons are peer reviewed in a journal-agnostic way.
- Upon transfer of the peer reviewed preprint to a journal, the referee reports will be available in full to the handling editor.
- The identity of the referees will NOT be communicated to the authors unless the reviewers choose to sign their report.
- The identity of the referee will be confidentially disclosed to any affiliate journals to which the manuscript is transferred.

GUIDELINES:

- For reviewers: https://www.reviewcommons.org/reviewers
- For authors: https://www.reviewcommons.org/authors

CONTACT:

The Review Commons office can be contacted directly at: office@reviewcommons.org

Identifying Genetic Variations in *emm*89 *Streptococcus pyogenes* Linked to Severe Invasive Infections

Masayuki Ono[1,2], Masaya Yamaguchi[1,2,3,4,19*], Daisuke Motooka[3,5,6], Yujiro Hirose[1], Kotaro Higashi[1], Tomoko Sumitomo[1,7], Tohru Miyoshi-Akiyama[8], Rumi Okuno[9], Takahiro Yamaguchi[10], Ryuji Kawahara[10], Hitoshi Otsuka[11], Noriko Nakanishi[12], Yu Kazawa[13], Chikara Nakagawa[14], Ryo Yamaguchi[15], Hiroo Sakai[16], Yuko Matsumoto[17], Tadayoshi Ikebe[18], and Shigetada Kawabata[1,4]

1. Department of Microbiology, Osaka University Graduate School of Dentistry, Osaka, Japan
2. Bioinformatics Research Unit, Osaka University Graduate School of Dentistry, Osaka, Japan
3. Bioinformatics Center, Research Institute for Microbial Diseases, Osaka University, Osaka, Japan
4. Center for Infectious Diseases Education and Research, Osaka University, Osaka, Japan
5. NGS Core Facility, Research Institute for Microbial Diseases, Osaka University,

Osaka, Japan

6. Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives (OTRI), Osaka University, Osaka, Japan

7. Department of Oral Microbiology, Graduate School of Biomedical Sciences, Tokushima University, Tokushima, Japan

8. Pathogenic Microbe Laboratory, Department of Infectious Diseases, National Center for Global Health and Medicine, Tokyo, Japan

9. Department of Microbiology, Tokyo Metropolitan Institute of Public Health, Tokyo, Japan

10. Department of Bacteriology, Osaka Institute of Public Health, Osaka, Japan

11. Department of Public Health Sciences, Yamaguchi Prefectural Institute Public Health and Environment, Yamaguchi, Japan

12. Department of Infectious Diseases, Kobe Institute for Health, Hyogo, Japan

13. Department of Microbiology, Fukushima Prefectural Institute for Public Health, Fukushima, Japan

14. Division of Microbiology, Kyoto City Institute of Health and Environmental Sciences, Kyoto, Japan

15. Sapporo Public Health Office, Hokkaido, Japan

[16] Niigata City Institute of Public Health and the Environment, Niigata, Japan

[17] Microbiological Testing and Research Division, Yokohama City Institute of Public Health, Kanagawa, Japan

[18] Department of Bacteriology I, National Institute of Infectious Diseases, Tokyo, Japan

[19] Lead contact

* Correspondence: yamaguchi.masaya.dent@osaka-u.ac.jp

Summary

Streptococcus pyogenes causes mild human infections as well as life-threatening invasive diseases. Since the mutations known to enhance virulence to date account for only half of the severe invasive infections, additional mechanisms/mutations need to be identified. Here, we conducted a genome-wide association study of *emm*89 *S. pyogenes* strains to comprehensively identify pathology-related bacterial genetic factors (SNPs, indels, genes, or k-mers). Japanese (*n*=311) and global (*n*=666) cohort studies of strains isolated from invasive or non-invasive infections revealed 17 and 1,075 SNPs/indels and 2 and 169 genes, respectively, that displayed associations with invasiveness. We validated one of them, a non-invasiveness-related point mutation, *fhuB* T218C, by structure predictions and introducing it into a severe invasive strain and confirmed that the mutant showed slower growth in human blood. Thus, we report novel mechanisms that convert *emm*89 *S. pyogenes* to an invasive phenotype and a platform for establishing novel treatments and prevention strategies.

Keywords: *Streptococcus pyogenes*, *emm*89, severe invasive infections, genome-wide association study, pan-genome analysis

62 **Introduction**

63 *Streptococcus pyogenes* is a human-restricted gram-positive pathogen associated with a
64 wide spectrum of diseases. While *S. pyogenes* often causes non-invasive diseases,
65 including pharyngitis and impetigo in children, it is also known as a "flesh-eating
66 bacterium" owing to its involvement in life-threatening invasive diseases, such as
67 necrotizing fasciitis and streptococcal toxic shock syndrome (STSS)[1,2]. In 2005, more
68 than 0.6 million people were estimated to have invasive *S. pyogenes* infections[3], and the
69 reported incidence of invasive *S. pyogenes* infections continues to increase globally[1]. In
70 cases of severe infection, rapid bacterial growth and profound metabolic acidosis
71 necessitate urgent surgical inspection and extended debridement with empiric
72 antibacterial chemotherapy[4,5]. However, even with proper treatment, the mortality rate of
73 patients with *S. pyogenes* infections remains 23–81%[6]. Moreover, although several
74 protective vaccine candidates against *S. pyogenes* exist, no safe and effective commercial
75 vaccine has yet been licensed for human use[2,7].

76 *S. pyogenes* has been classified into at least 240 *emm* types based on the *emm* gene
77 hypervariable region sequence[8]. Since the mid to late 2000s, *emm*89 strains have been
78 increasingly isolated from samples obtained from patients with invasive diseases,
79 becoming one of the most frequently identified lineages in developed countries[9,10]. For

80 example, in the United Kingdom, no more than ten cases of invasive diseases caused by
81 *emm*89 *S. pyogenes* were reported annually in the 1990s. In the 2010s, however,
82 approximately 150 cases were reported out of the 1,000–1,500 invasive cases caused by
83 all *emm* types of *S. pyogenes* in each year[11]. Similarly, in Japan, in the 2010s, there was
84 an increase in the number of patients with *emm*89 *S. pyogenes*-induced severe invasive
85 infections. In 2018, *emm*89 strains were isolated from 36% of patients diagnosed with *S.*
86 *pyogenes*-induced severe invasive infections, following *emm*1 strains, which were
87 isolated from 57% of patients[9].

88 *S. pyogenes emm*89 strains have been genetically sub-clustered into three clades
89 according to the *nga* promoter region patterns and the presence/absence of the *hasABC*
90 locus, which is responsible for hyaluronan capsule synthesis. Clade 3 is distinct from
91 clades 1 and 2 in terms of two features: overexpression of virulence factors NAD
92 glycohydrolase (NADase) and streptolysin O (SLO), owing to mutations in the promoter
93 region of the *nga-ifs-slo* operon, and the lack of a hyaluronan capsule[11-13]. Although clade
94 3 strains have frequently been isolated from invasive diseases, their numbers from non-
95 invasive infections have also increased. We previously reported no difference in the
96 isolation frequencies of clade 3 strains between invasive and non-invasive diseases, at
97 least in Japan, and concluded that the mutations in clade 3 are not responsible for the gain

98 of invasiveness[14]. Therefore, there must be other genetic features within the *emm*89

99 strains that determine their phenotypes.

100 During infections caused by *emm*1 *S. pyogenes*, mutations in the two-component

101 system, CovR/S, promote high virulence[15]. These mutations cause the upregulation of

102 DNase, hyaluronan capsule, IL-8 protease, C5a peptidase, streptokinase, NADase, SLO,

103 and superantigen SpeA, as well as the downregulation of cysteine protease SpeB and

104 streptolysin S[16,17]. The resulting mutants can prevent neutrophil death and subsequently

105 promote tissue destruction and systemic infections. Epidemiologically, Ikebe *et al.*

106 reported that nonsense mutations in *covR* and/or *covS* are present in 46.3% of *S. pyogenes*

107 strains isolated from severe invasive infections in Japan, but only 1.69% of isolates from

108 non-severe ones[18]. Moreover, these studies indicated that the *covR*/*S* mutation is not

109 responsible for all invasive clinical strains. Thus, we hypothesize that other mechanisms

110 are involved in the development of invasiveness.

111 In the present study, we aimed to explore novel hypervirulent mechanisms of *S.*

112 *pyogenes* by performing a genome-wide association study (GWAS) on S. *pyogenes*. Our

113 GWAS focused on the *emm*89 lineage of *S. pyogenes* to detect lineage-specific factors

114 and minimize false positives due to lineage differentiation. For a comprehensive analysis,

115 we constructed the core genome and pan-genome of *emm*89 *S. pyogenes* strains and

116 evaluated the effect of single-nucleotide polymorphisms (SNPs) on core gene alignment
117 and accessory clusters of orthologous genes (COGs). In addition, we performed a k-mers-
118 based GWAS to detect SNPs in the intergenic regions and multiple mutations. We
119 collected and sequenced *emm*89 clinical strains isolated in Japan during 2016–2021, in
120 addition to public *emm*89 genome sequences. Using these sequences, we investigated the
121 bacterial factors associated with severe invasive infections in Japan and globally, using
122 GWAS. Based on the bacterial protein structural predictions, we then selected candidates
123 with high potential relevance to the phenotype. Finally, we introduced an SNP related to
124 non-invasiveness into a clinical strain isolated from a severe invasive infection and
125 examined the alteration of the bacterial phenotype through an *ex vivo* infection assay.
126

127 ## Results

128 **Collection of *emm*89 *S. pyogenes* clinical isolates in Japan and construction of**
129 **cohorts**

130 We collected clinical *S. pyogenes* strains isolated between 2016 and 2021 from patients
131 with non-invasive and severe invasive infections in Japan. The Ministry of Health, Labour,
132 and Welfare of Japan has defined the clinical criteria of severe invasive β-hemolytic
133 streptococcal infections as STSS, based on the STSS 2010 Case Definition of the Centers

134 for Disease Control and Prevention in the US, with minor modifications, including the
135 addition of encompassing symptoms in the central nervous system (Table S1)[19,20].
136 For the *emm*89 clinical isolates, we collected T serotype TB3264 and untypable
137 strains, in addition to *emm* genotype-identified strains. T-typing is a serologically-based
138 approach that is often used as an alternative or supplement to *emm* typing. T-antigens are
139 trypsin-resistant surface antigens exhibiting extensive antigenic diversity[21]. Isolates of a
140 given *emm* type frequently share the same T serotype pattern[22,23]. The T serotype TB3264
141 corresponds to the genotype *emm*89 or *emm*94[21,24]. A total of 207 clinical isolates were
142 collected with the cooperation of the National Institute of Infectious Diseases and ten
143 public health institutes nationwide (Tables S2 and S3). We performed draft genome
144 sequencing of the strains and identified their *emm* types. In total, 150 of these were
145 determined as *emm*89, followed by 24 and 19 strains as *emm*4 and *emm*12, respectively.
146 To focus on the pathogenic mechanisms underlying severe invasive infections in the
147 *emm*89 cohort, we used 150 *emm*89 strains for subsequent analyses (Figure 1A, Tables
148 S2, and S3). We previously determined the draft genome sequences of 161 *emm*89 strains
149 isolated in Japan between 2011 and 2019 and determined their phenotypes using the same
150 criteria (Table S2)[14]. We combined these two sets and finally considered a total of 311
151 *emm*89 strains, including 135 severe invasive and 176 non-invasive isolates, as the

152 Japanese cohort.

153 We also collected public genome sequences of *emm*89 *S. pyogenes* strains isolated
154 from nine countries to further characterize the genetic properties of the Japanese cohort
155 (Table S2)[25-27]. In this study, the phenotypes of these strains were considered invasive if
156 the diagnoses included severe infections, STSS, invasive infections, necrotizing fasciitis,
157 bacteremia, or sepsis, and isolation sites were described as normally sterile sites, such as
158 the blood, brain, kidney, muscular tissue, or brain. Consequently, we identified 666 strains
159 in the global cohort, including 420 isolates from invasive cases and 246 from non-
160 invasive ones (Table S2).

161

162 **Pan-genome and phylogenetic analyses reveal both shared and distinct features in**
163 **the Japanese and global cohorts**

164 To determine the core genes and gene distribution in both cohorts, we performed pan-
165 genome analyses. In the Japanese cohort, 1,417 core genes common to more than 99% of
166 all isolates were determined out of the 3,334 different genes detected within the 311
167 strains. In contrast, the global cohort was more diverse, with 4,743 different genes, of
168 which 1,327 were core genes (Figure 1B).

169 Next, we calculated the phylogenetic relationships based on the maximum

170 likelihood of the core gene sequences (Figure 1C and S2). The tree for the global cohort
171 branched into four clusters, with clusters A, B1, B2, and B3. Cluster B3 included 640
172 genetically similar strains isolated mainly from Europe, North America, and Japan,
173 whereas cluster A comprised 19 strains isolated from Oceanian countries, Kenya,
174 Lebanon, the US, and Japan (Figure 1C). The phylogenetic tree for the Japanese cohort
175 could also be clustered as in the case of the global cohort, with no significant difference
176 in the proportions of strains classified into each sub-cluster (chi-square test, p=0.13;
177 Figure S2 and Table S4). Thus, we concluded that the overall phylogenetic features of
178 *emm*89 strains were distributed similarly in Japan and other areas, especially Europe and
179 North America. Within cluster B3, we identified a non-invasive strain from Japan that had
180 no identical pattern to the reported *nga* promoter variations (Figure 1C)[28]. This pattern is
181 likely a subtype of clade 3 as it shares the haplotype $A_{-27}G_{-22}T_{-18}$, which is distinctive of
182 clade 3, but has a mutation in the –10 box (Figure S2)[28]. Thus, we named this novel *nga*
183 promoter variation type 3.4[28]. Multilocus sequence typing (MLST) analysis revealed that
184 in cluster B3, 522 strains (80.9%) were ST101, and 96 strains (14.8%) were ST646
185 (Figure 1C). Notably, ST646 was the second most dominant type and a Japan-specific
186 lineage. Moreover, they only differed in the 295th nucleotide in the *murI* locus, one of the
187 seven loci that determine MLST, suggesting that both lineages have a genetically close

188 relationship. Eight novel MLSTs were determined (ST1450, 1451, 1452, 1454, 1455,
189 1456, 1461, and 1463) and 15 novel MLST strains were detected in the Japanese cohort
190 (Figure S1). Taken together, using phylogenetic approaches, we found that most strains
191 from Japan and countries in Europe and North America share genetically close
192 relationships, with only one unique lineage in Japan, ST646.
193

194 **GWASes detect SNPs/indels associated with invasiveness that are both common and**
195 **specific to Japan and other countries**

196 To discover all types of genetic variants in whole genes within *emm*89 *S. pyogenes*
197 associated with (severe) invasiveness, we applied pan-genome analysis and performed
198 three types of independent GWASes targeting SNPs in core genes, the presence or
199 absence of all genes, and other variants located in intergenic regions spanning several
200 nucleotides.

201 We extracted SNPs and single-nucleotide indels from core gene alignments and
202 detected 24,627 and 47,060 SNPs/indels in the Japanese and global cohorts, respectively.
203 Subsequent GWASes identified SNPs/indels associated with severe invasiveness in a
204 Japanese cohort and invasiveness in a global cohort. To control for population bias, we
205 calculated pairwise distance matrices and selected seven and three dimensions for the

206 analyses of the Japanese and global cohorts, respectively (Figure S3A and S3B). For each
207 cohort, we performed a permutation test by conducting 1,000 iterations of calculations
208 with randomly permuted phenotypes, with the significance level set at the 5th percentile
209 of the 1,000 minimal *p*-values (p=5.75×10^{-4} and p=5.75×10^{-4} for the Japanese and global
210 cohorts, respectively). The GWAS of the Japanese cohort detected 17 SNPs/indels in 13
211 core genes (Figure 2A and Table S5). Of the 17 significant variants, there were 7 single-
212 nucleotide deletions (SNDs), 7 SNPs causing non-synonymous amino acid substitutions,
213 and 3 SNPs causing synonymous substitutions. The *covS* gene (also known as *csrS*),
214 encoding a sensor kinase of the two-component system CovR/S, contains four SNDs with
215 the lowest *p*-values (p=1.16×10^{-7} for the 39th, 40th, and 46th nucleotides, and p=1.15×10^{-}
216 ^{6} for the 125th nucleotide). These four deletions were associated with severe invasive
217 infections.

218 We also performed a GWAS for the global cohort and detected 1,075 SNPs/indels
219 significantly related to invasive infections among the 360 core genes (Figure 2B, 2C and
220 Table S6). Among the significant SNPs/indels, 725 caused synonymous substitutions and
221 319 caused non-synonymous substitutions or frameshift mutations. Moreover, 19 SNPs
222 induced nonsense mutations, whereas the effects of 12 SNPs/indels were unpredictable
223 because of a lack of reference sequences (Table S3). Notably, 96 SNPs/indels

224 accumulated in a single gene, *murJ*, which is involved in peptidoglycan biosynthesis,
225 whereas 53 and 51 SNPs/indels were detected in *murE* and *group_1008*, respectively
226 (Figure S3C). The SNP with the lowest *p*-value (p=1.35×10^{-14}) was *lacE,* which encodes
227 the EIICB component of the lactose-specific phosphotransferase system (Figure 2C). This
228 mutation was associated with an invasive phenotype and waws mainly observed in strains
229 isolated in the US. Compared with the significant 17 SNPs/indels in the Japanese cohort,
230 10 SNPs/indels were also detected in the global cohort, including 4 SNDs in *covS* and 1
231 SNP each in 6 loci (Figure 2C). Deletions at the *covS* locus were common among strains
232 from several countries, including Japan. In contrast, SNPs in six loci, *gatA, group_1102*,
233 *group_647*, *iscS_1*, *recU*, and *fhuB*, were present exclusively in Japan (Figure 2C and
234 Table S6). These results suggest that several bacterial mechanisms cause severe invasive
235 *S. pyogenes* infections, and some prevail worldwide, such as *covS* mutations, whereas
236 others are specific to Japan.
237

238 **GWAS on COGs reveals 2 and 109 genes associated with severe invasiveness in the**
239 **Japanese cohort and invasiveness in the global cohort, respectively**

240 Next, we examined the associations of accessory COGs with severe invasiveness and
241 global invasiveness in Japanese patients. A permutation test determined significance

242 levels as *p*-values of 1.09×10^{-4} and 7.72×10^{-5} for the Japanese and global cohorts,
243 respectively. Two significant genes were detected in the GWAS for the Japanese cohort:
244 *group_184*, which encodes a hypothetical protein, and *divIC*, which encodes a septum
245 formation initiator protein ($p=8.81\times10^{-6}$ and $p=6.72\times10^{-6}$, respectively; Figure 3A and
246 Table S7). Although analysis of the global cohort revealed the presence of 169 genes that
247 were significantly related to invasiveness, no genes were identical or homologous to the
248 two genes detected in the Japanese cohort (Figure 3B, 3C, and Table S8). Among the 169
249 genes, 25 encoded phage-related genes and 14 encoded mobile genetic elements (MGEs)
250 such as transposase, integrase, and recombinase. The gene with the lowest *p*-value was
251 *group_829*, which encodes a transposase and is related to the invasive phenotype (Figure
252 3C and Table S8). In addition, *group_2689*, which encodes a multidrug efflux transporter
253 permease, *rhaR_3*, which encodes a transcriptional regulator involved in rhamnose
254 metabolism, and *group_1829*, which encodes a vitamin B12 import transporter permease,
255 were also substantially associated with the invasive phenotype. We identified several gene
256 distribution patterns associated with invasiveness, suggesting that multiple independent
257 genetic factors cause invasive infections (Figure 3C). Taken together, genes associated
258 with invasiveness were found to encode mobile genetic factors and transporters, whereas
259 major virulence factors were not significantly associated with invasiveness.

260

261 **K-mers-based GWAS detects both distinctive and identical variants compared to the**

262 **SNPs- and COGs-based GWASes**

263 To detect SNPs/indels and multiple mutations in the entire genome, we extracted 31-nt-

264 length k-mers from whole genomes and performed a GWAS. The k-mers-based GWAS

265 can handle polymorphisms spanning more than one base, such as indels, inversions, and

266 translocations, in both the coding and non-coding regions.

267 In the Japanese cohort, the k-mers-based GWAS detected two regions containing

268 causative variants associated with severe invasiveness (Table S9). As shown in the de-

269 Bruijn graphs, overlapping k-mers were concatenated into a single node, and edges

270 represent variability among the sequences (Figure 4A–F). The set of connected nodes and

271 edges comprising each de-Bruijn graph is called a complex. Nodes were determined to

272 be significant if their *q*-value was less than 0.05. Significant nodes are shown in red and

273 blue, indicating an association with severe invasive or non-invasive infections,

274 respectively. The complex comprising the nodes with the lowest *q*-value (q=1.49×10^{-2})

275 was Comp_11 in the *covS* locus (Table S9). The causative mutations in *covS* were an SND

276 and 10 nucleotide polymorphisms (Figure 4A). This deletion resulted in a frameshift

277 mutation that shortened the length of CovS from 500 to 35 amino acids, leading to

278 increased invasiveness, as previously reported in other *emm* types[17]. Another complex
279 significantly associated with severe invasiveness is Comp_2 (q=4.22×10^{-2}; Table S9).
280 Comp_2 is a highly variable region containing eight hypothetical protein-coding genes,
281 with high similarity within the first 75 bp. Significant k-mers were also mapped to the
282 first 26 bp of *group_184* and 20 bp upstream (Figure 4B). These findings suggest that
283 *group_184* possibly contributes to severe invasiveness through not only its presence but
284 also by that of the upstream region.

285 Next, we analyzed the global cohort and identified mutations that were
286 significantly associated with invasiveness in five regions (Table S10). The mutation with
287 the lowest q-value (q=1.90×10^{-2}) existed in Comp_7 and was identical to the SND in *covS*
288 present in the k-mers-based GWAS of the Japanese cohort. Thus, this SND is global, as
289 are the four SNDs detected in the SNPs/indels-based GWAS.

290 Two significant k-mers were present in Comp_6, which were found to be an
291 intergenic region of 270 bp (q=1.90×10^{-2}; Table S10 and Figure 4C). In Comp_24, with
292 a high sequence variation containing *group_141*, *group_142*, and *group_143*, which
293 encode transposases, the presence of a 281-bp sequence consisting of several k-mers was
294 significantly associated with the invasive phenotype (q=1.90×10^{-2}; Table S10 and Figure
295 4D). In Comp_10, n27458, the *sagG* locus encoding the ATP-binding protein of the efflux

296 transporter of SLS, was significantly correlated with the non-invasive phenotype
297 (q=2.40×10^{-2}; Table S10 and Figure 4E). However, a significant SNP in n27458, *sagG*
298 (A882G), was found to cause a synonymous mutation. The other significant mutation was
299 present in the *fhuB* locus, encoding a putative ferrichrome transport system permease
300 (q=2.40×10^{-2}; Table S10). This mutation was identical to SNP T218C detected in the
301 SNP/indels-based GWAS (Figure 4F). Moreover, this mutation changes the 73rd residue
302 from valine to alanine in FhuB, which is a putative ferrichrome transport system permease.
303 Therefore, the k-mer approach identified multiple variants, including the mutation
304 identified in the SNPs/indels-based GWAS, *fhu*B SNP T218C. In addition, while the
305 mutation detected in *covS* differed from that detected in the SNPs/indels-based GWAS,
306 both caused frameshift mutations.
307
308 **AlphaFold-based prediction of the impact of the identified mutations on function**
309 To assess the impact of mutations on protein function, we predicted the protein structure
310 using AlphaFold[29]. Here, we present structural predictions for three representative
311 proteins: LacE, whose mutation was observed mainly in invasive strains from the US;
312 CovS, whose invasion-related deletions prevailed worldwide; and FhuB, which carries a
313 prominent mutation in the Japanese cohort and is associated with non-invasiveness.

314 The invasive-related SNP in *lacE* substitutes the 554th glycine in LacE with valine.
315 LacE was predicted to be a membrane protein with nine transmembrane regions (Figure
316 S4A). LacE is the EIICB component of the lactose-specific phosphotransferase system,
317 and the EIIC and EIIB domains correspond to the transmembrane and intracellular
318 regions, respectively. As shown in magenta in the model, the 554th residue is predicted to
319 be in the intracellular EIIB domain (Figure S4B).

320 Next, we predicted a homodimerized CovS model using AlphaFold, as the CovS
321 of *S. pyogenes* forms homodimers[30]. SOSUI predicted that CovS has two transmembrane
322 regions (Figure 5A). Mutations detected in the SNP/indels- and k-mers-based GWASes
323 were predicted to shorten the CovS protein to 35 and 45 amino acids, respectively. As the
324 intracellular domain of CovS is in the C-terminal region and is involved in the
325 phosphorylation of the transcriptional regulator CovR, frameshift mutations leading to
326 CovS truncation would inactivate the protein, and thus, CovR function (Figure 5B).

327 As described above, the SNP T218C in the *fhuB* locus substitutes the 73rd valine of
328 FhuB with alanine. FhuB is a component of an ATP-binding cassette transporter system
329 that utilizes ferrichrome, which is one of the carriers of Fe^{3+}. FhuB is predicted to localize
330 to the cell membrane and form a channel with FhuG (Figure 5C). SOSUI suggested that
331 FhuB and FhuG are 9-transmembrane proteins. The FhuBG complex can bind to one

332 molecule of the extracellular ferrichrome-binding lipoprotein, FhuD, and two molecules
333 of the intracellular ATP-binding protein, FhuC. Therefore, we constructed a structural
334 model of the FhuBCCDG complex (Figure 5D), which implied that the 73rd residue of
335 FhuB exists in a region adjacent to FhuD. The hydrophobicity of the side chain was
336 attenuated by the mutation, which potentially affected ferrichrome transport (Figure 5E).
337 The SNP G538A in *fhuD* was Japan-specific, significantly related to severe invasiveness,
338 and caused the V180I mutation in FhuD (Figure S3C). The prediction suggested that the
339 180th residue is in an α-helix distant from the active site or the interactive sites with FhuB.
340 Both valine and isoleucine are branched-chain amino acids, and the amino acid residue is
341 located on a rigid structure, the α-helix (Figure 5D). Thus, we believe that this mutation
342 is less likely to cause structural changes than the other non-synonymous mutations.
343

344 **The *fhuB* T218C mutation inhibits the growth of a severe invasive strain in human**
345 **blood**

346 Based on the GWAS results and predicted protein structures, we focused on the SNP *fhuB*
347 T218C. We constructed a mutant strain, in which the SNP *fhuB* T218C was introduced,
348 to further investigate its potential virulence. We selected the strain TK02 which carries
349 the wild-type (WT) allele T218 in *fhuB* and was originally isolated from a sample

350 obtained from a patient with severe invasive infection in Japan. We used a several times-
351 passaged TK02 strain, TK02', as a WT strain and introduced the SNP *fhuB* T218C into it
352 *via* allelic exchange mutagenesis with a thermo-sensitive shuttle vector. We then
353 confirmed that there were no differences between the WT and *fhuB* T218C strains using
354 whole-genome resequencing.

355 To reveal the effects of the SNP on invasiveness, we performed a transcriptomic
356 analysis of the WT and *fhuB* T218C strains in THY broth and human blood. Principal
357 component analysis revealed that the differences in the overall transcriptional profiles
358 between the strains were more remarkable in blood than in THY (Figure 6A, 6B, and
359 Tables S11–14). Among the overlapping genes, we found that the expression of CovR-
360 regulating genes, including *speB*, *nga-ifs-slo* operon, and *sag* operon, was upregulated in
361 the blood, compared to that in THY, in both the strains (Figure 6C, Tables S12, and S13).
362 CovR has also been reported to regulate *sda1*, which plays a key role in invasive disease
363 progression in *emm*1 strains; however, *sda1* does not exist in the *emm*89 reference strain
364 MGAS27061 or WT strain [15]. In human blood, the mutant strain resulted in the
365 downregulation of *mga* and *emm* expression and upregulation of the expression of genes
366 encoding virulence factors, such as *speC*, *scpA/B*, *endoS*, *fba*, *ska*, and *sfbX*, compared to
367 those observed in the WT strain (Table S14). Although Mga regulates the expression of

368 surface and secreted molecules important for colonization and immune evasion[31], no
369 strong expression changes were observed in the Mga regulon, except in the *emm* gene.
370 Notably, *fhuB*, *fhuC*, and *fhuD* were upregulated in the *fhuB* mutant in both human blood
371 and THY (Tables S11 and S14). Moreover, the transcriptomic analysis revealed that both
372 WT and *fhuB* T218C strains displayed upregulation of the *speB, nga-ifs-slo* operon, and
373 *sag* operons in human blood. In addition, the *fhuB* T218C mutation caused the
374 upregulation of *fhuBCD* expression.

375 To determine whether *fhuB* T218C mutation affects iron transport, we measured
376 the intracellular free ferric ion concentration in each strain and observed no differences
377 among the strains in either environment (Figure 7A). Therefore, the upregulation of
378 *fhuBCD* may compensate for the impaired function mediated by SNP T218C.

379 Next, to investigate the effects of SNP on bacterial survival in human blood, we
380 performed a bactericidal assay using human blood. At 2 and 3 h after incubation, the *fhuB*
381 T218C mutant strain exhibited a significantly decreased survival rate than that of the WT
382 strain (Figure 7B). To further determine the blood components that the attenuated survival
383 of the mutant can be attributed to, we compared bacterial survival rates in erythrocytes,
384 polymorphonuclear cells, plasma, heat-inactivated plasma, and brain heart infusion broth
385 (Figure 7C–7G). Notably, after incubation with erythrocytes, polymorphonuclear cells,

386 or plasma, the *fhuB* T218C mutant strain exhibited a significantly lower survival index
387 than that of the WT strain, at 2 and 3 h after incubation, as observed in whole blood
388 (Figure 7C–7E and 7G). However, there were no significant differences between the
389 survival rates of the WT and mutant strains in heat-inactivated plasma, suggesting that
390 the mutant strain is susceptible to complement (Figure 7F). Taken together, the
391 polymorphism T218C in *fhuB* impaired the survival of severe invasive strains in human
392 blood through interactions with erythrocytes.
393

394 **Discussion**

395 The present study focused on *emm*89 *S. pyogenes* causing globally expanding invasive
396 infections. We constructed a workflow for bacterial GWAS and explored the bacterial
397 genetic factors related to severe or invasive infections to reveal the plausible mechanisms
398 of pathogenesis. We independently performed GWASes for Japanese and global cohorts
399 to investigate the associations of variants with strictly defined phenotypes in severe
400 invasive *S. pyogenes* infections and with phenotypes in a broader context of invasive
401 infections. Several GWASes have been conducted on *S. pyogenes* to date. Davies *et al.*
402 extracted k-mers from the genomes of 1,944 strains of *S. pyogenes* with broad *emm*
403 genotypes and assessed their associations with invasiveness, and Kachroo *et al.*

404 performed GWAS on k-mers using 442 *emm*28 strains isolated from either invasive or
405 non-invasive infections[27,32]. However, no studies have described the combined analysis
406 of SNPs, genes, and k-mers exclusively using *emm*89 strains that share genetically close
407 genomes and, thus, a large number of core genes. The investigations we performed led to
408 the discovery of a tremendous number of causative variants, including not only SNPs in
409 core genes but also accessory genes and insertions spanning intergenic regions.

410 Spontaneous mutations in *covR*/*S* genes potentiate the transition from localized to
411 invasive infection by M1T1 *S. pyogenes* [17]. We demonstrated that some SNDs in the *covS*
412 locus were significantly related to the phenotypes in both the SNPs/indels- and k-mers-
413 based GWASes. CovS forms a two-component system with CovR that regulates the
414 transcription of operons encoding virulence factors when phosphorylated by the histidine
415 sensor kinase CovS[33]. Sumby *et al.* suggested that systemic invasive infections are caused
416 by the overexpression of virulence factors through the inactivation of CovS, as a result of
417 the introduction of a depletive mutation in the *covS* locus after infection[15,16]. Several
418 mutation sites causing such overexpression, such as the 7 nt insertion in *covS* of *emm*1 *S.*
419 *pyogenes* and the insertion of adenine at the 877th nucleotide, have been reported[15,16].
420 Ikebe *et al.* demonstrated variations in the lengths of mutant CovS proteins by analyzing
421 the gene sequences of 164 STSS strains in Japan[18]. In line with their study, we found that

422 the shortening of CovS that potentially occurred owing to frameshifts caused by SNDs is
423 significantly associated with invasive infections, potentially contributing to the
424 upregulation of virulence factors and invasiveness.

425 In contrast, we observed several factors associated with invasiveness to be
426 independent of *covR*/*S* mutations. Based on conformational predictions, we selected SNPs
427 with non-synonymous substitutions that were likely to affect protein function. Our
428 transcriptomic analysis suggested that the Japan-specific *fhuB* mutation contributes to the
429 growth rate of *S. pyogenes* in human blood by adapting to the environment. In addition,
430 *emm*89 clade 3 carries the identical promoter region pattern of the *nga* operon as *emm*1
431 strains, and the pattern conferred similarly high expression of *nga* and *slo*[11,12,28]. Our
432 RNA-seq data demonstrated that a severe invasive Japanese strain without *covS* mutations
433 increased the expression of *speB*, *nga*, *slo*, and genes in the *sag* operon in the blood.
434 Although *covR*/*S* mutations downregulate the expression of SpeB and SLS, SpeB and
435 SLS act as virulence factors, allowing *S. pyogenes* to invade host tissues[34-37]. These data
436 suggest that severe invasive infections have multiple gene expression profiles in addition
437 to *covR*/*S* mutation-induced profiles within a single lineage, *emm*89, and that the synergy
438 between optimizing bacterial survival in human blood and upregulating multiple
439 virulence factors contributes to the development of severe invasive infections.

440 In addition, we propose two possible roles of the *FhuB* V73A mutation in the
441 pathogenesis of severe invasive infections. First, the mutation in *FhuB* could increase
442 bacterial susceptibility to free radicals generated in the presence of ferric ions in the blood.
443 Generally, ferric ions are essential for the survival of almost all living organisms.
444 Catalyzed by ferric ions, the Fenton and Haber–Weiss reactions generate hydroxyl
445 radicals in erythrocytes[38]. Previously, we provided evidence that the iron in erythrocytes
446 partially inhibits pneumococcal growth *via* a free radical-based mechanism[7]. Although no
447 significant difference was observed in the intracellular ferric ion concentration in the
448 present study, the structural changes in FhuB may have caused an increased generation of
449 free radicals in bacterial cells and the prevention of bacterial survival in human blood.
450 Second, the *FhuB* mutation could affect bacterial transcriptional profiles and result in
451 reduced fitness for survival in human blood. We also observed a lower survival rate of the
452 mutant strain with polymorphonuclear cells (Figure 7D). Lower *emm* transcription in the
453 mutant may lead to attenuated immune evasion. The M protein can bind C4-binding
454 protein and factor H and inactivate the deposited C4b and C3b, leading to limited surface
455 opsonization[39]. Furthermore, complements could have inhibited the proliferation of the
456 mutant strain, as suggested by the activation of the membrane attack complex (Figure 7D
457 and 7E). Although the mechanisms of the *FhuB* V73A mutation in the interaction with

458 each blood component must be further elucidated, our results indicated that the 73rd
459 residue of FhuB is a key factor for bacterial survival in human blood during systemic
460 infection.

461 Recently, He *et al.* reported that antibodies against the ferrichrome-binding protein
462 FtsB, also known as FhuD, decrease blood bacterial burden and skin abscess formation
463 in murine models infected with *emm*1 *S. pyogenes*[40]. FhuD is a lipoprotein that has been
464 studied as a potential vaccine candidate against *Staphylococcus aureus*[41]. The detailed
465 function and effect of the FhuBCCDG complex in *S. pyogenes* remain unknown; however,
466 at least in *emm*1 strains, the complex can be expressed and localized abundantly on the
467 cell membrane. Hence, this complex may be a promising target for vaccines against *S.*
468 *pyogenes*.

469 Our gene-based GWAS revealed no significant correlation with the distribution of
470 genes encoding virulence factors. This finding minimizes the possibility that invasive
471 infections are the result of the acquisition of virulence factors by non-invasive strains and
472 indirectly supports the hypothesis that changes in gene expression profiles cause invasive
473 infections. Although no vaccine has been commercialized against *S. pyogenes*, given the
474 possibility that the pathogen has multiple gene expression patterns, even within a single
475 lineage, it may be difficult to develop a universal vaccine with a single antigen, and a

476 vaccine containing multiple antigens may be effective.

477 In the global cohort, 169 genes were related to the phenotype, including 25 phage-
478 and 14 MGE-related genes. The genome of *S. pyogenes* is rich in prophages, phage-like
479 chromosomal islands, and MGEs [42]. Prophages and MGEs sometimes function as vectors
480 of virulence factors and antibiotic-resistance genes through interspecies and intraspecific
481 transmission. Common superantigen-coding genes, such as *speA*, *speC*, *speG*, *speH*, *speI*,
482 *speJ*, *speK*, and *speL*, and DNase-coding genes, such as *spd1*, are exogenous genes
483 derived from prophages[43]. In addition, in the genome of *S. pyogenes,* there are broad
484 MGEs where *erm* and *mefA* related to macrolide resistance and *tetM* involved in
485 resistance to tetracyclines are located[44]. In the present study, we observed no direct
486 association between invasiveness and virulence or antibiotic-resistance genes located in
487 the prophages and MGEs. One possible explanation is that significantly related genes
488 reflect the process of not only gain but also loss of factors affecting the fitness cost. Given
489 that we also identified hypothetical genes associated with this phenotype, we assumed
490 that there is an unknown mechanism contributing to the pathogenesis of invasive
491 infections.

492 In the present study, MLST analysis revealed that 522 of the 666 strains belonged
493 to ST101. Following ST101, ST646 was the second most prevalent lineage in the

494 Japanese cohort and was not detected in any other country. In addition, the difference
495 between ST101 and ST646 was found only at the 295th nucleotide in *mulI*, among the
496 seven genes subjected to MLST typing. Our phylogenetic trees also suggested that ST101
497 and ST646 have a close phylogenetic relationship. Five previous studies have described
498 ST646 in *S. pyogenes*, and all of them reported *emm*89 ST646 strains isolated in Japan[45-
499 49]. Furthermore, all ST646 strains were isolated from the 4,249 *emm*89 *S. pyogenes* strains
500 in the MLST database PubMLST[50]. Therefore, ST646 appears to be specific to *emm*89
501 strains and a unique lineage in Japan. Notably, Ubukata *et al.* also reported that ST646
502 strains began increasing after 2012, whereas the ST101 lineage was dominant until 2008[49].
503 Taken together, we conclude that ST646 is possibly a relatively new and Japan-specific
504 lineage of *emm*89 *S. pyogenes.*

505 Our study had three limitations. First, because we sequenced bacterial genomes
506 using short-read sequencing, we could not detect large bacterial genome rearrangements
507 by comparing complete genome sequences. Thus, we were unable to investigate the
508 effects of long genomic structural dynamics, such as inversions, on pathogenesis. Second,
509 the clinical information associated with the isolates was limited; therefore, our analyses
510 did not reflect host information, such as age, sex, and underlying health conditions. Third,
511 we only used bacterial genetic distances based on core genome sequences as covariates

512 in the GWAS. A combined GWAS of the host and pathogen in *S. pyogenes* infection would
513 highlight the relationship between host risk factors and bacterial genetic variants.
514 In this study, we revealed the genotype-phenotype associations found in not only
515 the known factors represented by *covS* but also factors that are related to invasiveness,
516 including *fhuB*. Moreover, we experimentally validated the contribution of the *fhuB*
517 mutation to bacterial survival in human blood. This study demonstrates the potential of
518 our genomic statistical approach for elucidating the pathogenesis of invasive infections.
519 Further analyses of the invasiveness-related factors identified in this study could provide
520 a platform for establishing novel treatments and preventive strategies against invasive
521 infections.
522

523 **Acknowledgments**

524 We would like to thank the NGS core facility of the Genome Information Research Center
525 at the Research Institute for Microbial Diseases of Osaka University for their support in
526 the DNA sequencing and data analysis and the Bioinformatic Research Unit of Osaka
527 University Graduate School of Dentistry for their support in the bioinformatics analysis.
528 This study was partially performed on the National Institute of Genetics (NIG)
529 supercomputer at the Research Organization of Information and Systems National

Institute of Genetics. This study was partly completed using SQUID at the Cybermedia Center, Osaka University, under the "Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures (JHPCN)" in Japan (Project ID: EX22701 and jh230035). Masayuki Ono and Kotaro Higashi were recipients of the Iwadare Scholarship from the Iwadare Scholarship Foundation. We wish to express our gratitude to Mami Tateshita (Sapporo City Institute of Public Health) as well as the medical institutions that participated in the collection of clinical strains.

This study was partly supported by AMED (JP19fk0108044, JP22fk0108130, and JP22wm0325001), the Japan Society for the Promotion of Science KAKENHI (grant numbers 20KK0210, 22H03262, 22K19618, 22K19619, 23H03073, and 23K19687), the Takeda Science Foundation, Naito Foundation, and Joint Research Program of the Research Center for GLOBAL and LOCAL Infectious Diseases, Oita University (2022B05). This study was conducted as part of "The Nippon Foundation - Osaka University Project for Infectious Disease Prevention." This study was supported by JST SPRING (grant number JPMJSP2138). The funders had no role in the study design, data collection or analysis, decision to publish, or preparation of the manuscript.

Author contributions

548 M.Y., M.O., and S.K. designed the study. M.O., M.Y., and K.H. designed and performed
549 bioinformatics analysis. D.M. performed the next-generation sequencing. M.O.
550 performed the in vitro and in vivo experiments. M.Y., Y.H., T.S., and S.K. contributed to
551 the setup of the experiments. R.O., T.Y., R.K., H.O., N.N., Y.K., C.N., R.Y., H.S., and
552 Y.M. collected clinical strains. M.O. and M.Y. wrote the original manuscript. All authors
553 reviewed and edited the manuscript.
554

555 **Declaration of interests**

556 The authors declare no competing interests.
557

558 **Data availability**

559 Data for the 207 sequenced *S. pyogenes* genomes were deposited in the DDBJ sequence
560 read archive, under BioProject PRJDB16457. The DRR run number is DRR511668-
561 DRR511874.
562

563 ## Figure titles and legends

564 **Figure 1. Pan-genomic and phylogenetic characterization of the collected emm89 S.**
565 **pyogenes isolates.**

566 (A) *emm* genotyping of the 207 clinical isolates collected in Japan. (B) Pan-genome
567 analysis of the Japanese and global cohorts. All genes detected in each cohort were
568 classified into four groups, according to prevalence: core, soft-core, shell, and cloud genes.
569 (C) Phylogenetic tree for the global cohort, based on the sequences of the core genes.
570 From the inside, the color bars show clusters, phenotypes, MLSTs, clades, countries
571 where strains were isolated, and global regions, in the order mentioned. The root of the
572 tree was set as the mid-point. The scale located upper left indicates 0.001 times
573 substitution of the bases on average. MLST, multilocus sequence typing.

574

575 **Figure 2. Genome-wide association study on SNPs/indels.**

576 (A–B) Manhattan plots for the Japanese (A) and global (B) cohorts. The X-axis shows
577 the location of each SNP/indel on the core gene alignment, while the Y-axis indicates the
578 *p*-value. For each cohort, a permutation test was performed by iterating the calculations
579 1,000 times with randomly permuted phenotypes, with the significance level set at the 5th
580 percentile of the 1,000 minimal *p*-values ($p=5.75\times10^{-4}$ and $p=1.05\times10^{-4}$ for the Japanese

581 and global cohorts, respectively). Plots with lower *p*-values than genome-wide significant
582 levels are colored magenta and blue, based on the direction of their effect size (positive
583 and negative, respectively). (C) Distribution heatmap for the global cohort, with the
584 strains possessing the significant SNPs/indels colored orange. Only the 20 SNPs with the
585 lowest *p*-values are shown in this heatmap. The color bars above indicate countries and
586 phenotypes. The position is the location of each SNP/indel on the core gene alignment.
587 SNP, single-nucleotide polymorphism; indel, insertion/deletion.
588
589 **Figure 3. Genome-wide association study on gene presence.**
590 (A–B) Volcano plots for the Japanese (A) and global (B) cohorts. The X-axis shows the
591 effect size, while the Y-axis indicates the *p*-value. Plots with a lower *p*-value than the
592 genome-wide significant levels (p=1.09×10^{-4} and p=7.72×10^{-5} for the Japanese and
593 global cohorts, respectively) have been colored magenta and blue, based on the direction
594 of their effect size (positive and negative, respectively). (C) Distribution heatmap for the
595 global cohort, with the strains possessing the significant genes colored orange. Only the
596 20 genes with the lowest *p*-values are shown in this heatmap. The color bars above
597 indicate countries and phenotypes.
598

599 **Figure 4. K-mers related to pathology.**

600 Detailed results of the GWAS on k-mers for the Japanese cohort in two genomic regions:
601 *covS* (A, top left) and *group_184* (B); and the global cohort in five regions: *covS* (A, top
602 right), an intergenic region (C), *group_141-143* (D), *sagG* (E), and *fhuB* (F). (A–F top)
603 de-Bruijn graphs generated using DBGWAS. The respective nodes in the graphs indicate
604 k-mers, with the significant nodes indicated using green arrowheads. The size of each
605 node corresponds to the allele frequency. (A–F, bottom) Alignments of k-mers or maps
606 of k-mers on the genome sequence of MGAS27061 around regions including significant
607 mutations. Each base is colored according to the interpreted amino acids. (A) The
608 arrowheads indicate significant k-mers that cause a frameshift mutation, truncating CovS
609 protein to 35 amino acids. (B) The significant k-mers indicate that the presence of a
610 sequence mapped on the first 26 bp of group_184 and its upstream 20 bp is related to the
611 pathology. (C) Two significant k-mers indicate the same intergenic region of 270 bps. (D)
612 The significant k-mers are indicated using arrowheads.

613

614 **Figure 5. Predicted protein structure models.**

615 (A) Snake-like plot of transmembrane regions of CovS estimated using SOSUI.
616 Frameshift mutations, detected by the SNPs/indel- and k-mer-based GWASes, cause

617 truncation of CovS at the indicated 35th and 45th residues, respectively. (B) Structural
618 model of the CovS homodimer (ipTM + pTM=0.614). The putative transmembrane
619 regions are colored orange. The upper part is the sensor domain, while the lower is the C-
620 terminal kinase domain involved in the phosphorylation of the transcriptional regulator
621 CovR. (C) Snake-like plot depicting the transmembrane regions of FhuB and FhuG. The
622 73rd residue of FhuB is indicated by an arrowhead and in magenta. (D) Structural model
623 of the FhuBDCCG complex (ipTM + pTM=0.791). The putative transmembrane regions
624 of FhuB and FhuG are colored green-yellow and peach, respectively. The upper part of
625 the model is located in the extracellular region, while the lower part is in the cytoplasm.
626 (E) The 73rd valine in FhuB, colored in magenta, was substituted with alanine.
627 ipTM + pTM: Weighted combination of interface-predicted TM and predicted TM scores.
628 ipTM is used to measure structural accuracy in the protein-protein interface, while ipTM
629 is a metric for overall topological accuracy.
630
631 **Figure 6. Transcriptome analysis of the fhuB T218 mutant strain in THY and human**
632 **blood.**
633 (A) Principal component analysis plot of RNA-seq data. (B) Differentially expressed
634 genes in four comparisons. (C) Plot of gene expression in the WT strain *vs*. *fhuB* T218C

635 mutant in human blood. Significantly upregulated and downregulated genes in the WT
636 are colored red and blue, respectively. The shapes of the plots indicate relative
637 transcriptional changes between THY and blood. Genes depicted with upward triangles
638 are either significantly upregulated in the WT or downregulated in the mutant strain, in
639 blood *vs.* THY. The downward triangle plots indicate genes that are either downregulated
640 in the WT or upregulated in the mutant strain, in blood *vs.* THY. WT, wild-type; THY,
641 Todd Hewitt broth supplemented with 0.2% yeast extract.
642
643 **Figure 7. Effects of the fhuB T218 mutant strain on ferric ion uptake and bacterial**
644 **survival in human blood.**
645 (A) Intracellular ferric ion assay. The WT and *fhuB* T218C mutant strains were incubated
646 in healthy human blood for 3 h, following which the intracellular ferric ion concentrations
647 were measured (*n*=18 technical replicates). (B) Bactericidal assay in human blood. The
648 WT, *fhuB* T218C mutant, and Wr strains were mixed with healthy human blood to
649 measure the ratio of bacterial counts at 1, 2, and 3 h to those at 0 h after infection (*n*=18
650 technical replicates). (C–F) Bactericidal assay in human erythrocyte-rich medium (C),
651 polymorphonuclear cell-rich medium (D), plasma (E), and plasma inactivated by heating
652 at 56ºC for 30 min (F) (*n*=18 technical replicates). (G) Bacterial growth in brain heart

653 infusion broth (*n*=18 technical replicates). Representative data from one of three
654 biological replicates are displayed. Thick bars and error bars indicate means and quartiles,
655 respectively. Statistical significance was determined using the Mann–Whitney test with
656 Benjamini–Hochberg's correction.
657

658 ## Materials and methods

659 **Clinical isolates in Japan**

660 Clinical isolates were collected from public health institutions in Tokyo, Osaka,
661 Yamaguchi, Fukushima, Kobe, Kyoto, Amagasaki, Sapporo, and Niigata, Japan. We
662 defined the strains collected as STSS according to the Infectious Diseases Control Law
663 in Japan. Non-invasive strains were defined based on diagnostic names, including
664 asymptomatic, pharyngitis, tonsillitis, or non-invasive infections. Strains with no
665 diagnostic names for the non-STSS strains were defined as non-invasive based on the
666 isolate sites. Information on all the strains included in this study is presented in Table S3.

667 The collected *S. pyogenes* strains were cultured at 37ºC in an atmosphere
668 containing 5% CO_2, in Todd Hewitt broth supplemented with 0.2% yeast extract (THY;
669 both from BD Biosciences, Franklin Lakes, NJ, USA) and stored in THY broth with 30%
670 glycerol (Nacalai Tesque, Kyoto, Japan), at –80°C.

671

672 **Genomic DNA sequencing of the clinical isolates**

673 The *S. pyogenes* strains were cultured until the exponential growth phase (OD_{600}=0.3–
674 0.4). Bacterial cells were lysed with T10E1N100 buffer (10 mM Tris-HCl buffer, 100
675 mM sodium chloride, and 1 mM EDTA), 10 units/mL mutanolysin (Sigma-Aldrich, St.
676 Louis, MO, USA), 10 mg/mL lysozyme (Fujifilm Wako Pure Chemical Co., Osaka,
677 Japan), 0.5 mg/mL achromopeptidase (Fujifilm Wako Pure Chemical Co.), and 0.3
678 mg/mL RNase A (Promega, Madison, WI, USA). Next, genomic DNA was extracted
679 from each lysate using a Maxwell® RSC instrument (Promega), according to the
680 manufacturer's instructions, and 250 bp paired-end libraries were then generated from the
681 extracted DNA using a Nextera XT DNA Kit (Illumina, San Diego, CA, USA). Libraries
682 were sequenced using a NovaSeq 6000 system (Illumina) at the Genome Information
683 Research Center, Research Institute for Microbial Diseases, Osaka University, Osaka,
684 Japan. On average, the number of reads was 5,433,301 (range 3,437,124–9,117,301).

685

686 **Collection of published genome sequences**

687 We previously sequenced the draft genomes of 161 *emm*89 clinical isolates collected in
688 Japan between 2011 and 2019[14]. We defined strains derived from STSS as "severe

689 invasive," and those obtained from pharyngitis, tonsillitis, and superficial skin lesions as
690 "non-invasive" phenotypes.
691 To obtain public genome sequences of *emm*89 strains isolated from other countries,
692 we downloaded draft genome sequences in FASTA format from the National Center for
693 Biotechnology Information (NCBI) database, using Fasterq-dump v.2.9.6. The phenotype
694 of each strain, whether invasive or non-invasive, was defined according to the definitions
695 in the respective references that reported the strains[25-27].
696
697 **Genomic data processing and pan-genome analysis**
698 All processes and analyses were performed using the National Institute of Genetics (NIG)
699 supercomputer and SQUID at the Cybermedia Center of Osaka University (Osaka, Japan).
700 We constructed a workflow for bacterial GWAS and other bioinformatic processes
701 (Figure 5). All collected sequences were subjected to quality checks using Fastp v.0.20.1[51].
702 For newly isolated strains in Japan, *emm* typing was performed using the emm-typing-
703 tool v.0.0.1, and only sequences of strains determined as *emm*89 were used for the
704 following analyses[52]. All *emm*89 sequence data were then subjected to *de novo* assembly
705 using SKESA v.2.4.0, with default parameters[53]. Next, the MLST of each sequence was
706 performed using MLST v.2.19.0[54,55]. Clade typing was performed using BLAST v.2.13.0,

707 with reference to the three *nga* promoter region sequences[13]. After the genes were
708 annotated with Prokka v.1.14.5, the pan-genome of all sequences was calculated using
709 Roary v.3.12.0, with the parameters "-e -mafft -r -qc -cd 99"[56,57]. Roary generated a core
710 gene alignment and the distribution of all genes among the strains. To extract SNPs/indels
711 from core genes, including single-nucleotide indels, snp-sites v.2.5.1 with the option "-v"
712 was used. The output files were further processed using BCFtools v.1.9, with the
713 parameter "norm -m –", enabling analysis of multiple alleles in the GWAS. In parallel, k-
714 mers were extracted using DBGWAS v.0.5.3, and the length of k-mers was set as 31 nt
715 using the "-k 31" parameter of DBGWAS[58].
716
717 **Phylogenetic analysis**
718 Phylogenetic relationships were calculated from the core gene alignment, using IQ-tree
719 v.1.16.12[59], based on maximum likelihood. The substitution model was automatically
720 selected considering the Akaike's and Bayesian information criteria by setting the "-m
721 MFP" parameter of IQ-tree[60]. Phylogenetic trees were constructed using iTOL v.6.6[61].
722 The similarity of clustering in the two phylogenetic trees was statistically examined using
723 Pearson's chi-square test with R v.4.0.3[62], followed by *post-hoc* analysis using residual
724 analysis adjusted with the Holm's method, if $p<0.05$.

725

726 **GWAS**

727 To investigate the associations between phenotypes and genotypes, including SNPs/indels
728 and genes, we performed a GWAS using pyseer v.1.3.4[63]. The VCF file of SNPs/indels or
729 the gene distribution matrix was designated as the genotype. To remove biases derived
730 from lineages, we added information on the genetic distances between all pairs of strains
731 as covariates using mash v.2.3[64]. Briefly, *de novo* assemblies were compressed through
732 conversion into minimum hash values using the command "mash sketch -s 10000."
733 Subsequently, the commands "mash dist" and "square_mash" were utilized to generate a
734 genetic distance matrix expressed with Jaccard coefficients[64]. The obtained matrix
735 underwent eigenvalue decomposition, and the number of eigenvalues used for
736 multidimensional scaling was visually determined using the plot of the relationships
737 between the eigenvalues and contribution ratios. The number of eigenvalues and the
738 distance matrix were then added as pyseer parameters. The pyseer calculation was iterated
739 1,000 times with randomized phenotypes, and the 5-percentile value of the minimal *p*-
740 value in each calculation was set at the significance level. Using R and the package
741 ggplot2, the results were visualized as a Manhattan plot for the SNPs/indels-based GWAS
742 and a volcano plot for the genes-based GWAS, respectively[65]. Heatmaps of the strains

743 possessing significant variants were generated using Excel (v.16.66.1; Microsoft,
744 Redmond, WA, USA).

745 The k-mers-based GWAS was carried out using DBGWAS[58]. K-mers were
746 considered significant at a false discovery rate (*q*-value) of <0.05. DBGWAS-calculated
747 complexes, which are regions encompassing the k-mers, were significantly related to
748 pathology, and de-Bruijn graphs were generated based on these complexes. The
749 sequences of the k-mers were outputted and mapped to a reference sequence using
750 Geneious Prime v.2022.0.1 (Biomatters, Auckland, New Zealand) to identify the
751 mutations indicated by the k-mers. For the reference strain, we adopted MGAS27061,
752 which was isolated from an invasive case in the USA and whose complete chromosomal
753 sequence has been used as the reference sequence of *emm*89 clade 3 [66].

754

755 **Protein structure prediction**

756 Significant variants found in the GWAS, resulting in non-synonymous substitutions in
757 proteins, were searched by converting nucleotide sequences into amino acid sequences
758 using EMBOSS Transeq v.6.6.0.0[67]. To assess whether these mutations affected the
759 protein function, protein structure prediction models were constructed using AlphaFold
760 v.2.2.2[29]. The calculations were performed five times for each model, by setting the option

761 multimer_predictions_per_model=5. We predicted multimer models using the option --
762 model_preset="multimer" if a protein is reported or anticipated to form a multimer. For
763 each monomer, we selected the model with the best predicted local difference distance
764 test (an indicator of local structural accuracy) score[68]. For each multimer, AlphaFold was
765 calculated and expressed as a weighted combination of the interface-predicted TM and
766 predicted TM scores (ipTM + pTM). pTM is a metric for overall topological accuracy
767 and ipTM is used to measure the structural accuracy of the protein-protein interface[69].
768 The transmembrane regions of the proteins were predicted using SOSUI
769 (https://harrier.nagahama-i-bio.ac.jp/sosui/mobile/)[70]. The structures of the obtained
770 model were visualized using PyMOL v.2.5 (Schrödinger, LLC., New York, NY, USA).
771

772 **Construction of the *fhuB* T218C mutant strains**

773 We used the several times-passaged *S. pyogenes* TK02 strain, TK02', as the WT strain.
774 The TK02 strain was originally isolated from a patient with severe invasive infection[14].
775 The whole genome of TK02' was sequenced, and the generated fasta file is available in
776 Document S1. A point mutation, *fhuB* T218C, was introduced using the temperature-
777 sensitive shuttle vector, pSET4s, as reported previously[71]. Sanger sequencing confirmed
778 the presence of the point mutation. In addition, we resequenced the draft genome to

779 confirm that there were no differences apart from the point mutation, as described above.

780 The bacterial strains, primers, and plasmids used in this study are listed in Tables
781 S15 and S16. *Escherichia coli* strain DH5α (Takara Bio, Shiga, Japan) was used as a host
782 for the plasmid derivatives. All *E. coli* strains were cultured in Luria Bertani broth, at
783 37°C, with agitation. For selection and maintenance of mutants, antibiotics were added
784 to the media at the following concentrations: carbenicillin (Nacalai Tesque), 100 μg/mL
785 for *E. coli*; and spectinomycin (Fujifilm Wako Pure Chemical Corporation), 100 μg/mL
786 for *E. coli* and *S. pyogenes.*

787

788 **Transcriptomic analysis**

789 The *fhuB* WT and mutant strains were harvested in 30 mL of THY broth, of which 1 mL
790 was dispensed into 10 mL of THY and the remainder was centrifuged and resuspended in
791 2 mL of heparinized human blood. Bacterial mixtures with THY or blood were dispensed
792 into three aliquots and incubated at 37ºC for 3 h. THY samples were centrifuged and
793 resuspended in RNA Shield (Zymo Research, Irvine, CA, USA). For each blood sample,
794 2 volumes (1 mL) of RNA protection bacteria reagent (Qiagen, Hilden, Germany) were
795 added. L5, included in the PureLink™ Total RNA Blood Purification Kit (Thermo Fisher
796 Scientific, Waltham, MA, USA), was added to remove erythrocytes. The bacterial cell

797 wall was mechanically lysed in Lysing Matrix B using a MagNA Lyser (Roche, Basel,
798 Switzerland). After centrifugation, the total bacterial RNA was extracted using a Quick-
799 RNA™ Miniprep Kit (Zymo Research), according to the manufacturer's instructions.
800 Full-length cDNA was generated using the SMART-Seq® HT Kit (Takara Bio), according
801 to the manufacturer's instructions. Pair-end libraries were generated using a Nextera XT
802 DNA Kit and sequenced using a NovaSeq 6000 system (both from Illumina, San Diego,
803 CA, USA). Sequenced data were preprocessed using Trimmomatic v.0.33 and FastQC
804 v.0.12.1. The reads were mapped to the complete MGAS27061 genome (NCBI reference
805 sequence: NZ_CP013840.1) using STAR v.2.7.0a. After a second quality check using
806 FastQC, read counting was performed using featureCounts v.1.5.2 [2]. Differentially
807 expressed genes were identified using iDEP v.0.96 and gene annotations from NCBI and
808 Prokka were combined[73]. Plots were created using iDEP and the R package ggplot2.
809

810 **Intracellular ferric ion assay**

811 Human plasma was obtained through centrifugation of heparinized human blood, after 30
812 min of incubation at 37ºC. The WT and *fhuB* mutant strains were harvested at the
813 exponential phase, resuspended into 1 mL of THY or serum, and then incubated at 37ºC
814 for 3 h. Viable bacterial cells were counted as colony-forming units (CFUs) by plating the

815 diluted samples onto THY agar plates. Intracellular ferric ions were measured using a
816 QuantiChrom™ Iron Assay Kit (BioAssay Systems, Hayward, CA, USA), according to
817 the manufacturer's instructions. Briefly, 50 μL of standards or samples in 96-well plates
818 were mixed with 200 μL QuantiChrom™ Working Reagent and incubated at 20–24˚C for
819 an hour. The optical density at the wavelength of 600 nm was measured using an Infinite®
820 200 Pro F Plex Instrument (TECAN, Männedorf, Switzerland). The assay was repeated
821 three times, and the results of the respective experiments were combined. Statistical
822 analyses were performed using the Mann–Whitney U test.
823
824 **Bactericidal assay**
825 The bactericidal assay was performed as described previously, with minor
826 modifications[34,74-76]. Briefly, whole blood was collected from healthy adults. Human
827 neutrophils and erythrocytes were prepared using PolymorphPrep™ (Serumwerk
828 Bernburg, Bernburg, Germany), according to the manufacturer's instructions.
829 Heparinized human blood was centrifuged at 500 × *g* for 30 min, to isolate erythrocytes
830 and polymorphonuclear cells, which were then suspended in Roswell Park Memorial
831 Institute (RPMI)-1640 medium containing L-glutamine and Phenol Red (Fujifilm Wako
832 Pure Chemical Corporation). Heat-inactivated plasma was prepared at 56ºC for 30 min.

833 Subsequently, 195 μL of heparinized human whole blood, erythrocytes in RPMI-1640,
834 polymorphonuclear leukocytes in RPMI-1640, plasma, heat-inactivated plasma, or brain
835 heart infusion broth (BD Biosciences), and 5 μL of early exponential phase bacteria with
836 $0.9–2.0\times10^4$ CFUs/well were mixed in 96-well plates and incubated at 37°C, in an
837 atmosphere containing 5% CO_2, for 1, 2, and 3 h. Viable bacterial cells were counted as
838 CFUs by plating the diluted samples onto THY agar plates. The growth index was
839 calculated as the number of CFUs at a specified time point divided by the number of
840 CFUs in the initial inoculum. The assay was repeated three times, and the results of the
841 respective experiments were combined. Statistical analyses were performed using the
842 Mann–Whitney U test. Differences were considered statistically significant at $p<0.05$,
843 using Prism v.7.0c (GraphPad, La Jolla, CA, USA).
844

845 **Ethical approval**

846 Studies involving human participants were reviewed and approved by the Institutional
847 Review Board of Osaka University Graduate School of Dentistry (approval nos. H26-E43
848 and H29-E16-2). The donors provided written informed consent to participate in the
849 human blood bactericidal assay. For the *S. pyogenes* collection, as we retrospectively
850 obtained clinical isolates of *S. pyogenes*, we utilized an opt-out consent procedure instead

of obtaining written informed consent from the patients.

Supplemental information titles and legends

Figures S1–S5.

Tables S1–S16. Excel file containing additional data too large to fit in a PDF.

Document S1. A pair of FASTA files containing short-read sequences.

References

1. Craik, N., Hla, T., Cannon, J., Moore, H., Carapetis, J.R., and Sanyahumbi, A. (2022). Global Disease Burden of *Streptococcus pyogenes*. In *Streptococcus pyogenes*: Basic Biology to Clinical Manifestations [Internet] J.J. Ferretti, D.L. Stevens, and V.A. Fischetti, eds. (University of Oklahoma Health Sciences Center), Chapter 21.
2. Brouwer, S., Rivera-Hernandez, T., Curren, B.F., Harbison-Price, N., De Oliveira, D.M.P., Jespersen, M.G., Davies, M.R., and Walker, M.J. (2023). Pathogenesis, epidemiology and control of Group A *Streptococcus* infection. Nat. Rev. Microbiol. 21, 431–447. 10.1038/s41579-023-00865-7.

868 3. Carapetis, J.R., Steer, A.C., Mulholland, E.K., and Weber, M. (2005). The global
869 burden of group A streptococcal diseases. Lancet Infect. Dis. 5, 685-694.
870 10.1016/S1473-3099(05)70267-X.
871 4. Stevens, D.L., Bisno, A.L., Chambers, H.F., Dellinger, E.P., Goldstein, E.J.,
872 Gorbach, S.L., Hirschmann, J.V., Kaplan, S.L., Montoya, J.G., Wade, J.C., et al.
873 (2014). Practice guidelines for the diagnosis and management of skin and soft
874 tissue infections: 2014 update by the Infectious Diseases Society of America. Clin.
875 Infect. Dis. 59, e10–e52. 10.1093/cid/ciu444.
876 5. Stevens, D.L., and Bryant, A.E. (2016). Severe Group A Streptococcal Infections.
877 In *Streptococcus pyogenes*: Basic Biology to Clinical Manifestations [internet],
878 J.J. Ferretti, D.L. Stevens, and V.A. Fischetti, eds. (University of Oklahoma
879 Health Sciences Center).
880 6. Walker, M.J., Barnett, T.C., McArthur, J.D., Cole, J.N., Gillen, C.M., Henningham,
881 A., Sriprakash, K.S., Sanderson-Smith, M.L., and Nizet, V. (2014). Disease
882 manifestations and pathogenic mechanisms of Group A Streptococcus. Clin.
883 Microbiol. Rev. 27, 264–301. 10.1128/CMR.00101-13.
884 7. Yamaguchi, M., Terao, Y., Mori-Yamaguchi, Y., Domon, H., Sakaue, Y., Yagi, T.,
885 Nishino, K., Yamaguchi, A., Nizet, V., and Kawabata, S. (2013). *Streptococcus*

886 *pneumoniae* invades erythrocytes and utilizes them to evade human innate
887 immunity. PLOS ONE 8, e77282. 10.1371/JOURNAL.PONE.0077282.
888 8. Bessen, D.E., Smeesters, P.R., and Beall, B.W. (2018). Molecular Epidemiology,
889 Ecology, and Evolution of Group A Streptococci. Microbiol. Spectr. 6.
890 10.1128/microbiolspec.CPP3-0009-2018.
891 9. Ikebe, T., Okuno, R., Kanda, Y., Sasaki, M., Yamaguchi, T., Otsuka, H., Kazawa,
892 Y., Suzuki, M., Ohya, H., Uchida, K., and Ohnishi, M. (2021). Molecular
893 characterization and antimicrobial resistance of group A streptococcus isolates in
894 streptococcal toxic shock syndrome cases in Japan from 2013 to 2018. Int. J. Med.
895 Microbiol. 311, 151496. 10.1016/j.ijmm.2021.151496.
896 10. Fay, K., Onukwube, J., Chochua, S., Schaffner, W., Cieslak, P., Lynfield, R., Muse,
897 A., Smelser, C., Harrison, L.H., Farley, M., et al. (2021). Patterns of Antibiotic
898 Nonsusceptibility Among Invasive Group A *Streptococcus* Infections-United
899 States, 2006–2017. Clin. Infect. Dis. 73, 1957–1964. 10.1093/cid/ciab575.
900 11. Turner, C.E., Abbott, J., Lamagni, T., Holden, M.T.G., David, S., Jones, M.D.,
901 Game, L., Efstratiou, A., and Sriskandan, S. (2015). Emergence of a new highly
902 successful acapsular group a Streptococcus clade of genotype *emm*89 in the
903 United Kingdom. mBio 6, e00622. 10.1128/mBio.00622-15.

904 12. Zhu, L., Olsen, R.J., Nasser, W., Beres, S.B., Vuopio, J., Kristinsson, K.G.,
905 Gottfredsson, M., Porter, A.R., DeLeo, F.R., and Musser, J.M. (2015). A molecular
906 trigger for intercontinental epidemics of group A Streptococcus. J. Clin. Invest.
907 125, 3545–3559. 10.1172/JCI82478.
908 13. Zhu, L., Olsen, R.J., Nasser, W., De La Riva Morales, I., and Musser, J.M. (2015).
909 Trading Capsule for Increased Cytotoxin Production: Contribution to Virulence
910 of a Newly Emerged Clade of *emm*89 *Streptococcus pyogenes*. mBio 6, e01378–
911 e01315. 10.1128/mbio.01378-15.
912 14. Hirose, Y., Yamaguchi, M., Takemoto, N., Miyoshi-Akiyama, T., Sumitomo, T.,
913 Nakata, M., Ikebe, T., Hanada, T., Yamaguchi, T., Kawahara, R., et al. (2020).
914 Genetic Characterization of *Streptococcus pyogenes emm*89 Strains Isolated in
915 Japan From 2011 to 2019. Infect. Microbes Dis. 2, 160–166.
916 10.1097/IM9.0000000000000038.
917 15. Walker, M.J., Hollands, A., Sanderson-Smith, M.L., Cole, J.N., Kirk, J.K.,
918 Henningham, A., McArthur, J.D., Dinkla, K., Aziz, R.K., Kansal, R.G., et al.
919 (2007). DNase Sda1 provides selection pressure for a switch to invasive group A
920 streptococcal infection. Nat. Med. 13, 981–985. 10.1038/nm1612.
921 16. Sumby, P., Whitney, A.R., Graviss, E.A., DeLeo, F.R., and Musser, J.M. (2006).

922 Genome-wide analysis of group a streptococci reveals a mutation that modulates
923 global phenotype and disease specificity. PLoS Pathog. 2, e5.
924 10.1371/journal.ppat.0020005.
925 17. Cole, J.N., Barnett, T.C., Nizet, V., and Walker, M.J. (2011). Molecular insight
926 into invasive group A streptococcal disease. Nat. Rev. Microbiol. 9, 724–736.
927 10.1038/nrmicro2648.
928 18. Ikebe, T., Ato, M., Matsumura, T., Hasegawa, H., and Sata, T. (2010). Highly
929 Frequent Mutations in Negative Regulators of Multiple Virulence Genes in Group
930 A Streptococcal Toxic Shock Syndrome Isolates. PLOS Pathog. 6, e1000832.
931 10.1371/journal.ppat.1000832.
932 19. Ministry of Health, Labour and Welfare. Severe Invasive Group A Streptococcal
933 Infections. https://www.mhlw.go.jp/bunya/kenkou/kekkaku-kansenshou11/01-
934 05-06.html.
935 20. Centers for Disease Control and Prevention, C.f.D.C.a. Streptococcal Toxic Shock
936 Syndrome (STSS) (*Streptococcus pyogenes*) 2010 Case Definition.
937 https://ndc.services.cdc.gov/case-definitions/streptococcal-toxic-shock-
938 syndrome-2010/.
939 21. Ikebe, T., Murai, N., Endo, M., Okuno, R., Murayama, S., Saitoh, K., Yamai, S.,

940 Suzuki, R., Isobe, J., Tanaka, D., et al. (2003). Changing prevalent T serotypes
941 and *emm* genotypes of *Streptococcus pyogenes* isolates from streptococcal toxic
942 shock-like syndrome (TSLS) patients in Japan. Epidemiol. Infect. 130, 569–572.
943 22. Beall, B., Facklam, R.R., Elliott, J.A., Franklin, A.R., Hoenes, T., Jackson, D.,
944 Laclaire, L., Thompson, T., and Viswanathan, R. (1998). Streptococcal *emm* types
945 associated with T-agglutination types and the use of conserved *emm* gene
946 restriction fragment patterns for subtyping group A streptococci. J. Med.
947 Microbiol. *47,* 893–898. 10.1099/00222615-47-10-893.
948 23. Johnson, D.R., and Kaplan, E.L. (1993). A review of the correlation of T-
949 agglutination patterns and M-protein typing and opacity factor production in the
950 identification of group A streptococci. J. Med. Microbiol. 38, 311–315.
951 10.1099/00222615-38-5-311.
952 24. Katsukawa, C., Tamaru, A., Morikawa, Y., and Oda, K. (2002). M protein gene
953 (*emm*) typing of *Streptococcus pyogenes*. Kansenshogaku Zasshi 76, 238–245.
954 10.11150/kansenshogakuzasshi1970.76.238.
955 25. Beres, S.B., Olsen, R.J., Saavedra, M.O., Ure, R., Reynolds, A., Lindsay, D.S.J.,
956 Smith, A.J., and Musser, J.M. (2017). Genome sequence analysis of *emm*89
957 *Streptococcus pyogenes* strains causing infections in Scotland, 2010–2016. J. Med.

958 Microbiol. 66, 1765–1773. 10.1099/jmm.0.000622.

959 26. Chochua, S., Metcalf, B.J., Li, Z., Rivers, J., Mathis, S., Jackson, D., Gertz, R.E.,
960 Jr., Srinivasan, V., Lynfield, R., Van Beneden, C., et al. (2017). Population and
961 Whole Genome Sequence Based Characterization of Invasive Group A
962 Streptococci Recovered in the United States during 2015. mBio 8, e01422-17.
963 10.1128/mBio.01422-17.

964 27. Davies, M.R., McIntyre, L., Mutreja, A., Lacey, J.A., Lees, J.A., Towers, R.J.,
965 Smeesters, P.R., Frost, H.R., Price, D.J., G Holden, M.T., et al. (2019). Atlas of
966 group A streptococcal vaccine candidates compiled using large-scale comparative
967 genomics. Nat. Genet. 51, 1035–1043. 10.1038/s41588-019-0417-8.

968 28. Turner, C.E., Holden, M.T.G., Blane, B., Horner, C., Peacock, S.J., and
969 Sriskandan, S. (2019). The Emergence of Successful *Streptococcus pyogenes*
970 Lineages through Convergent Pathways of Capsule Loss and Recombination
971 Directing High Toxin Expression. mBio 10, 1–20. 10.1128/mBio.02521-19.

972 29. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O.,
973 Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly
974 accurate protein structure prediction with AlphaFold. Nature, 583–589.
975 10.1038/s41586-021-03819-2.

976 30. Jain, I., Danger, J.L., Burgess, C., Uppal, T., and Sumby, P. (2020). The group A
977 Streptococcus accessory protein RocA: regulatory activity, interacting partners
978 and influence on disease potential. Mol. Microbiol. 113, 190–207.
979 10.1111/mmi.14410.
980 31. Ribardo, D.A., and McIver, K.S. (2006). Defining the Mga regulon: Comparative
981 transcriptome analysis reveals both direct and indirect regulation by Mga in the
982 group A streptococcus. Mol. Microbiol. 62, 491–508. 10.1111/j.1365-
983 2958.2006.05381.x.
984 32. Kachroo, P., Eraso, J.M., Beres, S.B., Olsen, R.J., Zhu, L., Nasser, W., Bernard,
985 P.E., Cantu, C.C., Saavedra, M.O., Arredondo, M.J., et al. (2019). Integrated
986 analysis of population genomics, transcriptomics and virulence provides novel
987 insights into *Streptococcus pyogenes* pathogenesis. Nat. Genet. 51, 548–559.
988 10.1038/S41588-018-0343-1.
989 33. Vega, L.A., Malke, H., and McIver, K.S. (2016). Virulence-Related
990 Transcriptional Regulators of *Streptococcus pyogenes*. In *Streptococcus*
991 *pyogenes* : Basic Biology to Clinical Manifestations [Internet], J.J. Ferretti, D.L.
992 Stevens, and V.A. Fischetti, eds (University of Oklahoma Health Sciences Center).
993 34. Terao, Y., Mori, Y., Yamaguchi, M., Shimizu, Y., Ooe, K., Hamada, S., and

994 Kawabata, S. (2008). Group A streptococcal cysteine protease degrades C3 (C3b)
995 and contributes to evasion of innate immunity. J. Biol. Chem. 283, 6253–6260.
996 10.1074/jbc.M704821200.
997 35. Sumitomo, T., Nakata, M., Higashino, M., Jin, Y., Terao, Y., Fujinaga, Y., and
998 Kawabata, S. (2011). Streptolysin S contributes to group A streptococcal
999 translocation across an epithelial barrier. J. Biol. Chem. 286, 2750–2761.
1000 10.1074/jbc.M110.171504.
1001 36. Honda-Ogawa, M., Ogawa, T., Terao, Y., Sumitomo, T., Nakata, M., Ikebe, K.,
1002 Maeda, Y., and Kawabata, S. (2013). Cysteine proteinase from Streptococcus
1003 pyogenes enables evasion of innate immunity via degradation of complement
1004 factors. J. Biol. Chem. 288, 15854–15864. 10.1074/jbc.M113.469106.
1005 37. Lin, A., Loughman, J.A., Zinselmeyer, B.H., Miller, M.J., and Caparon, M.G.
1006 (2009). Streptolysin S inhibits neutrophil recruitment during the early stages of
1007 *Streptococcus pyogenes* infection. Infect. Immun. 77, 5190–5201.
1008 10.1128/IAI.00420-09.
1009 38. Cimen, M.Y. (2008). Free radical metabolism in human erythrocytes. Clin. Chim.
1010 Acta. 390, 1–11. 10.1016/j.cca.2007.12.025.
1011 39. Laabei, M., and Ermert, D. (2019). Catch Me if You Can: *Streptococcus pyogenes*

1012 Complement Evasion Strategies. J. Innate. Immun. 11, 3–12. 10.1159/000492944.

1013 40. He, L.Y., Yu, Y.B., Liu, Y., Le, Y.J., Li, S., and Yang, X.Y. (2022). Immunization
1014 with the lipoprotein FtsB stimulates protective immunity against *Streptococcus*
1015 *pyogenes* infection in mice. Front. Microbiol. 13, 969490.
1016 10.3389/fmicb.2022.969490.

1017 41. Nguyen, M.T., and Gotz, F. (2016). Lipoproteins of Gram-Positive Bacteria: Key
1018 Players in the Immune Response and Virulence. Microbiol. Mol. Biol. Rev. 80,
1019 891–903. 10.1128/MMBR.00028-16.

1020 42. Bessen, D.E., McShan, W.M., Nguyen, S.V., Shetty, A., Agrawal, S., and Tettelin,
1021 H. (2015). Molecular epidemiology and genomics of group A Streptococcus.
1022 Infect. Genet. Evol. 33, 393–418. 10.1016/j.meegid.2014.10.011.

1023 43. McShan, W.M., McCullor, K.A., and Nguyen, S.V. (2019). The Bacteriophages of
1024 *Streptococcus pyogenes*. Microbiol. Spectr. 7. 10.1128/microbiolspec.GPP3-
1025 0059-2018.

1026 44. Berbel, D., Camara, J., Gonzalez-Diaz, A., Cubero, M., Lopez de Egea, G., Marti,
1027 S., Tubau, F., Dominguez, M.A., and Ardanuy, C. (2021). Deciphering mobile
1028 genetic elements disseminating macrolide resistance in *Streptococcus pyogenes*
1029 over a 21 year period in Barcelona, Spain. J. Antimicrob. Chemother. 76, 1991–

1030 2003. 10.1093/jac/dkab130.

1031 45. Wajima, T., Chiba, N., Morozumi, M., Shouji, M., Sunaoshi, K., Sugita, K.,
1032 Tajima, T., Ubukata, K., and Group A Streptococcus Surveillance Study Group.
1033 (2014). Prevalence of macrolide resistance among group A streptococci isolated
1034 from pharyngotonsillitis. Microb. Drug. Resist. 20, 431–435.
1035 10.1089/mdr.2013.0213.

1036 46. Hasegawa, T., Hata, N., Matsui, H., Isaka, M., and Tatsuno, I. (2017).
1037 Characterisation of clinically isolated *Streptococcus pyogenes* from
1038 balanoposthitis patients, with special emphasis on emm89 isolates. J. Med.
1039 Microbiol. 66, 511–516. 10.1099/jmm.0.000460.

1040 47. Tatara, K., Gotoh, K., Okumiya, K., Teramachi, M., Ishimoto, K., Tanaka, Y.,
1041 Iwahashi, J., Shindou, S., Yamashita, Y., and Watanabe, H. (2020). Molecular
1042 epidemiology, antimicrobial susceptibility, and characterization of
1043 fluoroquinolone non-susceptible *Streptococcus pyogenes* in Japan. J. Infect.
1044 Chemother. 26, 280–284. 10.1016/j.jiac.2019.10.004.

1045 48. Tanaka, Y., Gotoh, K., Teramachi, M., Ishimoto, K., Tsumura, N., Shindou, S., and
1046 Yamashita, Y. (2016). Molecular epidemiology, antimicrobial susceptibility, and
1047 characterization of macrolide-resistant *Streptococcus pyogenes* in Japan. J. Infect.

1048 Chemother. 22, 727–732. 10.1016/j.jiac.2016.06.013.

1049 49. Ubukata, K., Wajima, T., Morozumi, M., Sakuma, M., Tajima, T., Matsubara, K.,
1050 Itahashi, K., and Iwata, S. (2020). Changes in epidemiologic characteristics and
1051 antimicrobial resistance of *Streptococcus pyogenes* isolated over 10 years from
1052 Japanese children with pharyngotonsillitis. J. Med. Microbiol. 69, 443–450.
1053 10.1099/jmm.0.001158.

1054 50. Jolley, K.A., Bray, J.E., and Maiden, M.C.J. (2018). Open-access bacterial
1055 population genomics: BIGSdb software, the PubMLST.org website and their
1056 applications. Wellcome Open Res. 3, 124. 10.12688/wellcomeopenres.14826.1.

1057 51. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one
1058 FASTQ preprocessor. Bioinformatics 34, i884–i890.
1059 10.1093/bioinformatics/bty560.

1060 52. Kapatai, G., Coelho, J., Platt, S., and Chalker, V.J. (2017). Whole genome
1061 sequencing of group A Streptococcus: development and evaluation of an
1062 automated pipeline for *emm* gene typing. PeerJ 5, e3226. 10.7717/peerj.3226.

1063 53. Souvorov, A., Agarwala, R., and Lipman, D.J. (2018). SKESA: strategic k-mer
1064 extension for scrupulous assemblies. Genome Biol. 19, 153. 10.1186/s13059-018-
1065 1540-z.

1066 54. Jolley, K.A., and Maiden, M.C. (2010). BIGSdb: scalable analysis of bacterial
1067 genome variation at the population level. BMC Bioinformatics 11, 595.
1068 10.1186/1471-2105-11-595.
1069 55. Maiden, M.C.J., Jansen Van Rensburg, M.J., Bray, J.E., Earle, S.G., Ford, S.A.,
1070 Jolley, K.A., and McCarthy, N.D. (2013). MLST revisited: the gene-by-gene
1071 approach to bacterial genomics Europe PMC Funders Group. Nat. Rev. Microbiol.
1072 11, 728–736. 10.1038/nrmicro3093.
1073 56. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation.
1074 Bioinformatics 30, 2068–2069. 10.1093/bioinformatics/btu153.
1075 57. Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.,
1076 Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-
1077 scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693.
1078 10.1093/bioinformatics/btv421.
1079 58. Jaillard, M., Lima, L., Tournoud, M., Mahé, P., van Belkum, A., Lacroix, V., and
1080 Jacob, L. (2018). A fast and agnostic method for bacterial genome-wide
1081 association studies: Bridging the gap between k-mers and genetic events. PLOS
1082 Genet. 14, e1007758. 10.1371/journal.pgen.1007758.
1083 59. Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE:

1084 a fast and effective stochastic algorithm for estimating maximum-likelihood
1085 phylogenies. Mol. Biol. Evol. 32, 268–274. 10.1093/molbev/msu300.
1086 60. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin,
1087 L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic
1088 estimates. Nat. Methods 14, 587–589. 10.1038/nmeth.4285.
1089 61. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool
1090 for phylogenetic tree display and annotation. Nucleic Acids Res. 49, W293–W296.
1091 10.1093/nar/gkab301.
1092 62. R Core Team. (2022). A language and environment for statistical computing.
1093 https://www.R-project.org/.
1094 63. Lees, J.A., Galardini, M., Bentley, S.D., Weiser, J.N., and Corander, J. (2018).
1095 pyseer: A comprehensive tool for microbial pangenome-wide association studies.
1096 Bioinformatics 34, 4310–4312. 10.1093/bioinformatics/bty539.
1097 64. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren,
1098 S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance
1099 estimation using MinHash. Genome Biol. 17, 132. 10.1186/s13059-016-0997-x.
1100 65. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis, 2 Edition
1101 (Springer-Verlag New York). 10.1007/978-3-319-24277-4.

1102 66. Beres, S.B., Kachroo, P., Nasser, W., Olsen, R.J., Zhu, L., Flores, A.R., de la Riva,
1103 I., Paez-Mayorga, J., Jimenez, F.E., Cantu, C., et al. (2016). Transcriptome
1104 remodeling contributes to epidemic disease caused by the human pathogen
1105 *Streptococcus pyogenes*. mBio 7, 1–14. 10.1128/mBio.00403-16.
1106 67. Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular
1107 Biology Open Software Suite. Trends Genet. 16, 276–277. 10.1016/s0168-
1108 9525(00)02024-2.
1109 68. Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). lDDT: a local
1110 superposition-free score for comparing protein structures and models using
1111 distance difference tests. Bioinformatics 29, 2722–2728.
1112 10.1093/bioinformatics/btt473.
1113 69. Yin, R., Feng, B.Y., Varshney, A., and Pierce, B.G. (2022). Benchmarking
1114 AlphaFold for protein complex modeling reveals accuracy determinants. Protein
1115 Sci. 31, e4379. 10.1002/pro.4379.
1116 70. Hirokawa, T., Boon-Chieng, S., and Mitaku, S. (1998). SOSUI: classification and
1117 secondary structure prediction system for membrane proteins. Bioinformatics 14,
1118 378–379. 10.1093/bioinformatics/14.4.378.
1119 71. Takamatsu, D., Osaki, M., and Sekizaki, T. (2001). Thermosensitive suicide

1120 vectors for gene replacement in *Streptococcus suis*. Plasmid 46, 140–148.
1121 10.1006/PLAS.2001.1532.
1122 72. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general
1123 purpose program for assigning sequence reads to genomic features.
1124 Bioinformatics 30, 923–930. 10.1093/bioinformatics/btt656.
1125 73. Ge, S.X., Son, E.W., and Yao, R. (2018). iDEP: an integrated web application for
1126 differential expression and pathway analysis of RNA-Seq data. BMC
1127 Bioinformatics 19, 534. 10.1186/s12859-018-2486-6.
1128 74. Lancefield, R.C. (1957). Differentiation of group A streptococci with a common
1129 R antigen into three serological types, with special reference to the bactericidal
1130 test. J. Exp. Med. 106, 525–544. 10.1084/jem.106.4.525.
1131 75. Yamaguchi, M., Goto, K., Hirose, Y., Yamaguchi, Y., Sumitomo, T., Nakata, M.,
1132 Nakano, K., and Kawabata, S. (2019). Identification of evolutionarily conserved
1133 virulence factor by selective pressure analysis of *Streptococcus pneumoniae*.
1134 Commun. Biol. *2*, 96. 10.1038/s42003-019-0340-7.
1135 76. Takemura, M., Yamaguchi, M., Kobayashi, M., Sumitomo, T., Hirose, Y., Okuzaki,
1136 D., Ono, M., Motooka, D., Goto, K., Nakata, M., et al. (2022). Pneumococcal
1137 BgaA Promotes Host Organ Bleeding and Coagulation in a Mouse Sepsis Model.

1138 Front. Cell. Infect. Microbiol. 12, 844000. 10.3389/fcimb.2022.844000

1139

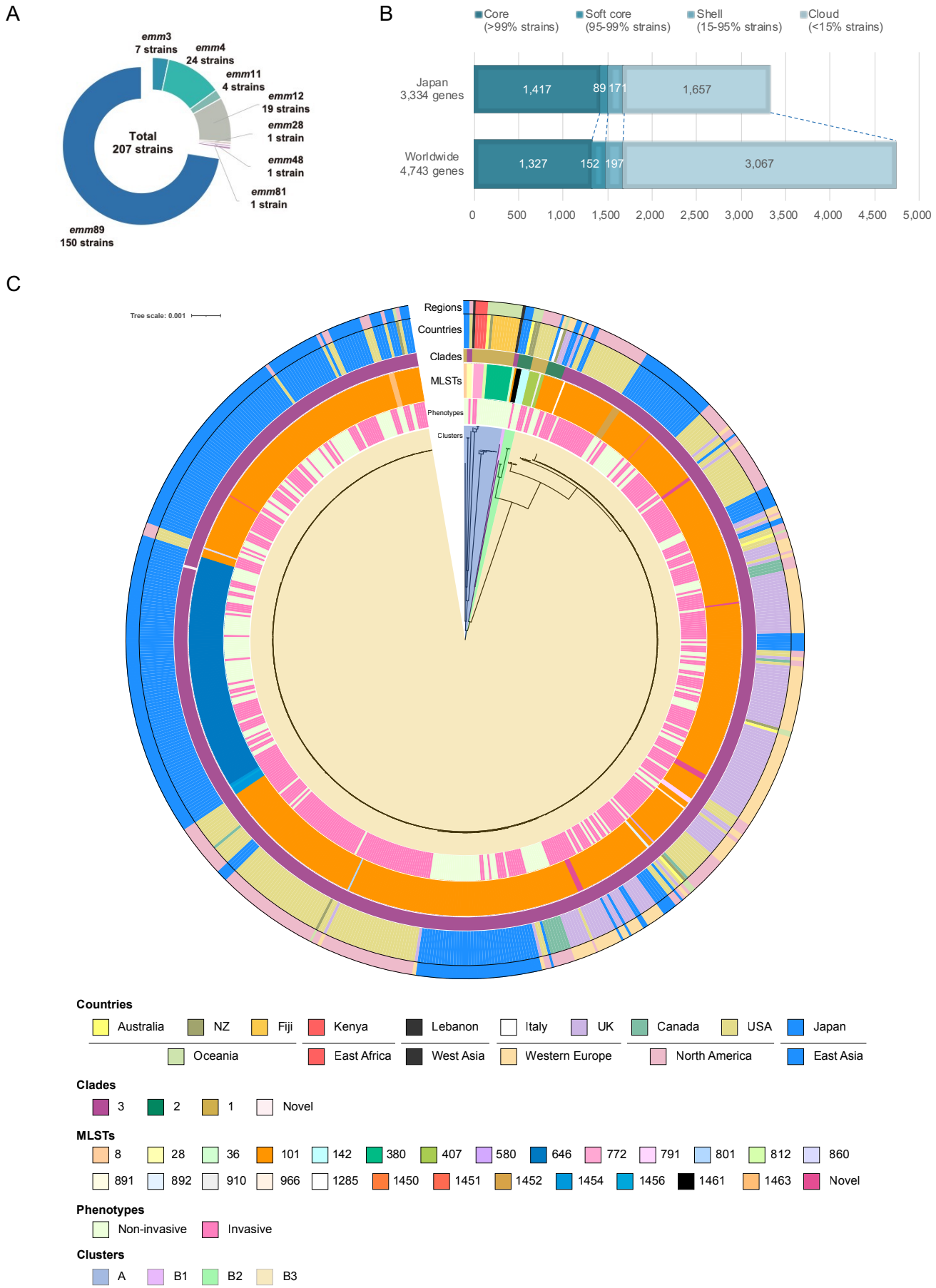
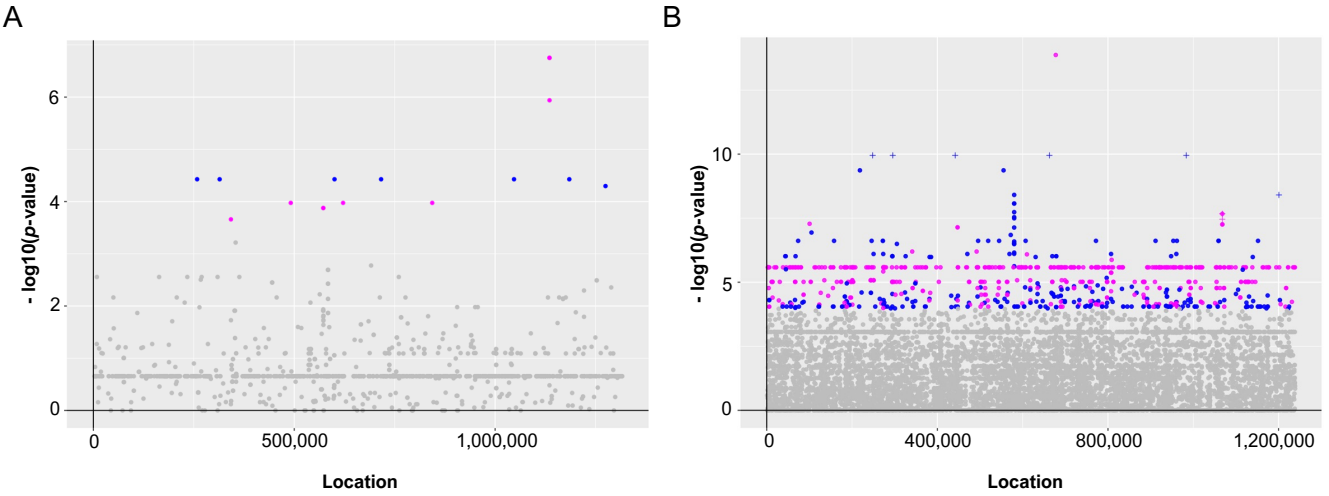


Fig. 1. Pan-genomic and phylogenetic characterization of the collected *emm*89 *S. pyogenes* isolates.

(A) *emm* genotyping of the 207 clinical isolates collected in Japan. (B) Pan-genome analysis of the Japanese and global cohorts. All genes detected in each cohort were classified into four groups, according to prevalence: core, soft-core, shell, and cloud genes. (C) Phylogenetic tree for the global cohort, based on the sequences of the core genes. From the inside, the color bars show clusters, phenotypes, MLSTs, clades, countries where strains were isolated, and global regions, in the order mentioned. The root of the tree was set as the mid-point. The scale located upper left indicates 0.001 times substitution of the bases on average.

MLST, multilocus sequence typing



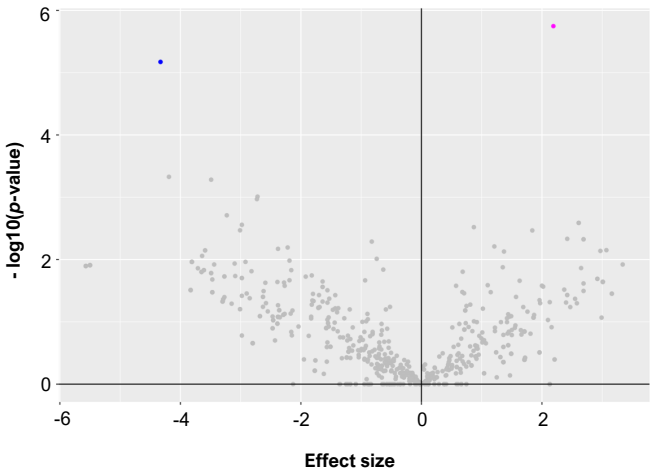
C

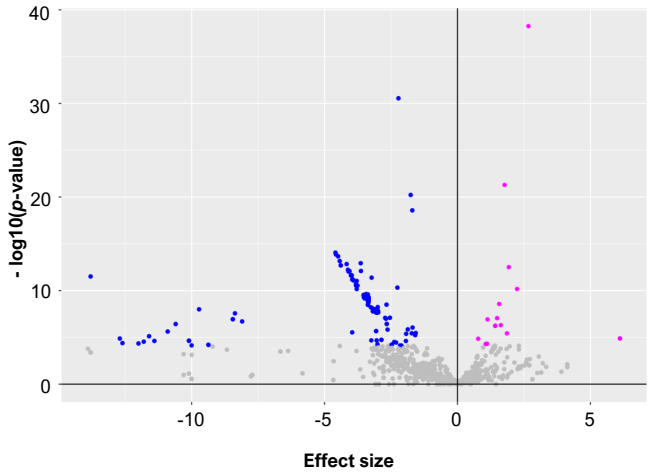
Genes	Position	SNP	*P*-value	Effect size
lacE	677614	G_T	1.35E-14	3.47E+00
gatA	248353	G_A	1.12E-10	-3.35E+00
group_1009	295467	G_A	1.12E-10	-3.35E+00
group_274	441823	C_A	1.12E-10	-3.35E+00
iscS_1	662673	C_A	1.12E-10	-3.35E+00
recU	983611	C_T	1.12E-10	-3.35E+00
fhuD	218429	G_A	4.30E-10	-1.84E+00
group_4706	555162	A_G	4.30E-10	-1.84E+00
group_589	580130	A_G	3.91E-09	-1.68E+00
fhuB	1200862	T_C	3.93E-09	-3.27E+00
group_589	580091	A_G	8.49E-09	-1.56E+00
	580123	G_T	8.49E-09	-1.56E+00
	580128	G_A	8.49E-09	-1.56E+00
	580108	T_*	1.84E-08	-1.54E+00
covS	1068053	T_*	2.14E-08	3.80E+00
	1068054	T_*	2.14E-08	3.80E+00
	1068060	C_*	2.14E-08	3.80E+00
	1068070	T_*	2.14E-08	3.80E+00
group_589	580129	G_A	2.89E-08	-1.58E+00
	580138	A_T	3.20E-08	-1.59E+00

Fig. 2. Genome-wide association study on SNPs/indels.

(A–B) Manhattan plots for the Japanese (A) and global (B) cohorts. The X-axis shows the location of each SNP/indel on the core gene alignment, while the Y-axis indicates the *p*-value. For each cohort, a permutation test was performed by iterating the calculations 1,000 times with randomly permuted phenotypes, with the significance level set at the 5th percentile of the 1,000 minimal *p*-values ($p=5.75\times10^{-4}$ and $p=1.05\times10^{-4}$ for the Japanese and global cohorts, respectively). Plots with lower *p*-values than genome-wide significant levels are colored magenta and blue, based on the direction of their effect size (positive and negative, respectively). (C) Distribution heatmap for the global cohort, with the strains possessing the significant SNPs/indels colored orange. Only the 20 SNPs with the lowest *p*-values are shown in this heatmap. The color bars above indicate countries and phenotypes. The position is the location of each SNP/indel on the core gene alignment.

SNP, single-nucleotide polymorphism; indel, insertion/deletion





C

Genes	*P*-value
group_829	6.41E-35
group_141	8.70E-23
group_468	2.74E-18
group_467	6.46E-13
group_139	8.19E-12
group_2718	6.59E-07
group_2687	1.67E-06
group_2464	2.05E-06
group_2689	4.01E-06
group_771	2.97E-03
rhaR_3	4.01E-06
group_1833	9.13E-06
group_2688	9.13E-06
group_75	1.04E-14
group_2692	1.98E-05
group_1660	1.69E-02
group_1830	9.13E-06
group_2696	1.98E-05
group_1829	4.09E-05
ykoD_2	4.09E-05

Fig. 3. Genome-wide association study on gene presence.

(A–B) Volcano plots for the Japanese (A) and global (B) cohorts. The X-axis shows the effect size, while the Y-axis indicates the *p*-value. Plots with a lower *p*-value than the genome-wide significant levels ($p=1.09\times10^{-4}$ and $p=7.72\times10^{-5}$) have been colored magenta and blue, based on the direction of their effect size (positive and negative, respectively). (C) Distribution heatmap for the global cohort, with the strains possessing the significant genes colored orange. Only the 20 genes with the lowest *p*-values are shown in this heatmap. The color bars above indicate countries and phenotypes.

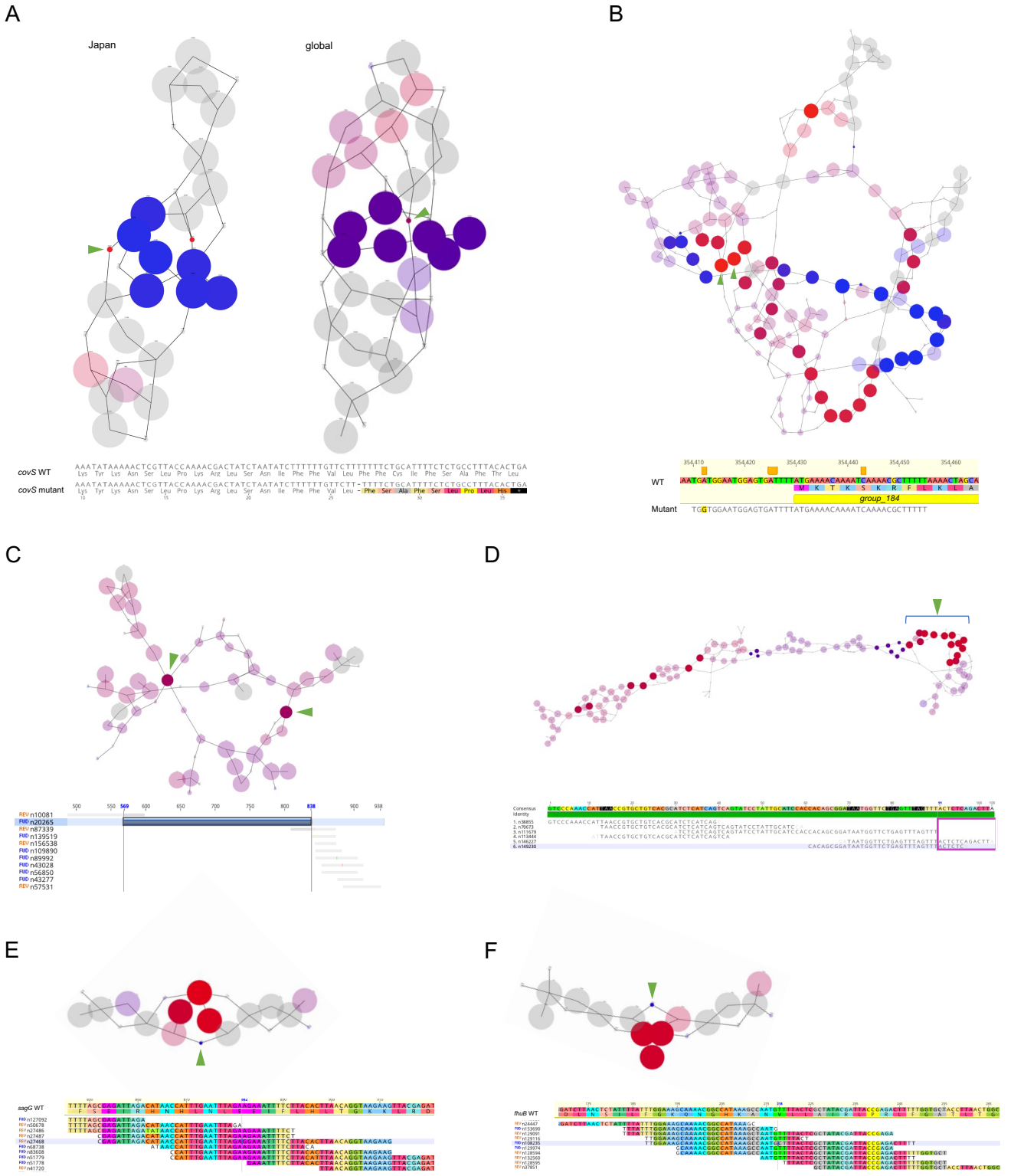


Fig. 4. K-mers related to pathology.

Detailed results of the GWAS on k-mers for the Japanese cohort in two genomic regions: *covS* (A, top left) and *group_184* (B); and the global cohort in five regions: *covS* (A, top right), an intergenic region (C), *group_141-143* (D), *sagG* (E), and *fhuB* (F). (A–F top) de-Bruijn graphs generated using DBGWAS. The respective nodes in the graphs indicate k-mers, with the significant nodes indicated using green arrowheads. The size of each node corresponds to the allele frequency. (A–F, bottom) Alignments of k-mers or maps of k-mers on the genome sequence of MGAS27061 around regions including significant mutations. Each base is colored according to the interpreted amino acids. (A) The significant k-mers pointed by arrowheads cause a frameshift mutation, truncating CovS protein to 35 amino acids. (B) The significant k-mers indicate that the presence of a sequence mapped on the first 26 bp of group_184 and its upstream 20 bp was related to the pathology. (C) Two significant k-mers indicate the same intergenic region of 270 bps. (D) The significant k-mers are indicated using arrowheads.

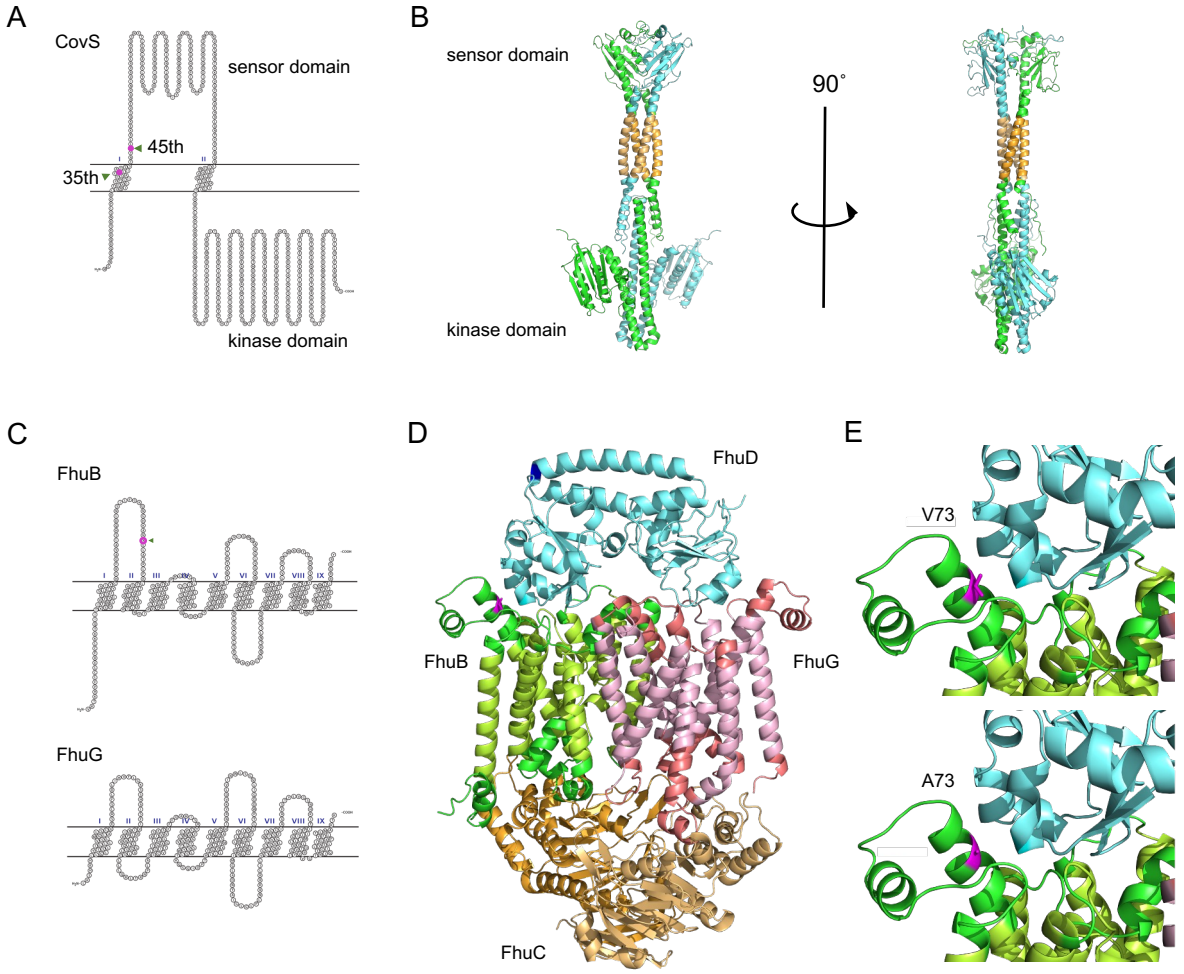


Fig. 5. Predicted protein structure models.

(A) Snake-like plot of transmembrane regions of CovS estimated using SOSUI. The frameshift mutations detected by the SNPs/indel- and k-mer-based GWASes cause truncation of CovS at the indicated 35th and 45th residues, respectively. (B) Structural model of the CovS homodimer (ipTM + pTM=0.614). The putative transmembrane regions are colored in orange. The upper part is the sensor domain, while the lower is the C-terminal kinase domain involved in phosphorylation of the transcriptional regulator CovR. (C) Snake-like plot depicting the transmembrane regions of FhuB and FhuG. The 73rd residue of FhuB has been indicated with an arrowhead and in magenta. (D) Structural model of the FhuBDCCG complex (ipTM + pTM=0.791). The putative transmembrane regions of FhuB and FhuG are colored in green-yellow and peach, respectively. The upper part of the model is located in the extracellular region, while the lower part is in the cytoplasm. (E) The 73rd valine in FhuB, colored in magenta, was substituted with alanine.

ipTM + pTM: Weighted combination of interface-predicted TM and predicted TM scores. ipTM is used to measure structural accuracy in the protein-protein interface, while ipTM is a metric for overall topological accuracy.

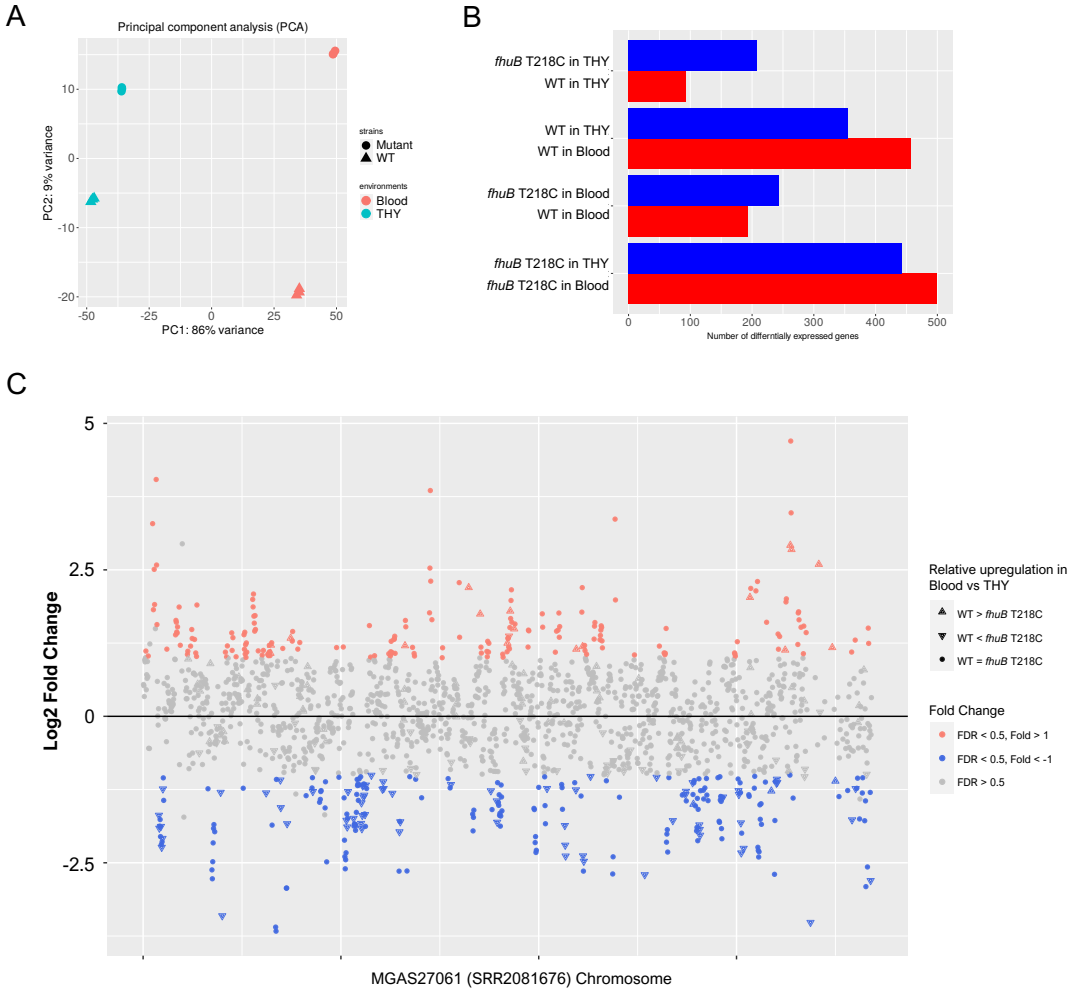


Fig. 6. Transcriptome analysis of the fhuB T218 mutant strain in THY and human blood.

(A) Principal component analysis plot of RNA-seq data. (B) Differentially expressed genes in four comparisons. (C) Plot of gene expression in the WT strain *vs. fhuB* T218C mutant in human blood. Significantly upregulated and downregulated genes in the WT have been colored red and blue, respectively. The shapes of the plots indicate relative transcriptional changes between THY and blood. Genes depicted with upward triangles were either significantly upregulated in the WT or downregulated in the mutant strain, in blood *vs.* THY. The downward triangle plots indicate genes that are either downregulated in the WT or upregulated in the mutant strain, in blood *vs.* THY.

WT, wild-type; THY, Todd Hewitt broth supplemented with 0.2% yeast extract

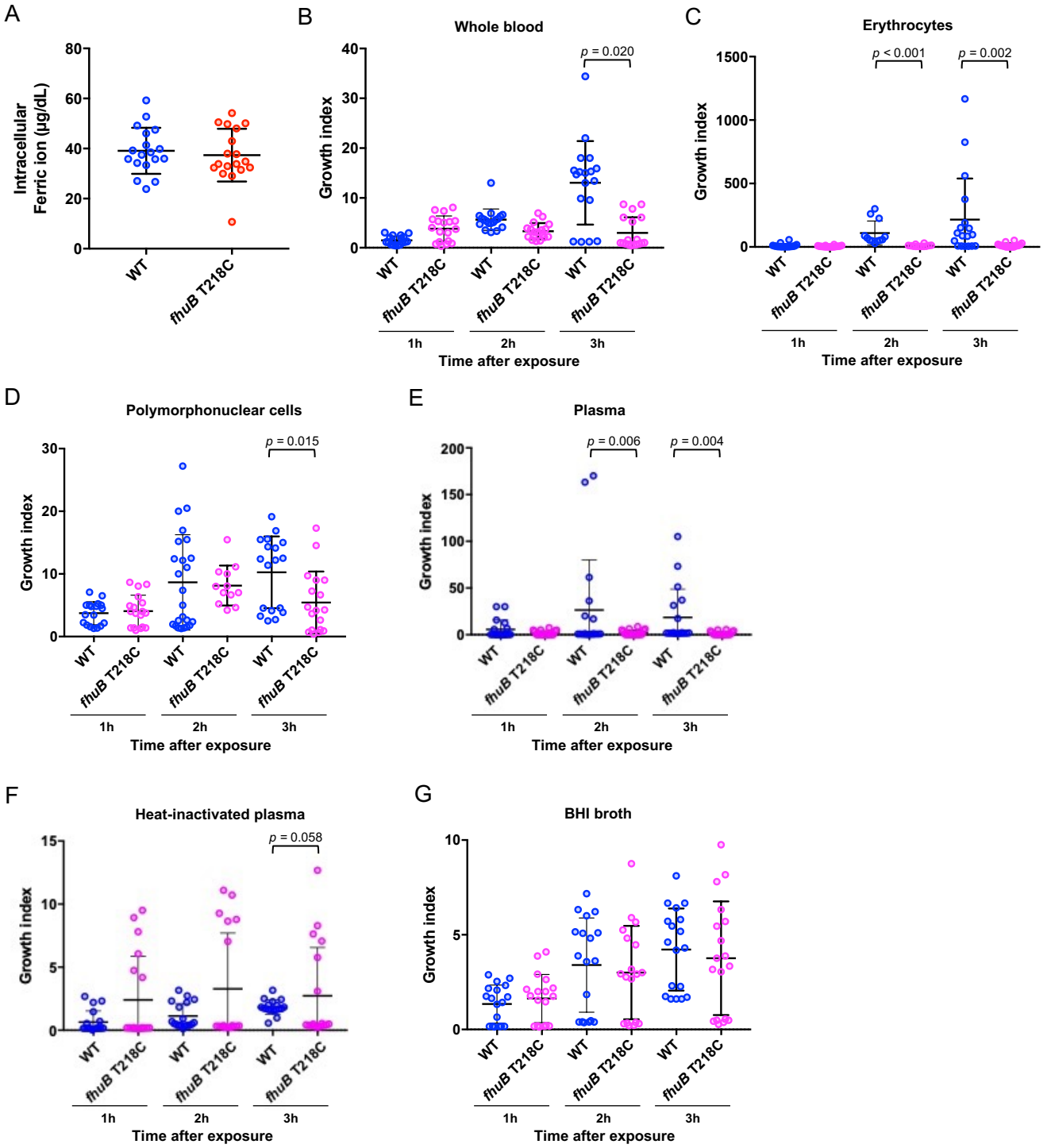


Fig. 7. Effects of the *fhuB* T218 mutant strain on ferric ion uptake and bacterial survival in human blood.

(A) Intracellular ferric ion assay. The WT and *fhuB* T218C mutant strains were incubated in healthy human blood for 3 h, following which the intracellular ferric ion concentrations were measured (n=18 technical replicates). (B) Bactericidal assay in human blood. The WT, *fhuB* T218C mutant, and Wr strains were mixed with healthy human blood to measure the ratio of bacterial counts at 1, 2, and 3 h to those at 0 h after infection (n=18 technical replicates). (C–F) Bactericidal assay in human erythrocyte-rich medium (C), polymorphonuclear cell-rich medium (D), plasma (E), and plasma inactivated by heating at 56ºC for 30 min (F) (n=18 technical replicates). (G) Bacterial growth in brain heart infusion broth (n=18 technical replicates). Representative data from one of three biological replicates are displayed. Thick bars and error bars indicate means and quartiles, respectively. Statistical significance was determined by means of the Mann–Whitney test with Benjamini–Hochberg's correction.