Identifying Genetic Variations in *emm*89 *Streptococcus pyogenes* Associated with Severe Invasive Infections

Masayuki Ono[1,2], Masaya Yamaguchi[1,2,3,4,5,20]*, Daisuke Motooka[3,6,7], Yujiro Hirose[1], Kotaro Higashi[1], Tomoko Sumitomo[1,8], Tohru Miyoshi-Akiyama[9], Rumi Okuno[10], Takahiro Yamaguchi[11], Ryuji Kawahara[11], Hitoshi Otsuka[12], Noriko Nakanishi[13], Yu Kazawa[14], Chikara Nakagawa[15], Ryo Yamaguchi[16], Hiroo Sakai[17], Yuko Matsumoto[18], Tadayoshi Ikebe[19], and Shigetada Kawabata[1,4]

1. Department of Microbiology, Graduate School of Dentistry, The University of Osaka, Osaka, Japan
2. Bioinformatics Research Unit, Graduate School of Dentistry, The University of Osaka, Osaka, Japan
3. Bioinformatics Center, Research Institute for Microbial Diseases, The University of Osaka, Osaka, Japan
4. Center for Infectious Diseases Education and Research, The University of Osaka, Osaka, Japan
5. Laboratory of Microbial Informatics, Microbial Research Center for Health and

19 Medicine, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka,
20 Japan
21 6. NGS Core Facility, Research Institute for Microbial Diseases, The University of
22 Osaka, Osaka, Japan
23 7. Integrated Frontier Research for Medical Science Division, Institute for Open and
24 Transdisciplinary Research Initiatives (OTRI), The University of Osaka, Osaka, Japan
25 8. Department of Oral Microbiology, Graduate School of Biomedical Sciences,
26 Tokushima University, Tokushima, Japan
27 9. Pathogenic Microbe Laboratory, Department of Infectious Diseases, National Center
28 for Global Health and Medicine, Tokyo, Japan
29 10. Department of Microbiology, Tokyo Metropolitan Institute of Public Health, Tokyo,
30 Japan
31 11. Department of Bacteriology, Osaka Institute of Public Health, Osaka, Japan
32 12. Department of Public Health Sciences, Yamaguchi Prefectural Institute of Public
33 Health and Environment, Yamaguchi, Japan
34 13. Department of Infectious Diseases, Kobe Institute of Health, Hyogo, Japan
35 14. Department of Microbiology, Fukushima Prefectural Institute for Public Health,
36 Fukushima, Japan

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

16. Sapporo Public Health Office, Hokkaido, Japan

17. Niigata City Institute of Public Health and the Environment, Niigata, Japan

18. Microbiological Testing and Research Division, Yokohama City Institute of Public Health, Kanagawa, Japan

19. Department of Bacteriology I, National Institute of Infectious Diseases, Tokyo, Japan

20. Lead contact.

*Correspondence: yamaguchi.masaya@nibn.go.jp

Abstract

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.

Introduction

Streptococcus pyogenes is a human-restricted gram-positive pathogen associated with a 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 STSS remains high, ranging from 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 a
77 hypervariable region sequence of the *emm* gene, which encodes the M protein. This
78 hypervariable region of the M protein is responsible for type-specific antigenicity and
79 binds with high affinity to C4b-binding protein, a major fluid phase inhibitor of the
80 classical and lectin pathways of the complement system that confers resistance to
81 opsonophagocytosis[8]. Since the mid to late 2000s, *emm*89 strains have been increasingly
82 isolated from samples obtained from patients with invasive diseases, becoming one of the

83 most frequently identified lineages in developed countries[9,10]. For example, in the United
84 Kingdom, no more than ten cases of invasive diseases caused by *emm*89 *S. pyogenes* were
85 reported annually in the 1990s. In the 2010s, however, approximately 150 cases were
86 reported out of the 1,000–1,500 invasive cases caused by all *emm* types of *S. pyogenes* in
87 each year[11]. Similarly, in Japan, in the 2010s, there was an increase in the number of
88 patients with *emm*89 *S. pyogenes*-induced severe invasive infections. In 2018, *emm*89
89 strains were isolated from 36% of patients diagnosed with *S. pyogenes*-induced severe
90 invasive infections, following *emm*1 strains, which were isolated from 57% of patients[9].

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

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

102 strains that determine their phenotypes.

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

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

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

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

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

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

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

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

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

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

113 are involved in the development of invasiveness.

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

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

116 *pyogenes*. Our pan-GWAS focused on the *emm*89 lineage of *S. pyogenes* to detect lineage-

117 specific factors and minimize false positives due to lineage differentiation. For a

118 comprehensive analysis, we constructed the core genome and pan-genome of *emm*89 *S.*

119 *pyogenes* strains and evaluated the effect of single-nucleotide polymorphisms (SNPs) on
120 core gene alignment and accessory clusters of orthologous genes (COGs). In addition, we
121 performed a k-mers-based GWAS to detect SNPs in the intergenic regions and multiple
122 mutations. We collected and sequenced *emm*89 clinical strains isolated in Japan during
123 2016–2021, in addition to public *emm*89 genome sequences. Using these sequences, we
124 investigated the bacterial factors associated with severe invasive infections in Japan and
125 globally, using pan-GWAS. Based on the bacterial protein structural predictions, we then
126 selected candidates with high potential relevance to the phenotype. Finally, we introduced
127 an SNP related to non-invasiveness into a clinical strain isolated from a severe invasive
128 infection and examined the alteration of the bacterial phenotype through an *ex vivo*
129 infection assay.
130

131 ## Results

132 **Collection of *emm*89 *S. pyogenes* clinical isolates in Japan and construction of**
133 **cohorts**

134 We collected clinical *S. pyogenes* strains isolated between 2016 and 2021 from patients
135 with non-invasive and severe invasive infections in Japan. The Ministry of Health, Labour,
136 and Welfare of Japan has defined the clinical criteria of severe invasive β-hemolytic

137 streptococcal infections as STSS, based on the STSS 2010 Case Definition of the Centers
138 for Disease Control and Prevention in the US, with minor modifications, including the
139 addition of encompassing symptoms in the central nervous system (Table S1)[19,20].

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

155 *emm*89 strains, including 135 severe invasive and 176 non-invasive isolates, as the
156 Japanese cohort.

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

165

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

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

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

191 seven loci that determine MLST, suggesting that both lineages have a genetically close
192 relationship. Eight novel MLSTs were determined (ST1450, 1451, 1452, 1454, 1455,
193 1456, 1461, and 1463) and 15 novel MLST strains were detected in the Japanese cohort
194 (Figure S1). Taken together, using phylogenetic approaches, we found that most strains
195 from Japan and countries in Europe and North America share genetically close
196 relationships, with only one unique lineage in Japan, ST646.
197

198 **Pan-GWASes detect SNPs/indels associated with invasiveness that are both common**
199 **and specific to Japan and other countries**

200 To discover all types of genetic variants in whole genes within *emm*89 *S. pyogenes*
201 associated with (severe) invasiveness, we applied pan-genome analysis and performed
202 three types of independent analyses, including pan-GWASes targeting SNPs in core genes
203 and the presence or absence of all genes, and a GWAS targeting other variants located in
204 intergenic regions spanning several nucleotides. The present pan-GWAS of bacteria
205 within a single *emm* type minimized lineage effects, thus reducing false positives.

206 We extracted SNPs and single-nucleotide indels from core gene alignments and
207 detected 24,627 and 47,060 SNPs/indels in the Japanese and global cohorts, respectively.
208 Subsequent pan-GWASes identified SNPs/indels associated with severe invasiveness in

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

223 We also performed a pan-GWAS for the global cohort and detected 1,075
224 SNPs/indels significantly related to invasive infections among the 360 core genes (Figure
225 2B, 2C and Table S6). Among the significant SNPs/indels, 725 caused synonymous
226 substitutions and 319 caused non-synonymous substitutions or frameshift mutations.

227 Moreover, 19 SNPs induced nonsense mutations, whereas the effects of 12 SNPs/indels
228 were unpredictable because of a lack of reference sequences (Table S3). Notably, 96
229 SNPs/indels accumulated in a single gene, *murJ*, which is involved in peptidoglycan
230 biosynthesis, whereas 53 and 51 SNPs/indels were detected in *murE* and *group_1008*,
231 respectively (Figure S3C). The SNP with the lowest *p*-value (p=1.35×10^{-14}) was *lacE,*
232 which encodes the EIICB component of the lactose-specific phosphotransferase system
233 (Figure 2C). This mutation was found to be associated with an invasive phenotype and
234 mainly observed in strains isolated in the US. Compared with the significant 17
235 SNPs/indels in the Japanese cohort, 10 SNPs/indels were also detected in the global
236 cohort, including 4 SNDs in *covS* and 1 SNP each in 6 loci (Figure 2C). Deletions at the
237 *covS* locus were common among strains from several countries, including Japan. In
238 contrast, SNPs in six loci, *gatA, group_1102*, *group_647*, *iscS_1*, *recU*, and *fhuB*, were
239 present exclusively in Japan (Figure 2C and Table S6). These results suggest that several
240 bacterial mechanisms cause severe invasive *S. pyogenes* infections, and some prevail
241 worldwide, such as *covS* mutations, whereas others are specific to Japan.
242
243 **Pan-GWAS on COGs reveals 2 and 109 genes associated with severe invasiveness in**
244 **the Japanese cohort and invasiveness in the global cohort, respectively**

245 Next, we examined the associations of accessory COGs with severe invasiveness
246 and global invasiveness in Japanese patients. A permutation test determined significance
247 levels as *p*-values of 1.09×10^{-4} and 7.72×10^{-5} for the Japanese and global cohorts,
248 respectively. Two significant genes were detected in the pan-GWAS for the Japanese
249 cohort: *group_184*, which encodes a hypothetical protein, and *divIC*, which encodes a
250 septum formation initiator protein ($p=8.81\times10^{-6}$ and $p=6.72\times10^{-6}$, respectively; Figure
251 3A and Table S7). Although analysis of the global cohort revealed the presence of 169
252 genes that were significantly related to invasiveness, no genes were identical or
253 homologous to the two genes detected in the Japanese cohort (Figure 3B, 3C, and Table
254 S8). Approximately 90% (152 of 169) of the significant COGs were found to be
255 associated with non-invasiveness. Among the 169 genes, 37 encoded phage-related genes
256 and 11 encoded mobile genetic elements (MGEs) such as transposase, integrase, and
257 recombinase. The gene with the lowest *p*-value was *group_829*, which encodes a
258 transposase and is related to the invasive phenotype (Figure 3C and Table S8). In addition,
259 *group_2689*, which encodes a multidrug efflux transporter permease, *rhaR_3*, which
260 encodes a transcriptional regulator involved in rhamnose metabolism, and *group_1829*,
261 which encodes a vitamin B12 import transporter permease, were also substantially
262 associated with the invasive phenotype. We identified several gene distribution patterns

263 associated with invasiveness, suggesting that multiple independent genetic factors cause
264 invasive infections (Figure 3C). On the other hand, the present pan-GWAS found no genes
265 encoding known virulence factors significantly associated with invasiveness, thus further
266 analysis of the relationships of detected distribution patterns with prophages and MGEs
267 was performed. For calculating the pairwise correlation of the presence of significant
268 COGs in the 666 strains, the COGs were clustered into eight coexisting groups, seven of
269 which contained phage- and/or MGEs-related genes (Figure S4). The largest group
270 comprised 65 genes including phage proteins, while the second largest with 42 genes was
271 found to be associated with non-invasive infections, and included *group_2689*,
272 *group_1833*, and *ermA1*, encoding TetR/AcrR family transcriptional regulator, multidrug
273 efflux system permease protein, and rRNA adenine N-6-methyltransferase, respectively.
274 Taken together, genes associated with invasiveness were found to encode mobile genetic
275 factors and transporters, whereas major virulence factors were not significantly associated
276 with invasiveness.
277

278 **K-mers-based GWAS detects both distinctive and identical variants compared to the**
279 **SNPs- and COGs-based pan-GWASes**

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

281 length k-mers from whole genomes and performed a GWAS. The k-mers-based GWAS
282 can handle polymorphisms spanning more than one base, such as indels, inversions, and
283 translocations, in both the coding and non-coding regions.

284 In the Japanese cohort, the k-mers-based GWAS detected two regions containing
285 causative variants associated with severe invasiveness (Table S9). As shown in the de-
286 Bruijn graphs, overlapping k-mers were concatenated into a single node, and edges
287 represent variability among the sequences (Figure 4A–F). The set of connected nodes and
288 edges comprising each de-Bruijn graph is called a complex. Nodes were determined to
289 be significant if their *q*-value was less than 0.05. Significant nodes are shown in red and
290 blue, indicating an association with severe invasive or non-invasive infections,
291 respectively. The complex comprising the nodes with the lowest *q*-value (q=1.49×10^{-2})
292 was Comp_11 in the *covS* locus (Table S9). The causative mutations in *covS* were an SND
293 and 10 nucleotide polymorphisms (Figure 4A). This deletion resulted in a frameshift
294 mutation that shortened the length of CovS from 500 to 35 amino acids, leading to
295 increased invasiveness, as previously reported in other *emm* types[17]. Another complex
296 significantly associated with severe invasiveness is Comp_2 (q=4.22×10^{-2}; Table S9).
297 Comp_2 is a highly variable region containing eight hypothetical protein-coding genes,
298 with high similarity within the first 75 bp. Significant k-mers were also mapped to the

299 first 26 bp of *group_184* and 20 bp upstream (Figure 4B). These findings suggest that
300 *group_184* possibly contributes to severe invasiveness through not only its presence but
301 also by that of the upstream region.

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

307 Two significant k-mers were present in Comp_6, which were found to be an
308 intergenic region of 270 bp (q=1.90×10^{-2}; Table S10 and Figure 4C). We speculated that
309 this region is related to regulation of gene expression. However, no promoter sequences
310 were identified by utilizing MLDSPP, BacPP, and BLAST, thus the significance of this
311 region remains to be clarified[29,30]. In Comp_24, with a high sequence variation containing
312 *group_141*, *group_142*, and *group_143*, which encode transposases, the presence of a
313 281-bp sequence consisting of several k-mers was significantly associated with the
314 invasive phenotype (q=1.90×10^{-2}; Table S10 and Figure 4D). Furthermore, *group_141*
315 was also identified in the COGs-based pan-GWAS as a non-invasiveness-related gene
316 along with *group_139* and *group_467*, which encode transposase and uncharacterized

317 protein, respectively (Table S8 and Figure S4). Taken together, the absence of an MGE
318 containing *group_141*, and the presence of another MGE harboring *group_142* and
319 *group_143* may result in an invasive phenotype. In Comp_10, n27458, the *sagG* locus
320 encoding the ATP-binding protein of the efflux transporter of SLS, was significantly
321 correlated with the non-invasive phenotype (q=2.40×10^{-2}; Table S10 and Figure 4E).
322 However, a significant SNP in n27458, *sagG* (A882G), was found to cause a synonymous
323 mutation. The other significant mutation was present in the *fhuB* locus, encoding a
324 putative ferrichrome transport system permease (q=2.40×10^{-2}; Table S10). This mutation
325 was identical to SNP T218C detected in the SNP/indels-based pan-GWAS (Figure 4F).
326 Moreover, this mutation changes the 73rd residue from valine to alanine in FhuB, which
327 is a putative ferrichrome transport system permease.
328 Therefore, the k-mer approach identified multiple variants, including the mutation
329 identified in the SNPs/indels-based pan-GWAS, *fhu*B SNP T218C. In addition, while the
330 mutation detected in *covS* differed from that detected in the SNPs/indels-based pan-
331 GWAS, both caused frameshift mutations.
332
333 **AlphaFold-based prediction of the impact of the identified mutations on function**
334 To assess the impact of mutations on protein function, we predicted the protein structure

335 using AlphaFold[31]. Here, we present structural predictions for three representative
336 proteins: LacE, whose mutation was observed mainly in invasive strains from the US;
337 CovS, whose invasion-related deletions prevailed worldwide; and FhuB, which carries a
338 prominent mutation in the Japanese cohort and is associated with non-invasiveness.

339 The invasive-related SNP in *lacE* substitutes the 554th glycine in LacE with valine.
340 LacE was predicted to be a membrane protein with nine transmembrane regions (Figure
341 S5A). LacE is the EIICB component of the lactose-specific phosphotransferase system,
342 and the EIIC and EIIB domains correspond to the transmembrane and intracellular
343 regions, respectively. As shown in magenta in the model, the 554th residue is predicted to
344 be in the intracellular EIIB domain (Figure S5B).

345 Next, we predicted a homodimerized CovS model using AlphaFold, as the CovS
346 of *S. pyogenes* forms homodimers[32]. SOSUI predicted that CovS has two transmembrane
347 regions (Figure 5A). Mutations detected in the SNP/indels- and k-mers-based GWASes
348 were predicted to shorten the CovS protein to 35 and 45 amino acids, respectively. As the
349 intracellular domain of CovS is in the C-terminal region and is involved in the
350 phosphorylation of the transcriptional regulator CovR, frameshift mutations leading to
351 CovS truncation would inactivate the protein, and thus, CovR function (Figure 5B).

352 As described above, the SNP T218C in the *fhuB* locus substitutes the 73rd valine of

353 FhuB with alanine. FhuB is a component of an ATP-binding cassette transporter system
354 that utilizes ferrichrome, which is one of the carriers of Fe^{3+}. FhuB is predicted to localize
355 to the cell membrane and form a channel with FhuG (Figure 5C). SOSUI suggested that
356 FhuB and FhuG are 9-transmembrane proteins. The FhuBG complex can bind to one
357 molecule of the extracellular ferrichrome-binding lipoprotein, FhuD, and two molecules
358 of the intracellular ATP-binding protein, FhuC. Therefore, we constructed a structural
359 model of the FhuBCCDG complex (Figure 5D), which implied that the 73rd residue of
360 FhuB exists in a region adjacent to FhuD. The mutation was predicted to induce formation
361 of a small indentation on the molecular surface, thus increasing the surface area accessible
362 to the solvent, and is considered to potentially affect the stability of the hydrophobic bond
363 between FhuB and FhuD, and thus ferrichrome transport (Figure 5E). The SNP G538A
364 in *fhuD* was Japan-specific, significantly related to severe invasiveness, and caused the
365 V180I mutation in FhuD (Figure S3C). The prediction suggested that the 180th residue is
366 in an α-helix distant from the active site or the interactive sites with FhuB. Both valine
367 and isoleucine are branched-chain amino acids, and the amino acid residue is located on
368 a rigid structure, the α-helix (Figure 5D). Thus, we believe that this mutation is less likely
369 to cause structural changes than the other non-synonymous mutations.
370

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

373 Based on the pan-GWAS and GWAS results and predicted protein structures, we focused
374 on the SNP *fhuB* T218C. We constructed a mutant strain, in which the SNP *fhuB* T218C
375 was introduced, to further investigate its potential virulence. We selected the strain TK02
376 which carries the wild-type (WT) allele T218 in *fhuB* and was originally isolated from a
377 sample obtained from a patient with severe invasive infection in Japan. We used a several
378 times-passaged TK02 strain, TK02', as a WT strain and introduced the SNP *fhuB* T218C
379 into it *via* allelic exchange mutagenesis with a thermo-sensitive shuttle vector. We then
380 confirmed that there were no differences between the WT and *fhuB* T218C strains using
381 whole-genome resequencing.

382 To reveal the effects of the SNP on invasiveness, we performed a transcriptomic
383 analysis of the WT and *fhuB* T218C strains in THY broth and human blood. Principal
384 component analysis revealed that the differences in the overall transcriptional profiles
385 between the strains were more remarkable in blood than in THY (Figure 6A, 6B, and
386 Tables S11–14). Among the overlapping genes, we found that the expression of CovR-
387 regulating genes, including *speB*, *nga-ifs-slo* operon, and *sag* operon, was upregulated in
388 the blood, compared to that in THY, in both the strains (Figure 6C, Tables S12, and S13).

389 CovR has also been reported to regulate *sda1*, which plays a key role in invasive disease
390 progression in *emm*1 strains; however, *sda1* does not exist in the *emm*89 reference strain
391 MGAS27061 or WT strain[15]. In human blood, the mutant strain resulted in the
392 downregulation of *mga* and *emm* expression and upregulation of the expression of genes
393 encoding virulence factors, such as *speC*, *scpA/B*, *endoS*, *fba*, *ska*, and *sfbX*, compared to
394 those observed in the WT strain (Table S14). Although Mga regulates the expression of
395 surface and secreted molecules important for colonization and immune evasion[33], no
396 strong expression changes were observed in the Mga regulon, except in the *emm* gene.
397 Notably, *fhuB*, *fhuC*, and *fhuD* were upregulated in the *fhuB* mutant in both human blood
398 and THY (Tables S11 and S14). Moreover, the transcriptomic analysis revealed that both
399 WT and *fhuB* T218C strains displayed upregulation of the *speB, nga-ifs-slo* operon, and
400 *sag* operons in human blood. In addition, the *fhuB* T218C mutation caused the
401 upregulation of *fhuBCD* expression. To determine whether *fhuB* T218C mutation affects
402 iron transport, we measured the intracellular free ferric ion concentration in each strain
403 and observed no differences among the strains in either environment (Figure 7A).
404 Next, to investigate the effects of SNP on bacterial survival in human blood, we
405 performed a bactericidal assay using human blood. At 2 and 3 h after incubation, the *fhuB*
406 T218C mutant strain exhibited a significantly decreased survival rate than that of the WT

407 strain (Figure 7B). To further determine the blood components that the attenuated survival
408 of the mutant can be attributed to, we compared bacterial survival rates in erythrocytes,
409 polymorphonuclear cells, plasma, heat-inactivated plasma, and brain heart infusion broth
410 (Figure 7C–7G). Notably, after incubation with erythrocytes, polymorphonuclear cells,
411 or plasma, the *fhuB* T218C mutant strain exhibited a significantly lower survival index
412 than that of the WT strain, at 2 and 3 h after incubation, as observed in whole blood
413 (Figure 7C–7E and 7G). However, there were no significant differences between the
414 survival rates of the WT and mutant strains in heat-inactivated plasma, suggesting that
415 the mutant strain is susceptible to complement (Figure 7F). Taken together, the
416 polymorphism T218C in *fhuB* impaired the survival of severe invasive strains in human
417 blood through interactions with erythrocytes.
418

419 **Discussion**

420 The present study focused on *emm*89 *S. pyogenes*, known to cause increasing rates of
421 invasive infections worldwide, and also assessed differences between *emm*89 strains
422 causing invasive and non-invasive infections. By focusing on bacteria within closed
423 phylogenies, false positives caused by confounding lineage effects were minimized, thus
424 contributing to a higher level of accuracy of the pan-GWAS. We constructed a workflow

425 for bacterial pan-GWAS and GWAS and explored the bacterial genetic factors related to
426 severe or invasive infections to reveal the plausible mechanisms of pathogenesis. We
427 independently performed pan-GWASes and GWASes for Japanese and global cohorts to
428 investigate the associations of variants with strictly defined phenotypes in severe invasive
429 *S. pyogenes* infections and with phenotypes in a broader context of invasive infections.
430 Several GWASes have been conducted on *S. pyogenes* to date. Davies *et al.* extracted k-
431 mers from the genomes of 1,944 strains of *S. pyogenes* with broad *emm* genotypes and
432 assessed their associations with invasiveness, and Kachroo *et al.* performed GWAS on k-
433 mers using 442 *emm*28 strains isolated from either invasive or non-invasive infections[27,34].
434 However, no studies have described the combined analysis of SNPs, genes, and k-mers
435 exclusively using *emm*89 strains that share genetically close genomes and, thus, a large
436 number of core genes. The investigations we performed led to the discovery of a
437 tremendous number of causative variants, including not only SNPs in core genes but also
438 accessory genes and insertions spanning intergenic regions.

439 Spontaneous mutations in *covR*/*S* genes potentiate the transition from localized to
440 invasive infection by M1T1 *S. pyogenes* [17]. We demonstrated that some SNDs in the *covS*
441 locus were significantly related to the phenotypes in both the SNPs/indels- and k-mers-
442 based GWASes. CovS forms a two-component system with CovR that regulates the

443 transcription of operons encoding virulence factors when phosphorylated by the histidine
444 sensor kinase CovS[35]. Sumby *et al.* suggested that systemic invasive infections are caused
445 by the overexpression of virulence factors through the inactivation of CovS, as a result of
446 the introduction of a depletive mutation in the *covS* locus after infection[15,16]. Several
447 mutation sites causing such overexpression, such as the 7 nt insertion in *covS* of *emm*1 *S.*
448 *pyogenes* and the insertion of adenine at the 877th nucleotide, have been reported[15,16].
449 Ikebe *et al.* demonstrated variations in the lengths of mutant CovS proteins by analyzing
450 the gene sequences of 164 STSS strains in Japan[18]. In line with their study, we found that
451 the shortening of CovS that potentially occurred owing to frameshifts caused by SNDs is
452 significantly associated with invasive infections, potentially contributing to the
453 upregulation of virulence factors and invasiveness.

454 In contrast, we observed several factors associated with invasiveness to be
455 independent of *covR*/*S* mutations. Based on conformational predictions, we selected SNPs
456 with non-synonymous substitutions that were likely to affect protein function.
457 Transcriptomic analysis findings suggested that the Japan-specific *fhuB* mutation
458 associated with non-severe invasive infections contributes to the growth rate of *S.*
459 *pyogenes* in human blood by adapting to the environment. In addition, *emm*89 clade 3
460 carries the identical promoter region pattern of the *nga* operon as *emm*1 strains, and the

461 pattern conferred similarly high expression of *nga* and *slo*[11,12,28]. Our RNA-seq data
462 demonstrated that a severe invasive Japanese strain without *covS* mutations increased the
463 expression of *speB*, *nga*, *slo*, and genes in the *sag* operon in the blood. Although *covR*/*S*
464 mutations downregulate the expression of SpeB and SLS, SpeB and SLS act as virulence
465 factors, allowing *S. pyogenes* to invade host tissues[36-39]. These data suggest that severe
466 invasive infections have multiple gene expression profiles in addition to *covR*/*S* mutation-
467 induced profiles within a single lineage, *emm*89, and that the synergy between optimizing
468 bacterial survival in human blood and upregulating multiple virulence factors contributes
469 to the development of severe invasive infections.

470 Two possible roles of the *FhuB* mutation in the pathogenesis of severe invasive
471 infections are thus proposed. First, the mutation in *FhuB* could increase bacterial
472 susceptibility to free radicals generated in the presence of ferric ions in the blood.
473 Generally, ferric ions are essential for the survival of almost all living organisms.
474 Catalyzed by ferric ions, the Fenton and Haber–Weiss reactions generate hydroxyl
475 radicals in erythrocytes[40]. Previously, we provided evidence that the iron in erythrocytes
476 partially inhibits pneumococcal growth *via* a free radical-based mechanism[7]. Although no
477 significant difference was observed in the intracellular ferric ion concentration in the
478 present study, the structural changes in FhuB may have caused an increased generation of

479 free radicals in bacterial cells and the prevention of bacterial survival in human blood.
480 Second, the *FhuB* mutation could affect bacterial transcriptional profiles and result in
481 reduced fitness for survival in human blood. We also observed a lower survival rate of the
482 mutant strain with polymorphonuclear cells (Figure 7D). Lower *emm* transcription in the
483 mutant may lead to attenuated immune evasion. The M protein can bind C4-binding
484 protein and factor H and inactivate the deposited C4b and C3b, leading to limited surface
485 opsonization[41]. Furthermore, complements could have inhibited the proliferation of the
486 mutant strain, as suggested by the activation of the membrane attack complex (Figure 7D
487 and 7E). Although the mechanisms of the *FhuB* V73A mutation in the interaction with
488 each blood component must be further elucidated, our results indicated that the 73rd
489 residue of FhuB is a key factor for bacterial survival in human blood during systemic
490 infection.

491 Recently, He *et al.* reported that antibodies against the ferrichrome-binding protein
492 FtsB, also known as FhuD, decrease blood bacterial burden and skin abscess formation
493 in murine models infected with *emm*1 *S. pyogenes*[42]. FhuD is a lipoprotein that has been
494 studied as a potential vaccine candidate against *Staphylococcus aureus*[43]. The detailed
495 function and effect of the FhuBCCDG complex in *S. pyogenes* remain unknown; however,
496 at least in *emm*1 strains, the complex can be expressed and localized abundantly on the

497 cell membrane. Hence, this complex may be a promising target for vaccines against *S.*
498 *pyogenes*.

499 Our gene-based pan-GWAS revealed no significant correlation with the
500 distribution of genes encoding virulence factors. This finding minimizes the possibility
501 that invasive infections are the result of the acquisition of virulence factors by non-
502 invasive strains and indirectly supports the hypothesis that changes in gene expression
503 profiles cause invasive infections. Although no vaccine has been commercialized against
504 *S. pyogenes*, given the possibility that the pathogen has multiple gene expression patterns,
505 even within a single lineage, it may be difficult to develop a universal vaccine with a
506 single antigen, and a vaccine containing multiple antigens may be effective.

507 In the global cohort, 169 genes were related to the phenotype, including 37 phage-
508 and 11 MGE-related genes. The genome of *S. pyogenes* is rich in prophages, phage-like
509 chromosomal islands, and MGEs[44]. Prophages and MGEs sometimes function as vectors
510 of virulence factors and antibiotic-resistance genes through interspecies and intraspecific
511 transmission. Common superantigen-coding genes, such as *speA*, *speC*, *speG*, *speH*, *speI*,
512 *speJ*, *speK*, and *speL*, and DNase-coding genes, such as *spd1*, are exogenous genes
513 derived from prophages[45]. In addition, in the genome of *S. pyogenes,* there are broad
514 MGEs where *erm* and *mefA* related to macrolide resistance and *tetM* involved in

515 resistance to tetracyclines are located[46]. In the present study, we observed no direct
516 association between invasiveness and virulence of genes located in the prophages and
517 MGEs. On the other hand, a cluster comprising 49 non-invasiveness-associated genes
518 including antibiotic-resistance genes was identified. Furthermore, among the genes
519 showing a significant correlation with the infectious phenotype, approximately 90% (152
520 of 169) were associated with non-invasiveness. One possible explanation is that
521 significantly related genes reflect the process of not only gain of factors but also loss of
522 those affecting fitness cost. Given that we also identified hypothetical genes associated
523 with this phenotype, we assumed that there is an unknown mechanism contributing to the
524 pathogenesis of invasive infections.

525 In the present study, MLST analysis revealed that 522 of the 666 strains belonged
526 to ST101. Following ST101, ST646 was the second most prevalent lineage in the
527 Japanese cohort and was not detected in any other country. In addition, the difference
528 between ST101 and ST646 was found only at the 295th nucleotide in *mulI*, among the
529 seven genes subjected to MLST typing. Our phylogenetic trees also suggested that ST101
530 and ST646 have a close phylogenetic relationship. Five previous studies have described
531 ST646 in *S. pyogenes*, and all of them reported *emm*89 ST646 strains isolated in Japan[47-
532 51]. Furthermore, all ST646 strains were isolated from the 4,249 *emm*89 *S. pyogenes* strains

533 in the MLST database PubMLST[52]. Therefore, ST646 appears to be specific to *emm*89
534 strains and a unique lineage in Japan. Notably, Ubukata *et al.* also reported that ST646
535 strains began increasing after 2012, whereas the ST101 lineage was dominant until 2008[51].
536 Taken together, we conclude that ST646 is possibly a relatively new and Japan-specific
537 lineage of *emm*89 *S. pyogenes.*

538 Our study had three limitations. First, because we sequenced bacterial genomes
539 using short-read sequencing, we could not detect large bacterial genome rearrangements
540 by comparing complete genome sequences. Thus, we were unable to investigate the
541 effects of long genomic structural dynamics, such as inversions, on pathogenesis. Second,
542 the clinical information associated with the isolates was limited; therefore, our analyses
543 did not reflect host information, such as age, sex, and underlying health conditions. Third,
544 we only used bacterial genetic distances based on core genome sequences as covariates
545 in the pan-GWAS. A combined GWAS of the host and pathogen in *S. pyogenes* infection
546 would highlight the relationship between host risk factors and bacterial genetic variants.

547 In this study, we revealed the genotype-phenotype associations found in not only
548 the known factors represented by *covS* but also factors that are related to invasiveness,
549 including *fhuB*. Moreover, we experimentally validated the contribution of the *fhuB*
550 mutation to bacterial survival in human blood. This study demonstrates the potential of

551 our genomic statistical approach for elucidating the pathogenesis of invasive infections.
552 Further analyses of the invasiveness-related factors identified in this study could provide
553 a platform for establishing novel treatments and preventive strategies against invasive
554 infections.
555

556 ## Materials and methods

557 **Clinical isolates in Japan**

558 Clinical isolates were collected from public health institutions in Tokyo, Osaka,
559 Yamaguchi, Fukushima, Kobe, Kyoto, Amagasaki, Sapporo, and Niigata, Japan. We
560 defined the strains collected as STSS according to the Infectious Diseases Control Law
561 in Japan. Non-invasive strains were defined based on diagnostic names, including
562 asymptomatic, pharyngitis, tonsillitis, or non-invasive infections. Strains with no
563 diagnostic names for the non-STSS strains were defined as non-invasive based on the
564 isolate sites. Information on all the strains included in this study is presented in Table S3.
565 The collected *S. pyogenes* strains were cultured at 37ºC in an atmosphere
566 containing 5% CO_2, in Todd Hewitt broth supplemented with 0.2% yeast extract (THY;
567 both from BD Biosciences, Franklin Lakes, NJ, USA) and stored in THY broth with 30%
568 glycerol (Nacalai Tesque, Kyoto, Japan), at –80°C.

569

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

571 The *S. pyogenes* strains were cultured until the exponential growth phase (OD_{600}=0.3–
572 0.4). Bacterial cells were lysed with T10E1N100 buffer (10 mM Tris-HCl buffer, 100
573 mM sodium chloride, and 1 mM EDTA), 10 units/mL mutanolysin (Sigma-Aldrich, St.
574 Louis, MO, USA), 10 mg/mL lysozyme (Fujifilm Wako Pure Chemical Co., Osaka,
575 Japan), 0.5 mg/mL achromopeptidase (Fujifilm Wako Pure Chemical Co.), and 0.3
576 mg/mL RNase A (Promega, Madison, WI, USA). Next, genomic DNA was extracted
577 from each lysate using a Maxwell® RSC instrument (Promega), according to the
578 manufacturer's instructions, and 250 bp paired-end libraries were then generated from the
579 extracted DNA using a Nextera XT DNA Kit (Illumina, San Diego, CA, USA). Libraries
580 were sequenced using a NovaSeq 6000 system (Illumina) at the Genome Information
581 Research Center, Research Institute for Microbial Diseases, The University of Osaka,
582 Osaka, Japan. On average, the number of reads was 5,433,301 (range 3,437,124–
583 9,117,301).

584

585 **Collection of published genome sequences**

586 We previously sequenced the draft genomes of 161 *emm*89 clinical isolates collected in

587 Japan between 2011 and 2019[14]. We defined strains derived from STSS as "severe
588 invasive," and those obtained from pharyngitis, tonsillitis, and superficial skin lesions as
589 "non-invasive" phenotypes.

590 To obtain public genome sequences of *emm*89 strains isolated from other countries,
591 we downloaded draft genome sequences in FASTA format from the National Center for
592 Biotechnology Information (NCBI) database, using Fasterq-dump v.2.9.6. The phenotype
593 of each strain, whether invasive or non-invasive, was defined according to the definitions
594 in the respective references that reported the strains[25-27].

595

596 **Genomic data processing and pan-genome analysis**

597 All processes and analyses were performed using the National Institute of Genetics (NIG)
598 supercomputer and SQUID at the D3 center of The University of Osaka (Osaka, Japan).
599 We constructed a workflow for bacterial pan-GWAS and other bioinformatic processes
600 (Figure S6). All collected sequences were subjected to quality checks using Fastp v.0.20.1,
601 with a default cutoff value of Q>15[53]. For newly isolated strains in Japan, *emm* typing
602 was performed using the emm-typing-tool v.0.0.1, and only sequences of strains
603 determined as *emm*89 were used for the following analyses[54]. All *emm*89 sequence data
604 were then subjected to *de novo* assembly using SKESA v.2.4.0, with default parameters[55].

605 Next, the MLST of each sequence was performed using MLST v.2.19.0[56,57]. Clade typing
606 was performed using BLAST v.2.13.0, with reference to the three *nga* promoter region
607 sequences[13]. After the genes were annotated with Prokka v.1.14.5, the pan-genome of all
608 sequences was calculated using Roary v.3.12.0, with the parameters "-e -mafft -r -qc -cd
609 99"[58,59]. Roary generated a core gene alignment and the distribution of all genes among
610 the strains. To extract SNPs/indels from core genes, including single-nucleotide indels,
611 snp-sites v.2.5.1 with the option "-v" was used. The output files were further processed
612 using BCFtools v.1.9, with the parameter "norm -m –", enabling analysis of multiple
613 alleles in the pan-GWAS. In parallel, k-mers were extracted using DBGWAS v.0.5.3, and
614 the length of k-mers was set as 31 nt using the "-k 31" parameter of DBGWAS[60].
615
616 **Phylogenetic analysis**
617 Phylogenetic relationships were calculated from the core gene alignment, using IQ-tree
618 v.1.16.12[61], based on maximum likelihood. The substitution model was automatically
619 selected considering the Akaike's and Bayesian information criteria by setting the "-m
620 MFP" parameter of IQ-tree[62]. Phylogenetic trees were constructed using iTOL v.6.6[63].
621 The similarity of clustering in the two phylogenetic trees was statistically examined using
622 Pearson's chi-square test with R v.4.0.3[64], followed by *post-hoc* analysis using residual

623 analysis adjusted with the Holm's method, if $p<0.05$.

624

625 **GWAS**

626 To investigate the associations between phenotypes and genotypes, including SNPs/indels
627 and genes, we performed a pan-GWAS using pyseer v.1.3.4[65]. The VCF file of
628 SNPs/indels or the gene distribution matrix was designated as the genotype. To remove
629 biases derived from lineages, we added information on the genetic distances between all
630 pairs of strains as covariates using mash v.2.3[66]. Briefly, *de novo* assemblies were
631 compressed through conversion into minimum hash values using the command "mash
632 sketch -s 10000." Subsequently, the commands "mash dist" and "square_mash" were
633 utilized to generate a genetic distance matrix expressed with Jaccard coefficients[66]. The
634 obtained matrix underwent eigenvalue decomposition, and the number of eigenvalues
635 used for multidimensional scaling was visually determined using the plot of the
636 relationships between the eigenvalues and contribution ratios. The number of eigenvalues
637 and the distance matrix were then added as pyseer parameters. The pyseer calculation was
638 iterated 1,000 times with randomized phenotypes, and the 5-percentile value of the
639 minimal *p*-value in each calculation was set at the significance level. Using R and the
640 package ggplot2, the results were visualized as a Manhattan plot for the SNPs/indels-

641 based pan-GWAS and a volcano plot for the genes-based pan-GWAS, respectively[67].
642 Heatmaps of the strains possessing significant variants were generated using Excel
643 (v.16.66.1; Microsoft, Redmond, WA, USA). The correlation of the presence of
644 significant COGs was calculated and visualized using the R program.

645 The k-mers-based GWAS was carried out using DBGWAS[60]. K-mers were
646 considered significant at a false discovery rate (*q*-value) of <0.05. DBGWAS-calculated
647 complexes, which are regions encompassing the k-mers, were significantly related to
648 pathology, and de-Bruijn graphs were generated based on these complexes. The
649 sequences of the k-mers were outputted and mapped to a reference sequence using
650 Geneious Prime v.2022.0.1 (Biomatters, Auckland, New Zealand) to identify the
651 mutations indicated by the k-mers. For the reference strain, we adopted MGAS27061,
652 which was isolated from an invasive case in the USA and whose complete chromosomal
653 sequence has been used as the reference sequence of *emm*89 clade 3[68]. Promoter
654 sequences in intergenic regions were predicted using web-based tools, MLDSPP and
655 BacPP [29,30]. Additionally, BLAST was employed to search the promoter sequences of *S.*
656 *pyogenes* strain SF370 registered in the CDBProm database
657 (https://aw.iimas.unam.mx/cdbprom/)[69].
658

659 **Protein structure prediction**

660 Significant variants found in the pan-GWAS, resulting in non-synonymous substitutions
661 in proteins, were searched by converting nucleotide sequences into amino acid sequences
662 using EMBOSS Transeq v.6.6.0.0[70]. To assess whether these mutations affected the
663 protein function, protein structure prediction models were constructed using AlphaFold
664 v.2.2.2[31]. The calculations were performed five times for each model, by setting the option
665 multimer_predictions_per_model=5. We predicted multimer models using the option --
666 model_preset="multimer" if a protein is reported or anticipated to form a multimer. For
667 each monomer, we selected the model with the best predicted local difference distance
668 test (an indicator of local structural accuracy) score[71]. For each multimer, AlphaFold was
669 calculated and expressed as a weighted combination of the interface-predicted TM and
670 predicted TM scores (ipTM + pTM). pTM is a metric for overall topological accuracy
671 and ipTM is used to measure the structural accuracy of the protein-protein interface[72].
672 The transmembrane regions of the proteins were predicted using SOSUI
673 (https://harrier.nagahama-i-bio.ac.jp/sosui/mobile/)[73]. The structures of the obtained
674 model were visualized using PyMOL v.2.5 (Schrödinger, LLC., New York, NY, USA).

675

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

677 We used the several times-passaged *S. pyogenes* TK02 strain, TK02', as the WT strain.
678 The TK02 strain was originally isolated from a patient with severe invasive infection[14].
679 The whole genome of TK02' was sequenced, and the generated fasta file is available in
680 Document S1. A point mutation, *fhuB* T218C, was introduced using the temperature-
681 sensitive shuttle vector, pSET4s, as reported previously[74]. Sanger sequencing confirmed
682 the presence of the point mutation. In addition, we resequenced the draft genome to
683 confirm that there were no differences apart from the point mutation, as described above.
684 The bacterial strains, primers, and plasmids used in this study are listed in Tables
685 S15 and S16. *Escherichia coli* strain DH5α (Takara Bio, Shiga, Japan) was used as a host
686 for the plasmid derivatives. All *E. coli* strains were cultured in Luria Bertani broth, at
687 37°C, with agitation. For selection and maintenance of mutants, antibiotics were added
688 to the media at the following concentrations: carbenicillin (Nacalai Tesque), 100 μg/mL
689 for *E. coli*; and spectinomycin (Fujifilm Wako Pure Chemical Corporation), 100 μg/mL
690 for *E. coli* and *S. pyogenes.*
691
692 **Transcriptomic analysis**
693 The *fhuB* WT and mutant strains were harvested in 30 mL of THY broth, of which 1 mL
694 was dispensed into 10 mL of THY and the remainder was centrifuged and resuspended in

695 2 mL of heparinized human blood. Bacterial mixtures with THY or blood were dispensed
696 into three aliquots (n=3 for each condition) and incubated at 37ºC for 3 h. THY samples
697 were centrifuged and resuspended in RNA Shield (Zymo Research, Irvine, CA, USA).
698 For each blood sample, 2 volumes (1 mL) of RNA protection bacteria reagent (Qiagen,
699 Hilden, Germany) were added. L5, included in the PureLink™ Total RNA Blood
700 Purification Kit (Thermo Fisher Scientific, Waltham, MA, USA), was added to remove
701 erythrocytes. The bacterial cell wall was mechanically lysed in Lysing Matrix B using a
702 MagNA Lyser (Roche, Basel, Switzerland). After centrifugation, the total bacterial RNA
703 was extracted using a Quick-RNA™ Miniprep Kit (Zymo Research), according to the
704 manufacturer's instructions. Full-length cDNA was generated using the SMART-Seq®
705 HT Kit (Takara Bio), according to the manufacturer's instructions. Pair-end libraries were
706 generated using a Nextera XT DNA Kit and sequenced using a NovaSeq 6000 system
707 (both from Illumina, San Diego, CA, USA). Sequenced data were preprocessed using
708 Trimmomatic v.0.33 and FastQC v.0.12.1. The reads were mapped to the complete
709 MGAS27061 genome (NCBI reference sequence: NZ_CP013840.1) using STAR
710 v.2.7.0a. After a second quality check using FastQC, read counting was performed using
711 featureCounts v.1.5.2[75]. Differentially expressed genes were identified using iDEP v.0.96
712 and gene annotations from NCBI and Prokka were combined[76]. Plots were created using

713 iDEP and the R package ggplot2.

714

715 **Intracellular ferric ion assay**

716 Human plasma was obtained through centrifugation of heparinized human blood, after 30
717 min of incubation at 37ºC. The WT and *fhuB* mutant strains were harvested at the
718 exponential phase, resuspended into 1 mL of THY or serum, and then incubated at 37ºC
719 for 3 h. Viable bacterial cells were counted as colony-forming units (CFUs) by plating the
720 diluted samples onto THY agar plates. Intracellular ferric ions were measured using a
721 QuantiChrom™ Iron Assay Kit (BioAssay Systems, Hayward, CA, USA), according to
722 the manufacturer's instructions. Briefly, 50 μL of standards or samples in 96-well plates
723 were mixed with 200 μL QuantiChrom™ Working Reagent and incubated at 20–24˚C for
724 an hour. The optical density at the wavelength of 600 nm was measured using an Infinite®
725 200 Pro F Plex Instrument (TECAN, Männedorf, Switzerland). The assay was repeated
726 three times, and the results of the respective experiments were combined. Statistical
727 analyses were performed using the Mann–Whitney U test.

728

729 **Bactericidal assay**

730 The bactericidal assay was performed as described previously, with minor

731 modifications[36,77-79]. Briefly, whole blood was collected from healthy adults. Human
732 neutrophils and erythrocytes were prepared using PolymorphPrep™ (Serumwerk
733 Bernburg, Bernburg, Germany), according to the manufacturer's instructions.
734 Heparinized human blood was centrifuged at 500 × *g* for 30 min, to isolate erythrocytes
735 and polymorphonuclear cells, which were then suspended in Roswell Park Memorial
736 Institute (RPMI)-1640 medium containing L-glutamine and Phenol Red (Fujifilm Wako
737 Pure Chemical Corporation). Heat-inactivated plasma was prepared at 56ºC for 30 min.
738 Subsequently, 195 μL of heparinized human whole blood, erythrocytes in RPMI-1640,
739 polymorphonuclear leukocytes in RPMI-1640, plasma, heat-inactivated plasma, or brain
740 heart infusion broth (BD Biosciences), and 5 μL of early exponential phase bacteria with
741 $0.9–2.0\times10^4$ CFUs/well were mixed in 96-well plates and incubated at 37°C, in an
742 atmosphere containing 5% CO_2, for 1, 2, and 3 h. Viable bacterial cells were counted as
743 CFUs by plating the diluted samples onto THY agar plates. The growth index was
744 calculated as the number of CFUs at a specified time point divided by the number of
745 CFUs in the initial inoculum. The assay was repeated three times, and the results of the
746 respective experiments were combined. Statistical analyses were performed using the
747 Mann–Whitney U test. Differences were considered statistically significant at $p<0.05$,
748 using Prism v.7.0c (GraphPad, La Jolla, CA, USA).

749

750 **Ethical approval**

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

757

758 **Data availability**

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

762

763 ## Acknowledgments

764 We would like to thank the NGS core facility of the Genome Information Research Center
765 at the Research Institute for Microbial Diseases of The University of Osaka for their
766 support in the DNA sequencing and data analysis and the Bioinformatic Research Unit of

767 Graduate School of Dentistry, The University of Osaka for their support in the
768 bioinformatics analysis. This study was partially performed on the National Institute of
769 Genetics (NIG) supercomputer at the Research Organization of Information and Systems
770 National Institute of Genetics. This study was partly completed using SQUID at the D3
771 center, The University of Osaka, Japan, under the "Joint Usage/Research Center for
772 Interdisciplinary Large-scale Information Infrastructures (JHPCN)" (Project ID:
773 EX22701, jh230035, jh240003, and jh250016). Masayuki Ono and Kotaro Higashi were
774 recipients of the Iwadare Scholarship from the Iwadare Scholarship Foundation. We wish
775 to express our gratitude to Mami Tateshita (Sapporo City Institute of Public Health) as
776 well as the medical institutions that participated in the collection of clinical strains.

777 This study was partly supported by AMED (JP17fk0108044, JP20fk0108130,
778 JP20wm0325001, and JP243fa727001h), the Japan Society for the Promotion of Science
779 KAKENHI (grant numbers 20KK0210, 22H03262, 22K19618, 22K19619, 23H03073,
780 23K19687, and 24K19854), the Takeda Science Foundation, Naito Foundation, and Joint
781 Research Program of the Research Center for GLOBAL and LOCAL Infectious Diseases,
782 Oita University (2022B05, 2025B01, and 2025B13). This study was conducted as part of
783 "The Nippon Foundation - Osaka University Project for Infectious Disease Prevention."
784 This study was supported by JST SPRING (grant number JPMJSP2138). The funders had

785 no role in the study design, data collection or analysis, decision to publish, or preparation
786 of the manuscript.
787

788 ## Author contributions

789 M.Y., M.O., and S.K. designed the study. M.O., M.Y., and K.H. designed and performed
790 bioinformatics analysis. D.M. performed the next-generation sequencing. M.O.
791 performed the in vitro and in vivo experiments. M.Y., Y.H., T.S., and S.K. contributed to
792 the setup of the experiments. R.O., T.Y., R.K., H.O., N.N., Y.K., C.N., R.Y., H.S., and
793 Y.M. collected clinical strains. M.O. and M.Y. wrote the original manuscript. All authors
794 reviewed and edited the manuscript.
795

796 ## Declaration of interests

797 The authors declare no competing interests.
798

799 ## References

800 1. Craik, N., Hla, T., Cannon, J., Moore, H., Carapetis, J.R., and Sanyahumbi, A.
801 (2022). Global Disease Burden of *Streptococcus pyogenes*. In *Streptococcus*
802 *pyogenes*: Basic Biology to Clinical Manifestations [Internet] J.J. Ferretti, D.L.

803 Stevens, and V.A. Fischetti, eds. (University of Oklahoma Health Sciences
804 Center), Chapter 21.
805 2. Brouwer, S., Rivera-Hernandez, T., Curren, B.F., Harbison-Price, N., De Oliveira,
806 D.M.P., Jespersen, M.G., Davies, M.R., and Walker, M.J. (2023). Pathogenesis,
807 epidemiology and control of Group A *Streptococcus* infection. Nat. Rev.
808 Microbiol. 21, 431–447. 10.1038/s41579-023-00865-7.
809 3. Carapetis, J.R., Steer, A.C., Mulholland, E.K., and Weber, M. (2005). The global
810 burden of group A streptococcal diseases. Lancet Infect. Dis. 5, 685-694.
811 10.1016/S1473-3099(05)70267-X.
812 4. Stevens, D.L., Bisno, A.L., Chambers, H.F., Dellinger, E.P., Goldstein, E.J.,
813 Gorbach, S.L., Hirschmann, J.V., Kaplan, S.L., Montoya, J.G., Wade, J.C., et al.
814 (2014). Practice guidelines for the diagnosis and management of skin and soft
815 tissue infections: 2014 update by the Infectious Diseases Society of America. Clin.
816 Infect. Dis. 59, e10–e52. 10.1093/cid/ciu444.
817 5. Stevens, D.L., and Bryant, A.E. (2016). Severe Group A Streptococcal Infections.
818 In *Streptococcus pyogenes*: Basic Biology to Clinical Manifestations [internet],
819 J.J. Ferretti, D.L. Stevens, and V.A. Fischetti, eds. (University of Oklahoma
820 Health Sciences Center).

821 6. Walker, M.J., Barnett, T.C., McArthur, J.D., Cole, J.N., Gillen, C.M., Henningham,
822 A., Sriprakash, K.S., Sanderson-Smith, M.L., and Nizet, V. (2014). Disease
823 manifestations and pathogenic mechanisms of Group A Streptococcus. Clin.
824 Microbiol. Rev. 27, 264–301. 10.1128/CMR.00101-13.
825 7. Yamaguchi, M., Terao, Y., Mori-Yamaguchi, Y., Domon, H., Sakaue, Y., Yagi, T.,
826 Nishino, K., Yamaguchi, A., Nizet, V., and Kawabata, S. (2013). *Streptococcus*
827 *pneumoniae* invades erythrocytes and utilizes them to evade human innate
828 immunity. PLOS ONE 8, e77282. 10.1371/JOURNAL.PONE.0077282.
829 8. Bessen, D.E., Smeesters, P.R., and Beall, B.W. (2018). Molecular Epidemiology,
830 Ecology, and Evolution of Group A Streptococci. Microbiol. Spectr. 6.
831 10.1128/microbiolspec.CPP3-0009-2018.
832 9. Ikebe, T., Okuno, R., Kanda, Y., Sasaki, M., Yamaguchi, T., Otsuka, H., Kazawa,
833 Y., Suzuki, M., Ohya, H., Uchida, K., and Ohnishi, M. (2021). Molecular
834 characterization and antimicrobial resistance of group A streptococcus isolates in
835 streptococcal toxic shock syndrome cases in Japan from 2013 to 2018. Int. J. Med.
836 Microbiol. 311, 151496. 10.1016/j.ijmm.2021.151496.
837 10. Fay, K., Onukwube, J., Chochua, S., Schaffner, W., Cieslak, P., Lynfield, R., Muse,
838 A., Smelser, C., Harrison, L.H., Farley, M., et al. (2021). Patterns of Antibiotic

839 Nonsusceptibility Among Invasive Group A *Streptococcus* Infections-United
840 States, 2006–2017. Clin. Infect. Dis. 73, 1957–1964. 10.1093/cid/ciab575.
841 11. Turner, C.E., Abbott, J., Lamagni, T., Holden, M.T.G., David, S., Jones, M.D.,
842 Game, L., Efstratiou, A., and Sriskandan, S. (2015). Emergence of a new highly
843 successful acapsular group a Streptococcus clade of genotype *emm*89 in the
844 United Kingdom. mBio 6, e00622. 10.1128/mBio.00622-15.
845 12. Zhu, L., Olsen, R.J., Nasser, W., Beres, S.B., Vuopio, J., Kristinsson, K.G.,
846 Gottfredsson, M., Porter, A.R., DeLeo, F.R., and Musser, J.M. (2015). A molecular
847 trigger for intercontinental epidemics of group A Streptococcus. J. Clin. Invest.
848 125, 3545–3559. 10.1172/JCI82478.
849 13. Zhu, L., Olsen, R.J., Nasser, W., De La Riva Morales, I., and Musser, J.M. (2015).
850 Trading Capsule for Increased Cytotoxin Production: Contribution to Virulence
851 of a Newly Emerged Clade of *emm*89 *Streptococcus pyogenes*. mBio 6, e01378–
852 e01315. 10.1128/mbio.01378-15.
853 14. Hirose, Y., Yamaguchi, M., Takemoto, N., Miyoshi-Akiyama, T., Sumitomo, T.,
854 Nakata, M., Ikebe, T., Hanada, T., Yamaguchi, T., Kawahara, R., et al. (2020).
855 Genetic Characterization of *Streptococcus pyogenes emm*89 Strains Isolated in
856 Japan From 2011 to 2019. Infect. Microbes Dis. 2, 160–166.

857 10.1097/IM9.0000000000000038.

858 15. Walker, M.J., Hollands, A., Sanderson-Smith, M.L., Cole, J.N., Kirk, J.K.,
859 Henningham, A., McArthur, J.D., Dinkla, K., Aziz, R.K., Kansal, R.G., et al.
860 (2007). DNase Sda1 provides selection pressure for a switch to invasive group A
861 streptococcal infection. Nat. Med. 13, 981–985. 10.1038/nm1612.

862 16. Sumby, P., Whitney, A.R., Graviss, E.A., DeLeo, F.R., and Musser, J.M. (2006).
863 Genome-wide analysis of group a streptococci reveals a mutation that modulates
864 global phenotype and disease specificity. PLoS Pathog. 2, e5.
865 10.1371/journal.ppat.0020005.

866 17. Cole, J.N., Barnett, T.C., Nizet, V., and Walker, M.J. (2011). Molecular insight
867 into invasive group A streptococcal disease. Nat. Rev. Microbiol. 9, 724–736.
868 10.1038/nrmicro2648.

869 18. Ikebe, T., Ato, M., Matsumura, T., Hasegawa, H., and Sata, T. (2010). Highly
870 Frequent Mutations in Negative Regulators of Multiple Virulence Genes in Group
871 A Streptococcal Toxic Shock Syndrome Isolates. PLOS Pathog. 6, e1000832.
872 10.1371/journal.ppat.1000832.

873 19. Ministry of Health, Labour and Welfare. Severe Invasive Group A Streptococcal
874 Infections. https://www.mhlw.go.jp/bunya/kenkou/kekkaku-kansenshou11/01-

875 05-06.html.

876 20. Centers for Disease Control and Prevention, C.f.D.C.a. Streptococcal Toxic Shock
877 Syndrome (STSS) (*Streptococcus pyogenes*) 2010 Case Definition.
878 https://ndc.services.cdc.gov/case-definitions/streptococcal-toxic-shock-
879 syndrome-2010/.

880 21. Ikebe, T., Murai, N., Endo, M., Okuno, R., Murayama, S., Saitoh, K., Yamai, S.,
881 Suzuki, R., Isobe, J., Tanaka, D., et al. (2003). Changing prevalent T serotypes
882 and *emm* genotypes of *Streptococcus pyogenes* isolates from streptococcal toxic
883 shock-like syndrome (TSLS) patients in Japan. Epidemiol. Infect. 130, 569–572.

884 22. Beall, B., Facklam, R.R., Elliott, J.A., Franklin, A.R., Hoenes, T., Jackson, D.,
885 Laclaire, L., Thompson, T., and Viswanathan, R. (1998). Streptococcal *emm* types
886 associated with T-agglutination types and the use of conserved *emm* gene
887 restriction fragment patterns for subtyping group A streptococci. J. Med.
888 Microbiol. *47,* 893–898. 10.1099/00222615-47-10-893.

889 23. Johnson, D.R., and Kaplan, E.L. (1993). A review of the correlation of T-
890 agglutination patterns and M-protein typing and opacity factor production in the
891 identification of group A streptococci. J. Med. Microbiol. 38, 311–315.
892 10.1099/00222615-38-5-311.

893 24. Katsukawa, C., Tamaru, A., Morikawa, Y., and Oda, K. (2002). M protein gene
894 (*emm*) typing of *Streptococcus pyogenes*. Kansenshogaku Zasshi 76, 238–245.
895 10.11150/kansenshogakuzasshi1970.76.238.
896 25. Beres, S.B., Olsen, R.J., Saavedra, M.O., Ure, R., Reynolds, A., Lindsay, D.S.J.,
897 Smith, A.J., and Musser, J.M. (2017). Genome sequence analysis of *emm*89
898 *Streptococcus pyogenes* strains causing infections in Scotland, 2010–2016. J. Med.
899 Microbiol. 66, 1765–1773. 10.1099/jmm.0.000622.
900 26. Chochua, S., Metcalf, B.J., Li, Z., Rivers, J., Mathis, S., Jackson, D., Gertz, R.E.,
901 Jr., Srinivasan, V., Lynfield, R., Van Beneden, C., et al. (2017). Population and
902 Whole Genome Sequence Based Characterization of Invasive Group A
903 Streptococci Recovered in the United States during 2015. mBio 8, e01422-17.
904 10.1128/mBio.01422-17.
905 27. Davies, M.R., McIntyre, L., Mutreja, A., Lacey, J.A., Lees, J.A., Towers, R.J.,
906 Smeesters, P.R., Frost, H.R., Price, D.J., G Holden, M.T., et al. (2019). Atlas of
907 group A streptococcal vaccine candidates compiled using large-scale comparative
908 genomics. Nat. Genet. 51, 1035–1043. 10.1038/s41588-019-0417-8.
909 28. Turner, C.E., Holden, M.T.G., Blane, B., Horner, C., Peacock, S.J., and
910 Sriskandan, S. (2019). The Emergence of Successful *Streptococcus pyogenes*

911 Lineages through Convergent Pathways of Capsule Loss and Recombination
912 Directing High Toxin Expression. mBio 10, 1–20. 10.1128/mBio.02521-19.
913 29. Paul, S., Olymon, K., Martinez, G.S., Sarkar, S., Yella. V.R., and Kumar, A. (2024).
914 MLDSPP: Bacterial Promoter Prediction Tool Using DNA Structural Properties
915 with Machine Learning and Explainable AI. J Chem Inf Model. 64, 7, 2705–2719.
916 10.1021/acs.jcim.3c02017
917 30. de Avila e Silva, S., Echeverrigaray, S., and Gerhardt G.J. (2011). BacPP:
918 Bacterial promoter prediction - A tool for accurate sigma-factor specific
919 assignment in enterobacteria. J Theor Biol. 287, 92–99.
920 31. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O.,
921 Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly
922 accurate protein structure prediction with AlphaFold. Nature, 583–589.
923 10.1038/s41586-021-03819-2.
924 32. Jain, I., Danger, J.L., Burgess, C., Uppal, T., and Sumby, P. (2020). The group A
925 Streptococcus accessory protein RocA: regulatory activity, interacting partners
926 and influence on disease potential. Mol. Microbiol. 113, 190–207.
927 10.1111/mmi.14410.
928 33. Ribardo, D.A., and McIver, K.S. (2006). Defining the Mga regulon: Comparative

929 transcriptome analysis reveals both direct and indirect regulation by Mga in the
930 group A streptococcus. Mol. Microbiol. 62, 491–508. 10.1111/j.1365-
931 2958.2006.05381.x.
932 34. Kachroo, P., Eraso, J.M., Beres, S.B., Olsen, R.J., Zhu, L., Nasser, W., Bernard,
933 P.E., Cantu, C.C., Saavedra, M.O., Arredondo, M.J., et al. (2019). Integrated
934 analysis of population genomics, transcriptomics and virulence provides novel
935 insights into *Streptococcus pyogenes* pathogenesis. Nat. Genet. 51, 548–559.
936 10.1038/S41588-018-0343-1.
937 35. Vega, L.A., Malke, H., and McIver, K.S. (2016). Virulence-Related
938 Transcriptional Regulators of *Streptococcus pyogenes*. In *Streptococcus*
939 *pyogenes* : Basic Biology to Clinical Manifestations [Internet], J.J. Ferretti, D.L.
940 Stevens, and V.A. Fischetti, eds (University of Oklahoma Health Sciences Center).
941 36. Terao, Y., Mori, Y., Yamaguchi, M., Shimizu, Y., Ooe, K., Hamada, S., and
942 Kawabata, S. (2008). Group A streptococcal cysteine protease degrades C3 (C3b)
943 and contributes to evasion of innate immunity. J. Biol. Chem. 283, 6253–6260.
944 10.1074/jbc.M704821200.
945 37. Sumitomo, T., Nakata, M., Higashino, M., Jin, Y., Terao, Y., Fujinaga, Y., and
946 Kawabata, S. (2011). Streptolysin S contributes to group A streptococcal

947 translocation across an epithelial barrier. J. Biol. Chem. 286, 2750–2761.
948 10.1074/jbc.M110.171504.
949 38. Honda-Ogawa, M., Ogawa, T., Terao, Y., Sumitomo, T., Nakata, M., Ikebe, K.,
950 Maeda, Y., and Kawabata, S. (2013). Cysteine proteinase from Streptococcus
951 pyogenes enables evasion of innate immunity via degradation of complement
952 factors. J. Biol. Chem. 288, 15854–15864. 10.1074/jbc.M113.469106.
953 39. Lin, A., Loughman, J.A., Zinselmeyer, B.H., Miller, M.J., and Caparon, M.G.
954 (2009). Streptolysin S inhibits neutrophil recruitment during the early stages of
955 *Streptococcus pyogenes* infection. Infect. Immun. 77, 5190–5201.
956 10.1128/IAI.00420-09.
957 40. Cimen, M.Y. (2008). Free radical metabolism in human erythrocytes. Clin. Chim.
958 Acta. 390, 1–11. 10.1016/j.cca.2007.12.025.
959 41. Laabei, M., and Ermert, D. (2019). Catch Me if You Can: *Streptococcus pyogenes*
960 Complement Evasion Strategies. J. Innate. Immun. 11, 3–12. 10.1159/000492944.
961 42. He, L.Y., Yu, Y.B., Liu, Y., Le, Y.J., Li, S., and Yang, X.Y. (2022). Immunization
962 with the lipoprotein FtsB stimulates protective immunity against *Streptococcus*
963 *pyogenes* infection in mice. Front. Microbiol. 13, 969490.
964 10.3389/fmicb.2022.969490.

965 43. Nguyen, M.T., and Gotz, F. (2016). Lipoproteins of Gram-Positive Bacteria: Key
966 Players in the Immune Response and Virulence. Microbiol. Mol. Biol. Rev. 80,
967 891–903. 10.1128/MMBR.00028-16.
968 44. Bessen, D.E., McShan, W.M., Nguyen, S.V., Shetty, A., Agrawal, S., and Tettelin,
969 H. (2015). Molecular epidemiology and genomics of group A Streptococcus.
970 Infect. Genet. Evol. 33, 393–418. 10.1016/j.meegid.2014.10.011.
971 45. McShan, W.M., McCullor, K.A., and Nguyen, S.V. (2019). The Bacteriophages of
972 *Streptococcus pyogenes*. Microbiol. Spectr. 7. 10.1128/microbiolspec.GPP3-
973 0059-2018.
974 46. Berbel, D., Camara, J., Gonzalez-Diaz, A., Cubero, M., Lopez de Egea, G., Marti,
975 S., Tubau, F., Dominguez, M.A., and Ardanuy, C. (2021). Deciphering mobile
976 genetic elements disseminating macrolide resistance in *Streptococcus pyogenes*
977 over a 21 year period in Barcelona, Spain. J. Antimicrob. Chemother. 76, 1991–
978 2003. 10.1093/jac/dkab130.
979 47. Wajima, T., Chiba, N., Morozumi, M., Shouji, M., Sunaoshi, K., Sugita, K.,
980 Tajima, T., Ubukata, K., and Group A Streptococcus Surveillance Study Group.
981 (2014). Prevalence of macrolide resistance among group A streptococci isolated
982 from pharyngotonsillitis. Microb. Drug. Resist. 20, 431–435.

983 10.1089/mdr.2013.0213.

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

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

993 50. Tanaka, Y., Gotoh, K., Teramachi, M., Ishimoto, K., Tsumura, N., Shindou, S., and
994 Yamashita, Y. (2016). Molecular epidemiology, antimicrobial susceptibility, and
995 characterization of macrolide-resistant *Streptococcus pyogenes* in Japan. J. Infect.
996 Chemother. 22, 727–732. 10.1016/j.jiac.2016.06.013.

997 51. Ubukata, K., Wajima, T., Morozumi, M., Sakuma, M., Tajima, T., Matsubara, K.,
998 Itahashi, K., and Iwata, S. (2020). Changes in epidemiologic characteristics and
999 antimicrobial resistance of *Streptococcus pyogenes* isolated over 10 years from
1000 Japanese children with pharyngotonsillitis. J. Med. Microbiol. 69, 443–450.

1001 10.1099/jmm.0.001158.

1002 52. Jolley, K.A., Bray, J.E., and Maiden, M.C.J. (2018). Open-access bacterial
1003 population genomics: BIGSdb software, the PubMLST.org website and their
1004 applications. Wellcome Open Res. 3, 124. 10.12688/wellcomeopenres.14826.1.

1005 53. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one
1006 FASTQ preprocessor. Bioinformatics 34, i884–i890.
1007 10.1093/bioinformatics/bty560.

1008 54. Kapatai, G., Coelho, J., Platt, S., and Chalker, V.J. (2017). Whole genome
1009 sequencing of group A Streptococcus: development and evaluation of an
1010 automated pipeline for *emm* gene typing. PeerJ 5, e3226. 10.7717/peerj.3226.

1011 55. Souvorov, A., Agarwala, R., and Lipman, D.J. (2018). SKESA: strategic k-mer
1012 extension for scrupulous assemblies. Genome Biol. 19, 153. 10.1186/s13059-018-
1013 1540-z.

1014 56. Jolley, K.A., and Maiden, M.C. (2010). BIGSdb: scalable analysis of bacterial
1015 genome variation at the population level. BMC Bioinformatics 11, 595.
1016 10.1186/1471-2105-11-595.

1017 57. Maiden, M.C.J., Jansen Van Rensburg, M.J., Bray, J.E., Earle, S.G., Ford, S.A.,
1018 Jolley, K.A., and McCarthy, N.D. (2013). MLST revisited: the gene-by-gene

1019 approach to bacterial genomics Europe PMC Funders Group. Nat. Rev. Microbiol.
1020 11, 728–736. 10.1038/nrmicro3093.
1021 58. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation.
1022 Bioinformatics 30, 2068–2069. 10.1093/bioinformatics/btu153.
1023 59. Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.,
1024 Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-
1025 scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693.
1026 10.1093/bioinformatics/btv421.
1027 60. Jaillard, M., Lima, L., Tournoud, M., Mahé, P., van Belkum, A., Lacroix, V., and
1028 Jacob, L. (2018). A fast and agnostic method for bacterial genome-wide
1029 association studies: Bridging the gap between k-mers and genetic events. PLOS
1030 Genet. 14, e1007758. 10.1371/journal.pgen.1007758.
1031 61. Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE:
1032 a fast and effective stochastic algorithm for estimating maximum-likelihood
1033 phylogenies. Mol. Biol. Evol. 32, 268–274. 10.1093/molbev/msu300.
1034 62. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin,
1035 L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic
1036 estimates. Nat. Methods 14, 587–589. 10.1038/nmeth.4285.

1037 63. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool
1038 for phylogenetic tree display and annotation. Nucleic Acids Res. 49, W293–W296.
1039 10.1093/nar/gkab301.
1040 64. R Core Team. (2022). A language and environment for statistical computing.
1041 https://www.R-project.org/.
1042 65. Lees, J.A., Galardini, M., Bentley, S.D., Weiser, J.N., and Corander, J. (2018).
1043 pyseer: A comprehensive tool for microbial pangenome-wide association studies.
1044 Bioinformatics 34, 4310–4312. 10.1093/bioinformatics/bty539.
1045 66. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren,
1046 S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance
1047 estimation using MinHash. Genome Biol. 17, 132. 10.1186/s13059-016-0997-x.
1048 67. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis, 2 Edition
1049 (Springer-Verlag New York). 10.1007/978-3-319-24277-4.
1050 68. Beres, S.B., Kachroo, P., Nasser, W., Olsen, R.J., Zhu, L., Flores, A.R., de la Riva,
1051 I., Paez-Mayorga, J., Jimenez, F.E., Cantu, C., et al. (2016). Transcriptome
1052 remodeling contributes to epidemic disease caused by the human pathogen
1053 *Streptococcus pyogenes*. mBio 7, 1–14. 10.1128/mBio.00403-16.
1054 69. Martinez, G.S., Perez-Rueda, E., Kumar, A., Dutt, M., Maya, C.R., Ledesma-

1055 Dominguez, L., Casa, P.L., Kumar, A., de Avila e Silva, S., and Kelvin, D.J.
1056 (2024). CDBProm: the Comprehensive Directory of Bacterial Promoters. NAR
1057 Genom Bioinform. 6, 1, lqae018. 10.1093/nargab/lqae018.
1058 70. Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular
1059 Biology Open Software Suite. Trends Genet. 16, 276–277. 10.1016/s0168-
1060 9525(00)02024-2.
1061 71. Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). lDDT: a local
1062 superposition-free score for comparing protein structures and models using
1063 distance difference tests. Bioinformatics 29, 2722–2728.
1064 10.1093/bioinformatics/btt473.
1065 72. Yin, R., Feng, B.Y., Varshney, A., and Pierce, B.G. (2022). Benchmarking
1066 AlphaFold for protein complex modeling reveals accuracy determinants. Protein
1067 Sci. 31, e4379. 10.1002/pro.4379.
1068 73. Hirokawa, T., Boon-Chieng, S., and Mitaku, S. (1998). SOSUI: classification and
1069 secondary structure prediction system for membrane proteins. Bioinformatics 14,
1070 378–379. 10.1093/bioinformatics/14.4.378.
1071 74. Takamatsu, D., Osaki, M., and Sekizaki, T. (2001). Thermosensitive suicide
1072 vectors for gene replacement in *Streptococcus suis*. Plasmid 46, 140–148.

1073 10.1006/PLAS.2001.1532.

1074 75. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general
1075 purpose program for assigning sequence reads to genomic features.
1076 Bioinformatics 30, 923–930. 10.1093/bioinformatics/btt656.

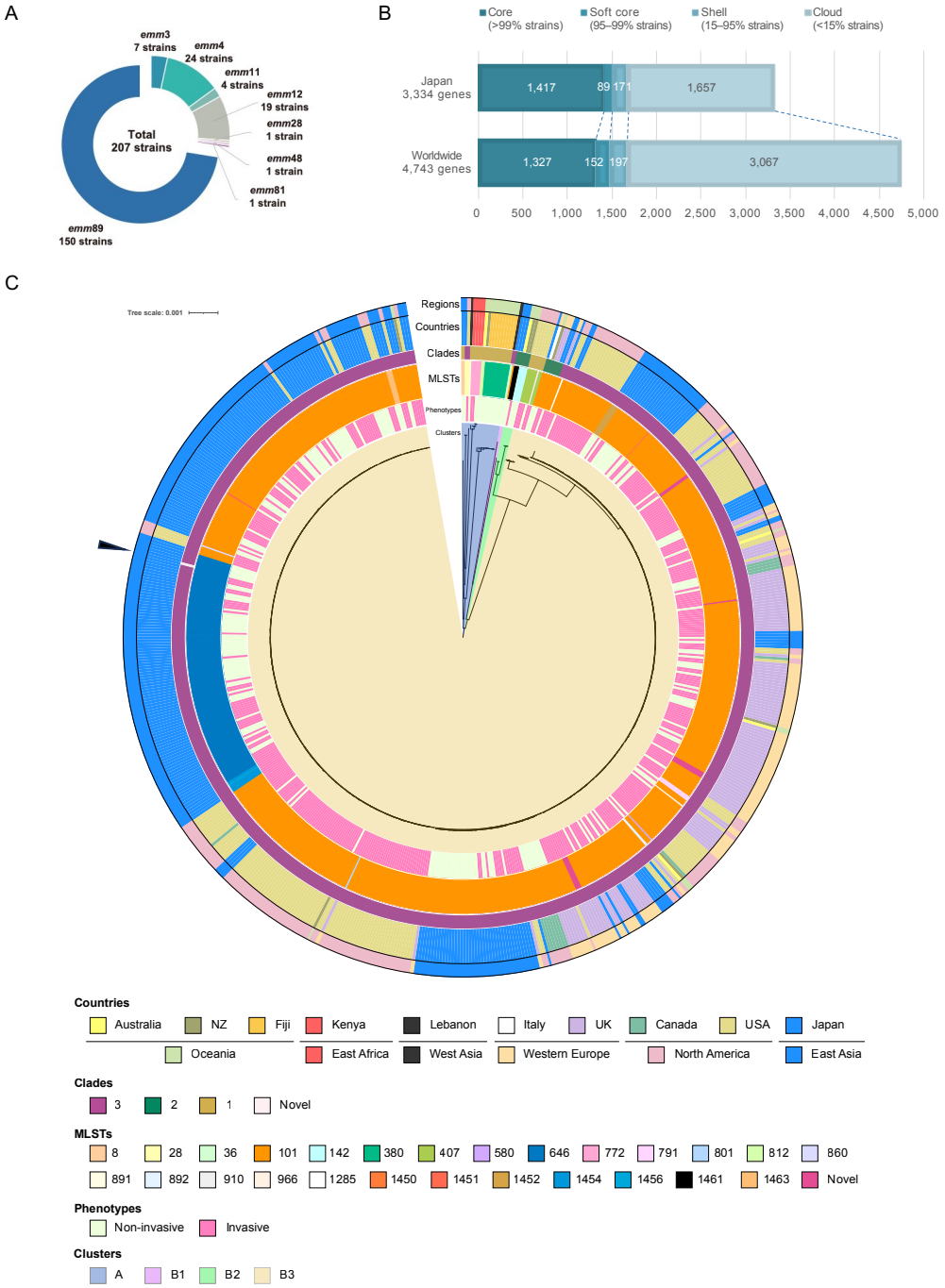
1077 76. Ge, S.X., Son, E.W., and Yao, R. (2018). iDEP: an integrated web application for
1078 differential expression and pathway analysis of RNA-Seq data. BMC
1079 Bioinformatics 19, 534. 10.1186/s12859-018-2486-6.

1080 77. Lancefield, R.C. (1957). Differentiation of group A streptococci with a common
1081 R antigen into three serological types, with special reference to the bactericidal
1082 test. J. Exp. Med. 106, 525–544. 10.1084/jem.106.4.525.

1083 78. Yamaguchi, M., Goto, K., Hirose, Y., Yamaguchi, Y., Sumitomo, T., Nakata, M.,
1084 Nakano, K., and Kawabata, S. (2019). Identification of evolutionarily conserved
1085 virulence factor by selective pressure analysis of *Streptococcus pneumoniae*.
1086 Commun. Biol. *2*, 96. 10.1038/s42003-019-0340-7.

1087 79. Takemura, M., Yamaguchi, M., Kobayashi, M., Sumitomo, T., Hirose, Y., Okuzaki,
1088 D., Ono, M., Motooka, D., Goto, K., Nakata, M., et al. (2022). Pneumococcal
1089 BgaA Promotes Host Organ Bleeding and Coagulation in a Mouse Sepsis Model.
1090 Front. Cell. Infect. Microbiol. 12, 844000. 10.3389/fcimb.2022.844000

1091 **Figures**



1092

1093 **Figure 1. Pan-genomic and phylogenetic characterization of the collected *emm*89 *S.***

1094 ***pyogenes* isolates.**

1095 (A) *emm* genotyping of the 207 clinical isolates collected in Japan. (B) Pan-genome

1096 analysis of the Japanese and global cohorts. All genes detected in each cohort were

1097 classified into four groups, according to prevalence: core, soft-core, shell, and cloud genes.

1098 (C) Phylogenetic tree for the global cohort, based on the sequences of the core genes.

1099 From the inside, the color bars show clusters, phenotypes, MLSTs, clades, countries

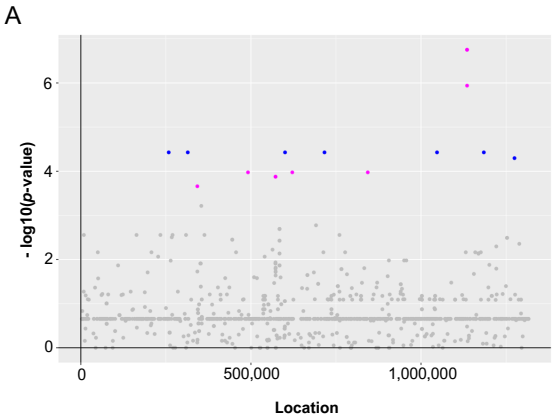
1100 where strains were isolated, and global regions, in the order mentioned. The root of the

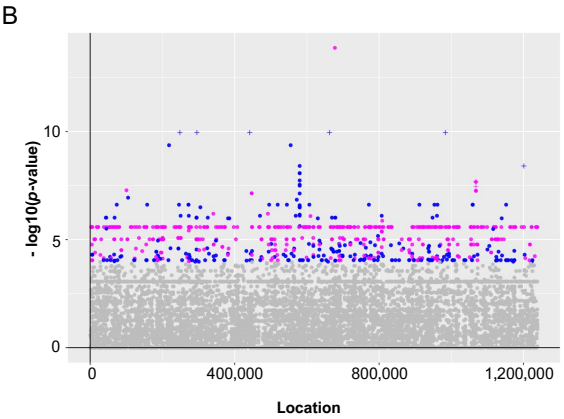
1101 tree was set as the mid-point. The scale located upper left indicates 0.001 times

1102 substitution of the bases on average. Arrowhead indicates strain belonging to the novel

1103 clade.

1104 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

Figure 2. Pan-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

1111 percentile of the 1,000 minimal *p*-values (p=5.75×10^{-4} and p=1.05×10^{-4} for the Japanese

1112 and global cohorts, respectively). Plots with lower *p*-values than genome-wide significant

1113 levels are colored magenta and blue, based on the direction of their effect size (positive

1114 and negative, respectively). (C) Distribution heatmap for the global cohort, with the

1115 strains possessing the significant SNPs/indels colored orange. Only the 20 SNPs with the

1116 lowest *p*-values are shown in this heatmap. Colored bars above indicate countries and

1117 phenotypes, and magenta bars represent invasive phenotypes. Using the Roary program,

1118 gene names starting with "Group_" were automatically assigned. Position indicates the

1119 location of each SNP/indel on the core gene alignment. The full results are shown in Table

1120 S6.

1121 SNP, single-nucleotide polymorphism; indel, insertion/deletion.

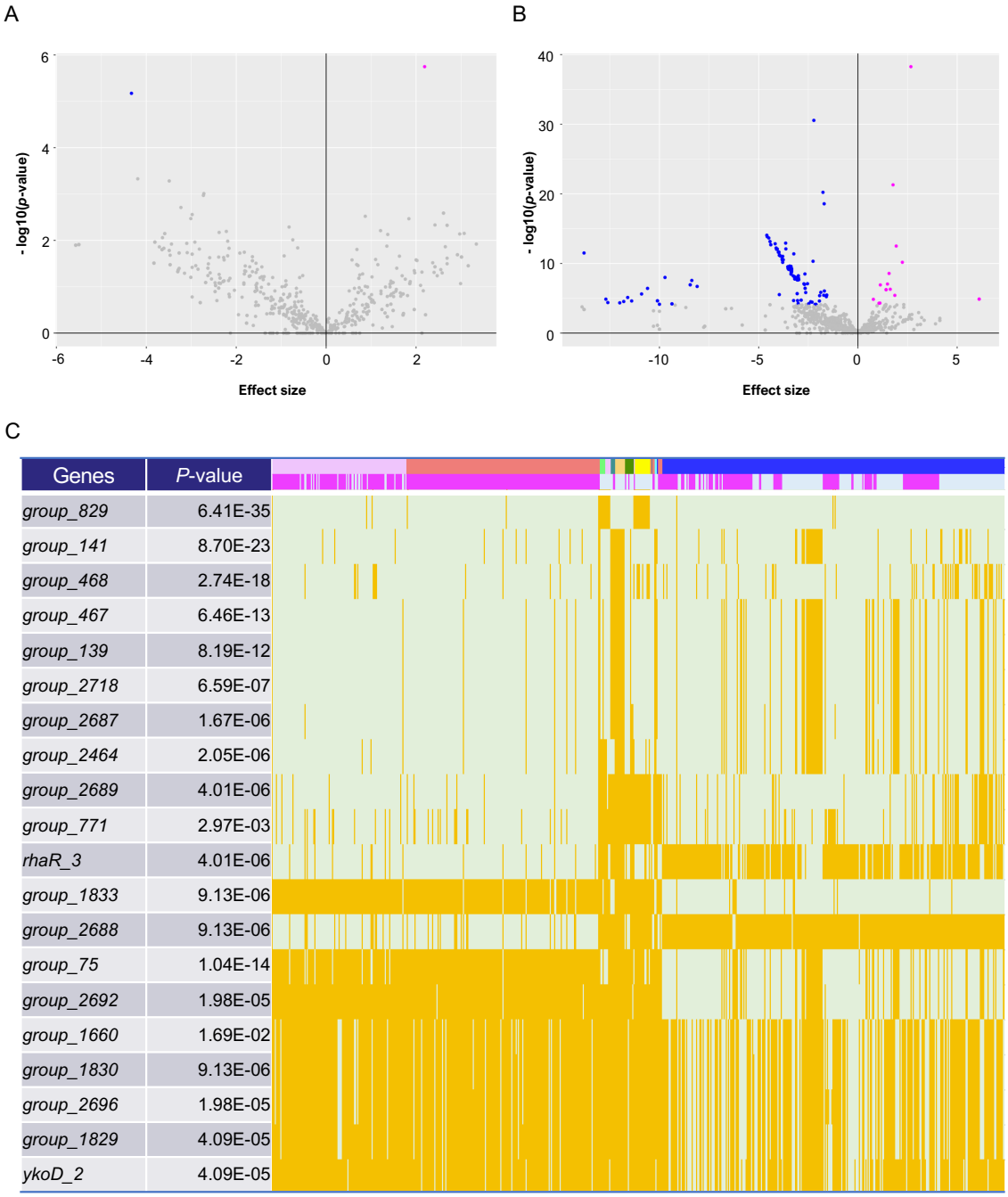


Figure 3. Pan-genome-wide association study on gene presence.

(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=1.09×10^{-4} and p=7.72×10^{-5} for the Japanese and

1127 global cohorts, respectively) have been colored magenta and blue, based on the direction

1128 of their effect size (positive and negative, respectively). (C) Distribution heatmap for the

1129 global cohort, with the strains possessing the significant genes colored orange. Only the

1130 20 genes with the lowest *p*-values are shown in this heatmap. Colored bars above indicate

1131 countries and phenotypes, and magenta bars represent invasive phenotypes. Using the

1132 Roary program, gene names starting with "Group_" were automatically assigned. The full

1133 results are shown in Table S8.

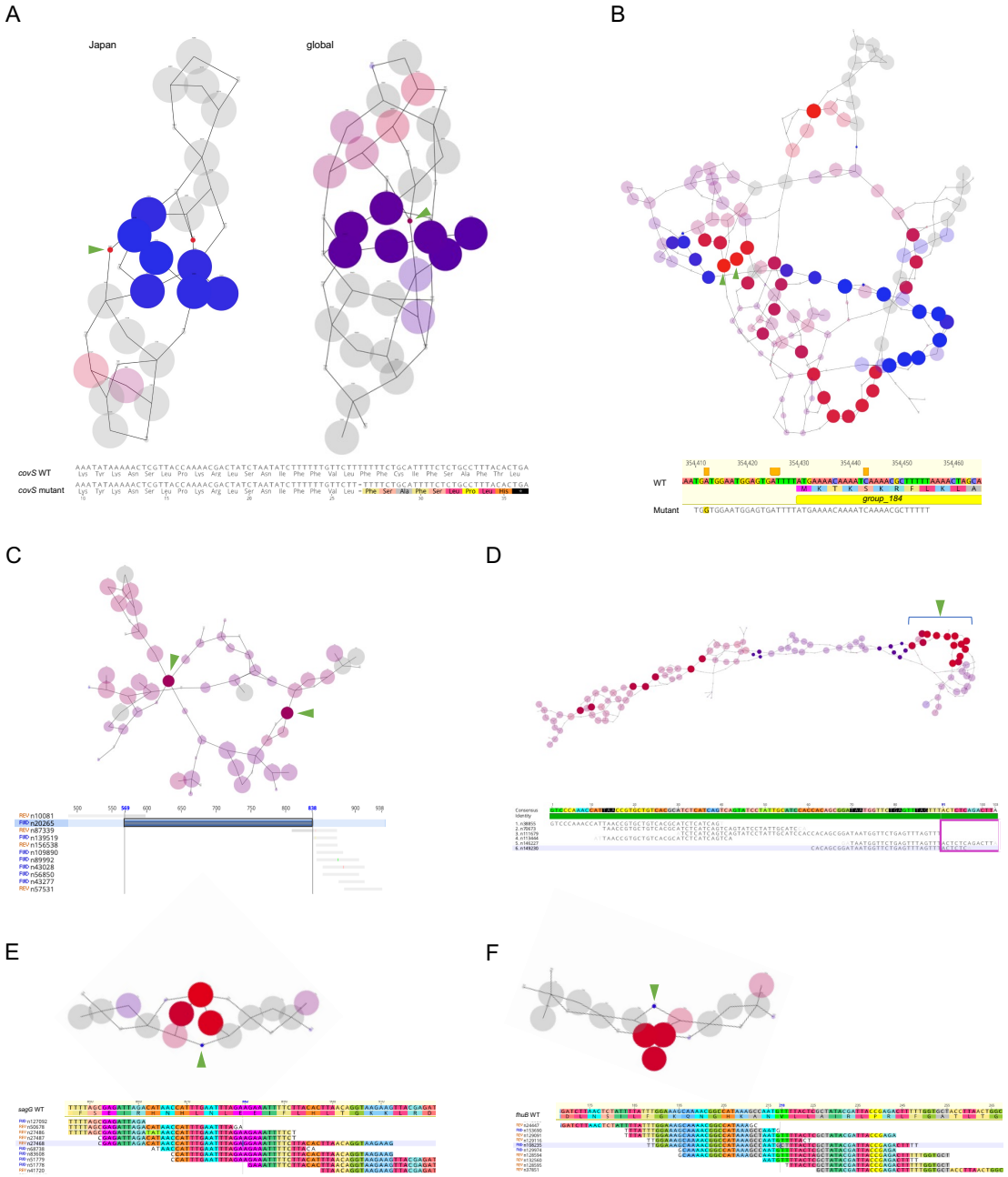
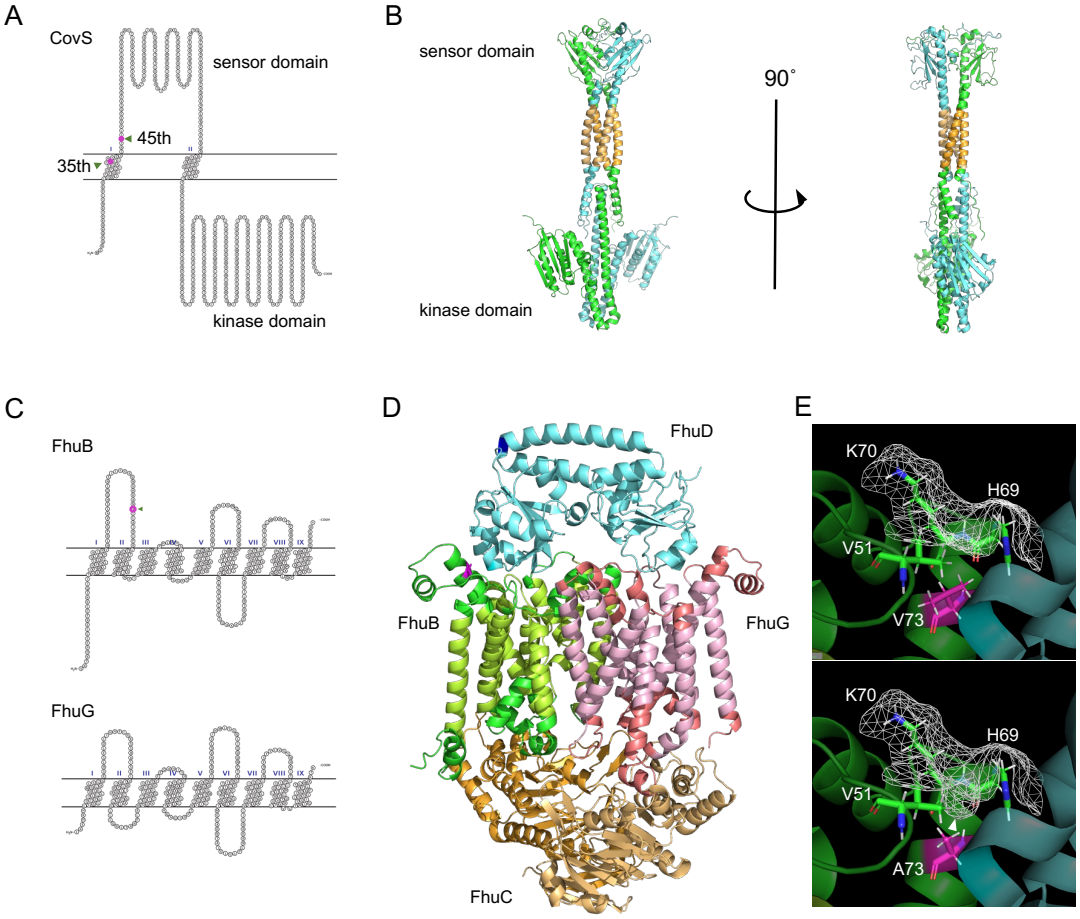


Figure 4. K-mers related to pathology.

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)

1139 de-Bruijn graphs generated using DBGWAS. The respective nodes in the graphs indicate
1140 k-mers, with the significant nodes indicated using green arrowheads. The size of each
1141 node corresponds to the allele frequency. (A–F, bottom) Alignments of k-mers or maps
1142 of k-mers on the genome sequence of MGAS27061 around regions including significant
1143 mutations. Each base is colored according to the interpreted amino acids. (A) The
1144 arrowheads indicate significant k-mers that cause a frameshift mutation, truncating CovS
1145 protein to 35 amino acids. (B) The significant k-mers indicate that the presence of a
1146 sequence mapped on the first 26 bp of group_184 and its upstream 20 bp is related to the
1147 pathology. (C) Two significant k-mers indicate the same intergenic region of 270 bps. (D)
1148 The significant k-mers are indicated using arrowheads.



1149

1150 **Figure 5. Predicted protein structure models.**

1151 (A) Snake-like plot of transmembrane regions of CovS estimated using SOSUI.

1152 Frameshift mutations, detected by the SNPs/indel- and k-mer-based GWASes, cause

1153 truncation of CovS at the indicated 35th and 45th residues, respectively. (B) Structural

1154 model of the CovS homodimer (ipTM + pTM=0.614). The putative transmembrane

1155 regions are colored orange. The upper part is the sensor domain, while the lower is the C-

1156 terminal kinase domain involved in the phosphorylation of the transcriptional regulator

1157 CovR. (C) Snake-like plot depicting the transmembrane regions of FhuB and FhuG. The

1158 73rd residue of FhuB is indicated by an arrowhead and in magenta. (D) Structural model

1159 of the FhuBDCCG complex (ipTM + pTM=0.791). The putative transmembrane regions

1160 of FhuB and FhuG are colored green-yellow and peach, respectively. The upper part of

1161 the model is located in the extracellular region, while the lower part is in the cytoplasm.

1162 (E) The 73rd valine in FhuB, shown in magenta, was substituted with alanine. The

1163 molecular surface is illustrated with a wireframe and that of the predicted indentation is

1164 shown with an arrowhead.

1165 ipTM + pTM: Weighted combination of interface-predicted TM and predicted TM scores.

1166 ipTM is used to measure structural accuracy in the protein-protein interface, while ipTM

1167 is a metric for overall topological accuracy.

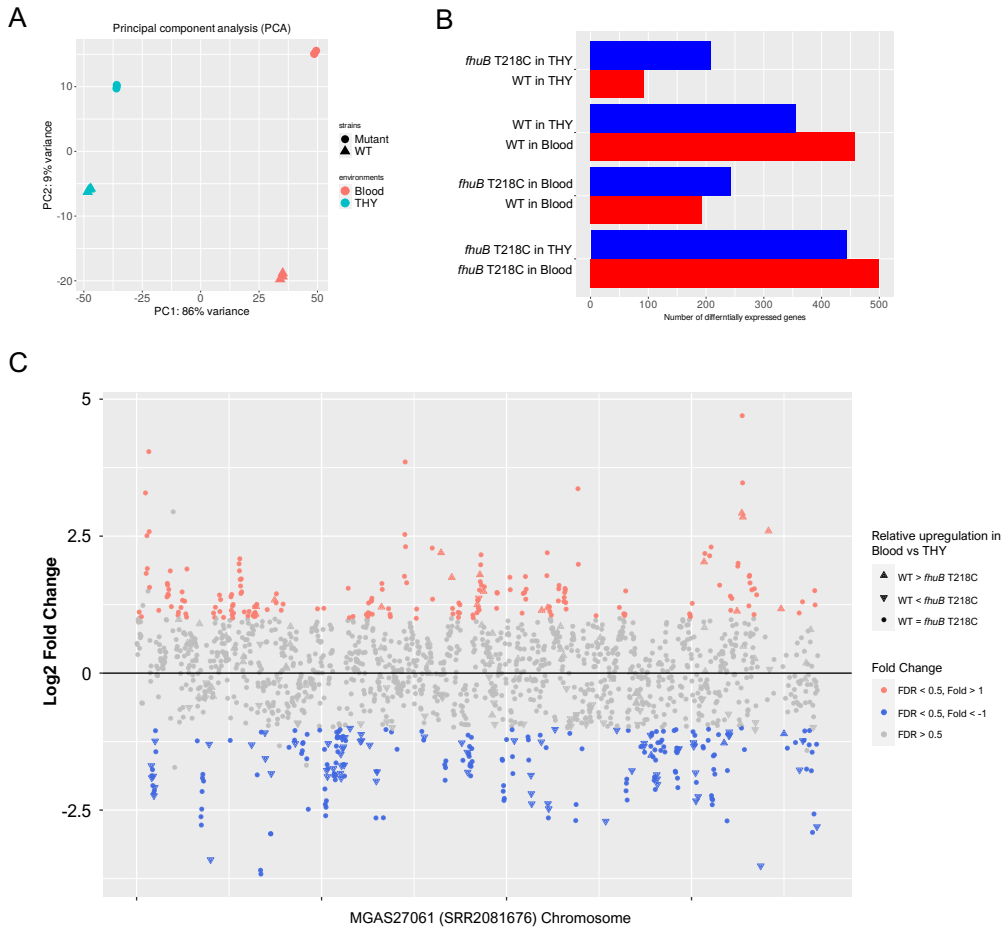


Figure 6. Transcriptome analysis of the *fhuB* T218C mutant strain in THY and human blood.

(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 are colored red and blue, respectively. The shapes of the plots indicate relative transcriptional changes between THY and blood. Genes depicted with upward triangles are either significantly upregulated in the WT or downregulated in the mutant strain, in

1177 blood *vs.* THY. The downward triangle plots indicate genes that are either downregulated

1178 in the WT or upregulated in the mutant strain, in blood *vs.* THY. WT, wild-type; THY,

1179 Todd Hewitt broth supplemented with 0.2% yeast extract.

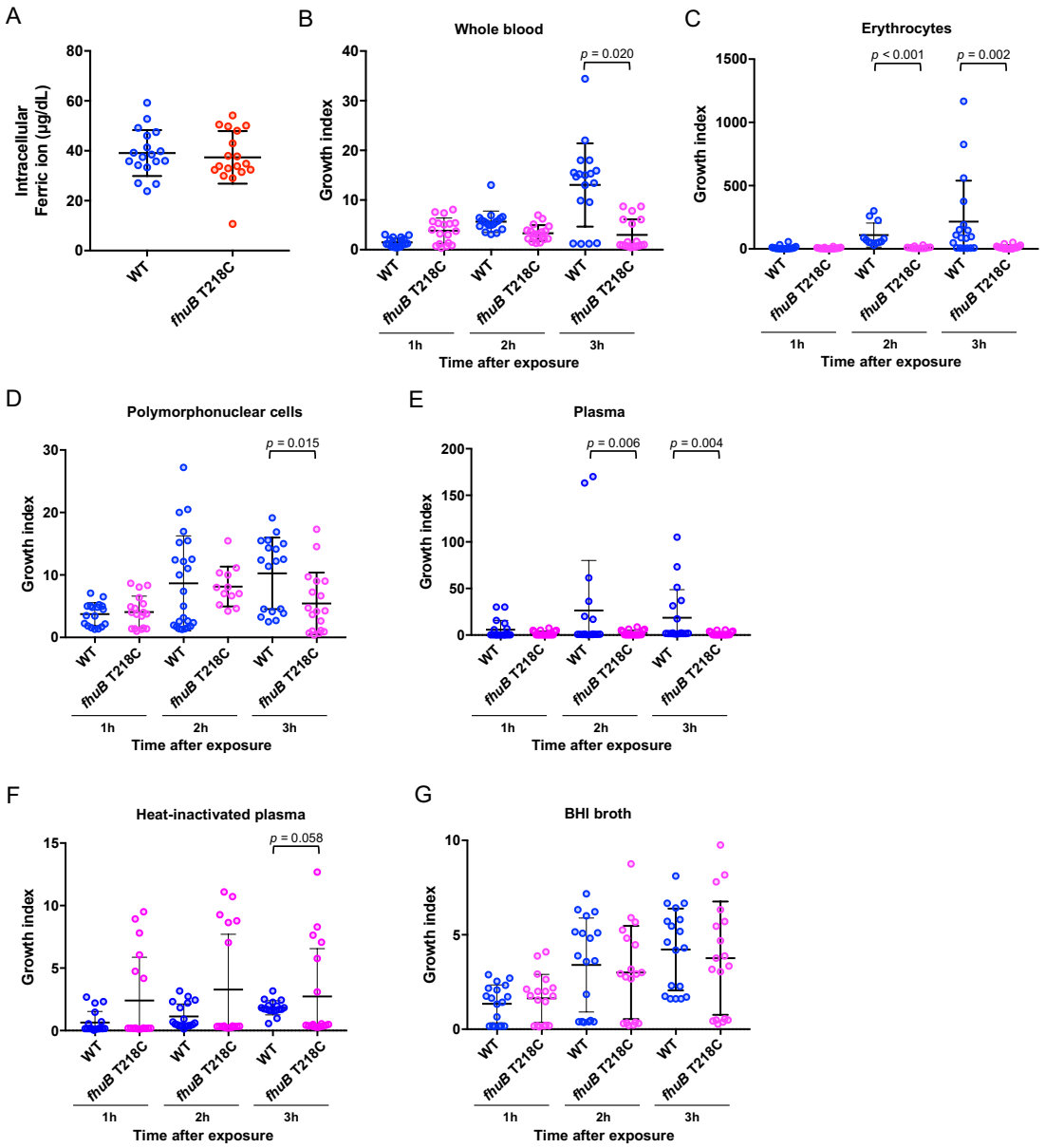


Figure 7. Effects of the *fhuB* T218C mutant strain on ferric ion uptake and bacterial survival in human blood.

(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. (B) Bactericidal assay in human blood. The WT and *fhuB* T218C mutant

1186 were mixed with healthy human blood to measure the ratio of bacterial counts at 1, 2, and
1187 3 h to those at 0 h after infection. (C–F) Bactericidal assay in human erythrocyte-rich
1188 medium (C), polymorphonuclear cell-rich medium (D), plasma (E), and plasma
1189 inactivated by heating at 56ºC for 30 min (F). (G) Bacterial growth in brain heart infusion
1190 broth. The data were pooled from three independent experiments, each performed in
1191 sextuplicate. Thick bars and error bars indicate means and quartiles, respectively.
1192 Statistical significance was determined using the Mann–Whitney test with Benjamini–
1193 Hochberg's correction.
1194

Table 1. Comparison of clustering on the phylogenetic trees for the Japanese and global cohorts.

Clusters	A	B1	B2	B3
Japanese	2	2	3	416
Global	19	2	5	639
Standard residuals	-4.33	-1.85	-2	5.14
Chi-squared test		p = 0.13		

Table 2. Complexes including k-mers significantly related to severe invasiveness detected in the Japanese cohort.

Complexes	*p* -value	*q* -value	Genes in complexes
Comp_11	4.41E-05	1.49E-02	*covS*
Comp_2	3.91E-05	4.22E-02	*group_184, group_185, group_186, group_1278, group1443, group_1519, group_1520, group_2169*

Table 3. Complexes including k-mers significantly related to invasiveness detected in the global cohort.

Complexes	*p* -value	*q* -value	Genes in complexes
Comp_7	2.35E-06	1.90E-02	*group_130, covS*
Comp_6	1.05E-05	1.90E-02	(intergenic region)
Comp_24	1.96E-06	1.90E-02	*group_141, group_142, group_142*
Comp_10	2.20E-05	2.40E-02	*sagG, sagI*
Comp_15	2.05E-05	2.40E-02	*fhuB, fhuD*

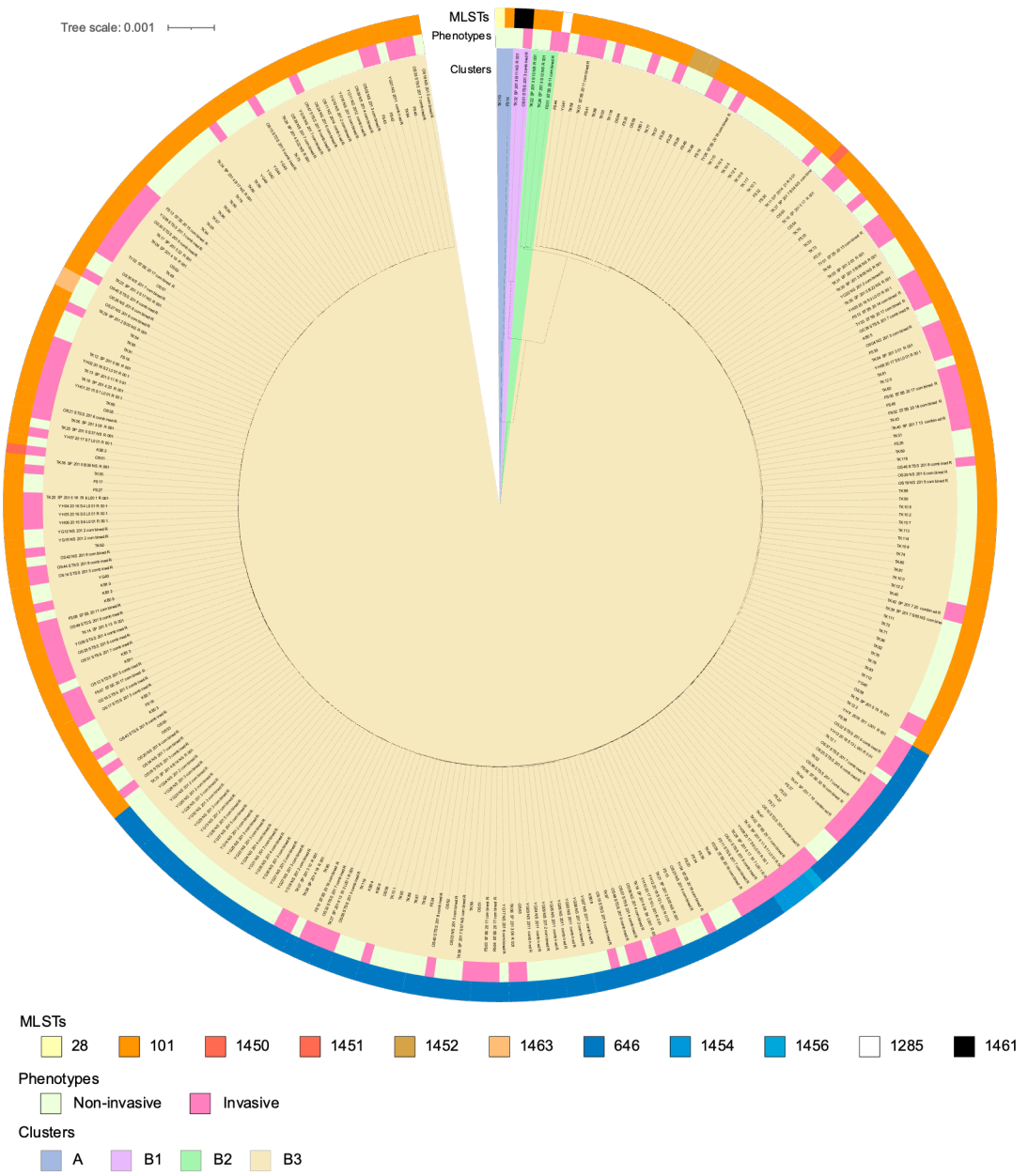


Figure S1. Phylogenetic tree for the Japanese cohort.
From the inside, color bars show clusters, phenotypes, and MLSTs, in the order mentioned. The root of the tree was set as the mid-point. The scale located in the upper left indicates 0.001 times substitution of the bases on average.

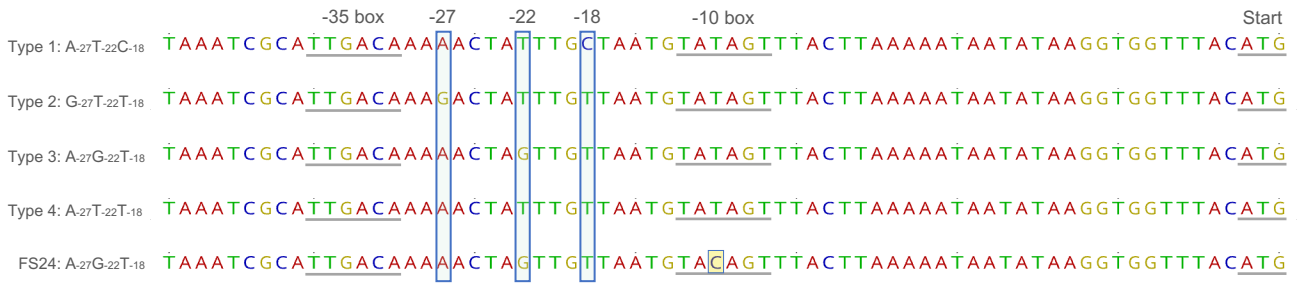


Figure S2. Variations in the *nga* promoter region.
Sequences of the *nga* promoter are classified into four major genotypes, according to three key residues (shown as -27, -22, and -18 in the figure) affecting promotor activity. Turner *et al.* further determined subtypes using mutations other than the key residues. FS24 strain has the same allele $A_{-27}G_{-22}T_{-18}$ as the type 3 or clade 3 but has an SNP highlighted in yellow in the -10 box.

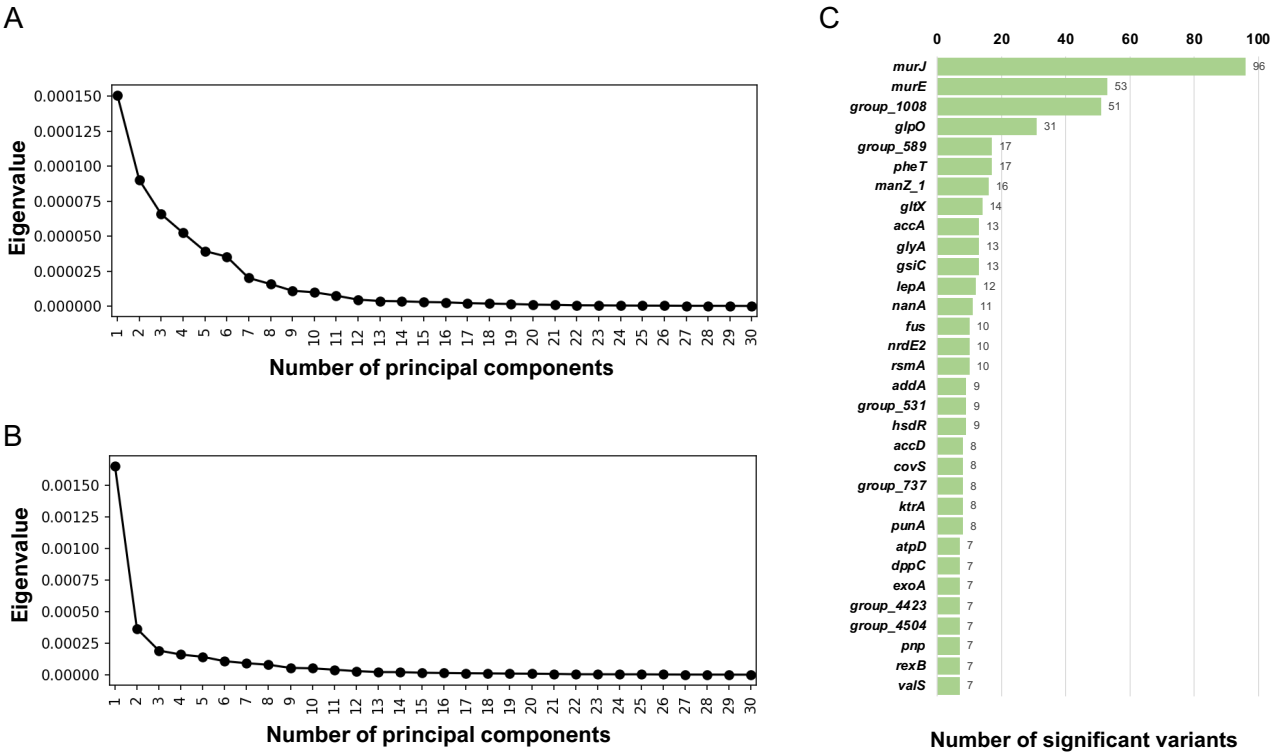


Figure S3. Pan-genome-wide association study on SNPs/indels.
(A–B) Scree plot of eigenvalues generated by decomposition of the genetic distance matrix for the Japanese (A) and global (B) cohort. To determine the number of principal components used as covariates for pan-GWAS, the plot in which the subsequent principal components seem to have relatively small eigenvalues was sought. Consequently, we used seven and three principal components as covariates for the Japanese and global cohorts, respectively. (C) Number of significant SNPs accumulated in single genes found in the pan-GWAS for the global cohort. Only the top 30 genes are shown in this graph.

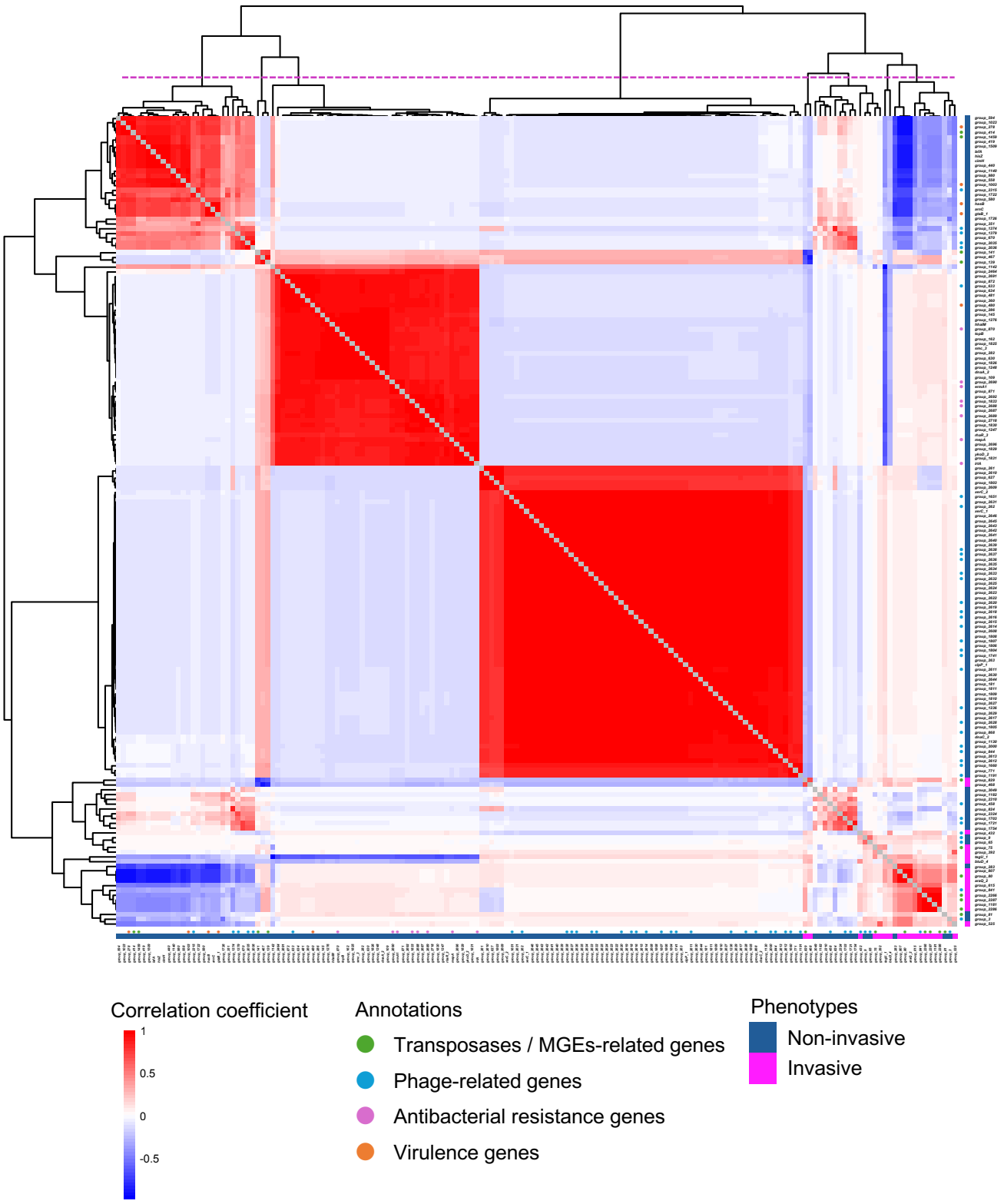


Figure S4. Pairwise correlation of significant COGs.
Each cell indicates the correlation coefficient of each significant COG pair on its presence in the global cohort. The COGs are sorted based on hierarchical clustering. Annotations and phenotypes are represented by circles and cells located on the right and bottom of the heatmap. The dotted line on the tree divides the COGs into eight clusters.

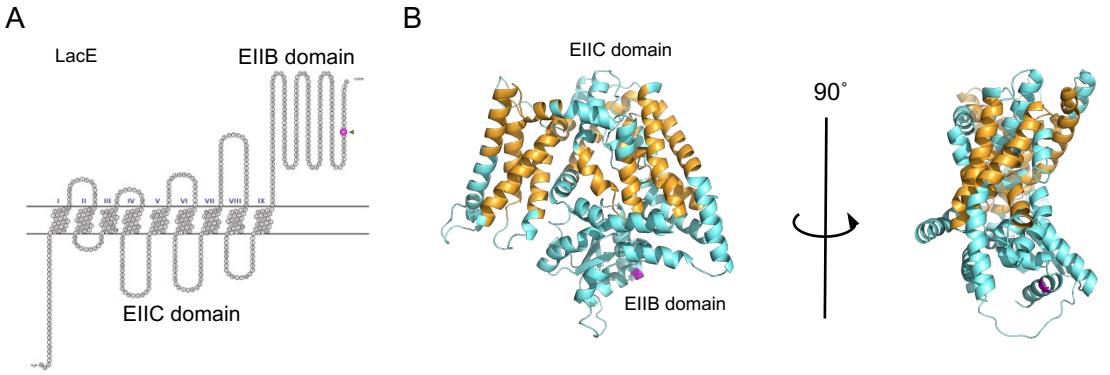


Figure S5. Predicted protein structure model of LacE.

(A) Snake-like plot of transmembrane regions of LacE estimated using SOSUI. The 554th residue has been indicated with an arrowhead and in magenta. (B) Structure model of LacE generated with AlphaFold (pLDDT=0.902). The putative transmembrane regions are colored in orange. LacE has two domains: the transmembrane domain EIIC and intracellular domain EIIB. The 554th glycine substituted with valine is colored in magenta.

pLDDT: The best predicted local difference distance test indicating local structural accuracy.

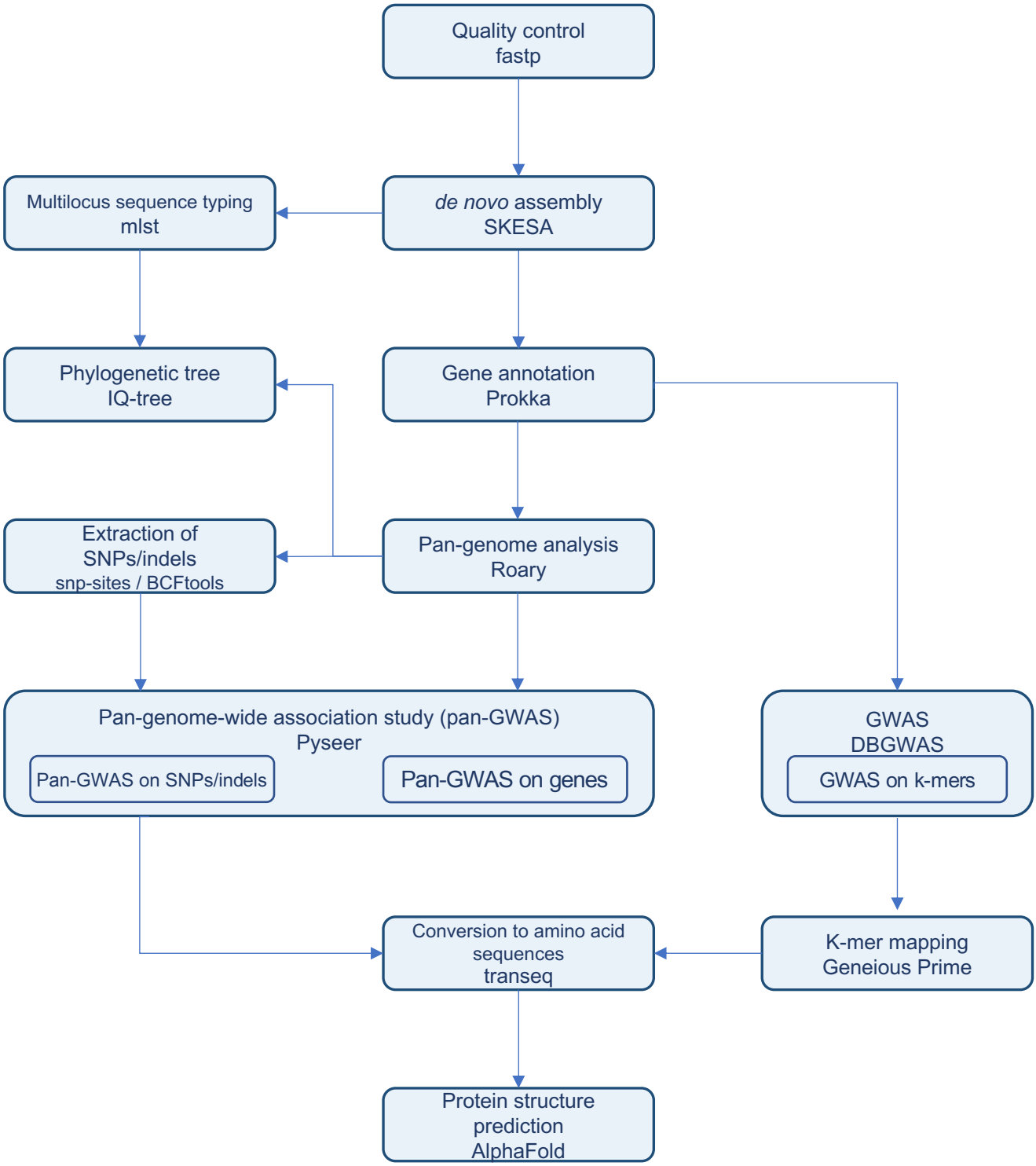


Figure S6. Pipeline for bacterial pan-genome analysis, pan-GWAS, and GWAS constructed in this study.
We developed a pipeline for whole genome analyses working on the National Institute for Genetics (NIG) Supercomputer and SQUID at the Cybermedi Center, Osaka University.