

Identifying Genetic Variations in *emm89 Streptococcus pyogenes* Associated with Severe Invasive Infections

Reviewed Preprint

v2 • May 20, 2025

Revised by authors

Reviewed Preprint

v1 • January 3, 2025

Masayuki Ono, Masaya Yamaguchi , Daisuke Motooka, Yujiro Hirose, Kotaro Higashi, Tomoko Sumitomo, Tohru Miyoshi-Akiyama, Rumi Okuno, Takahiro Yamaguchi, Ryuji Kawahara, Hitoshi Otsuka, Noriko Nakanishi, Yu Kazawa, Chikara Nakagawa, Ryo Yamaguchi, Hiroo Sakai, Yuko Matsumoto, Tadayoshi Ikebe, Shigetada Kawabata

Department of Microbiology, Graduate School of Dentistry, The University of Osaka, Osaka, Japan • Bioinformatics Research Unit, Graduate School of Dentistry, The University of Osaka, Osaka, Japan • Bioinformatics Center, Research Institute for Microbial Diseases, The University of Osaka, Osaka, Japan • Center for Infectious Diseases Education and Research, The University of Osaka, Osaka, Japan • Laboratory of Microbial Informatics, Microbial Research Center for Health and Medicine, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan • NGS Core Facility, Research Institute for Microbial Diseases, The University of Osaka, Osaka, Japan • Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives (OTRI), The University of Osaka, Osaka, Japan • Department of Oral Microbiology, Graduate School of Biomedical Sciences, Tokushima University, Tokushima, Japan • Pathogenic Microbe Laboratory, Department of Infectious Diseases, National Center for Global Health and Medicine, Tokyo, Japan • Department of Microbiology, Tokyo Metropolitan Institute of Public Health, Tokyo, Japan • Department of Bacteriology, Osaka Institute of Public Health, Osaka, Japan • Department of Public Health Sciences, Yamaguchi Prefectural Institute of Public Health and Environment, Yamaguchi, Japan • Department of Infectious Diseases, Kobe Institute of Health, Hyogo, Japan • Department of Microbiology, Fukushima Prefectural Institute for Public Health, Fukushima, Japan • Division of Microbiology, Kyoto City Institute of Health and Environmental Sciences, Kyoto, Japan • Sapporo Public Health Office, Hokkaido, Japan • Niigata City Institute of Public Health and the Environment, Niigata, Japan • Microbiological Testing and Research Division, Yokohama City Institute of Public Health, Kanagawa, Japan • Department of Bacteriology I, National Institute of Infectious Diseases, Tokyo, Japan

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

eLife Assessment

This study provides an **important** and timely analysis of invasive and non-invasive *Streptococcus pyogenes* *emm89* isolates, which have become a dominant serotype in the past decade. Using genome sequencing of 311 strains from Japan and comparing them with 666 global strains, the authors present **compelling** evidence in support of the identification of genetic factors linked to the invasive phenotype of *emm89*. The findings are both theoretically and practically significant in medical microbiology.

<https://doi.org/10.7554/eLife.101938.2.sa2>

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 *emm89 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 *emm89 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, including pharyngitis and impetigo in children, it is also known as a “flesh-eating bacterium” owing to its involvement in life-threatening invasive diseases, such as necrotizing fasciitis and streptococcal toxic shock syndrome (STSS)^{1,2}. In 2005, more than 0.6 million people were estimated to have invasive *S. pyogenes* infections³, and the reported incidence of invasive *S. pyogenes* infections continues to increase globally¹. In cases of severe infection, rapid bacterial growth and profound metabolic acidosis necessitate urgent surgical inspection and extended debridement with empiric antibacterial chemotherapy^{4,5}. However, even with proper treatment, the mortality rate of patients with STSS remains high, ranging from 23–81%⁶. Moreover, although several protective vaccine candidates against *S. pyogenes* exist, no safe and effective commercial vaccine has yet been licensed for human use^{2,7}.

S. pyogenes has been classified into at least 240 *emm* types based on a hypervariable region sequence of the *emm* gene, which encodes the M protein. This hypervariable region of the M protein is responsible for type-specific antigenicity and binds with high affinity to C4b-binding protein, a major fluid phase inhibitor of the classical and lectin pathways of the complement system that confers resistance to opsonophagocytosis⁸. Since the mid to late 2000s, *emm89* strains have been increasingly isolated from samples obtained from patients with invasive diseases, becoming one of the most frequently identified lineages in developed countries^{9,10}. For example, in the United Kingdom, no more than ten cases of invasive diseases caused by *emm89 S. pyogenes* were reported annually in the 1990s. In the 2010s, however, approximately 150 cases were reported out of the 1,000–1,500 invasive cases caused by all *emm* types of *S. pyogenes* in each year¹¹. Similarly, in Japan, in the 2010s, there was an increase in the number of patients with *emm89 S. pyogenes*-induced severe invasive infections. In 2018, *emm89* strains were isolated from 36% of patients diagnosed with *S. pyogenes*-induced severe invasive infections, following *emm1* strains, which were isolated from 57% of patients⁹.

S. pyogenes emm89 strains have been genetically sub-clustered into three clades according to the *nga* promoter region patterns and the presence/absence of the *hasABC* locus, which is responsible for hyaluronan capsule synthesis. Clade 3 is distinct from clades 1 and 2 in terms of two features: overexpression of virulence factors NAD glycohydrolase (NADase) and streptolysin O (SLO), owing to mutations in the promoter region of the *nga-ifs-slo* operon, and the lack of a hyaluronan

capsule.^{11,12,13} Although clade 3 strains have frequently been isolated from invasive diseases, their numbers from non-invasive infections have also increased. We previously reported no difference in the isolation frequencies of clade 3 strains between invasive and non-invasive diseases, at least in Japan, and concluded that the mutations in clade 3 are not responsible for the gain of invasiveness.¹⁴ Therefore, there must be other genetic features within the *emm89* strains that determine their phenotypes.

During infections caused by *emm1* *S. pyogenes*, mutations in the two-component system, CovR/S, promote high virulence.¹⁵ These mutations cause the upregulation of DNase, hyaluronan capsule, IL-8 protease, C5a peptidase, streptokinase, NADase, SLO, and superantigen SpeA, as well as the downregulation of cysteine protease SpeB and streptolysin S.^{16,17} The resulting mutants can prevent neutrophil death and subsequently promote tissue destruction and systemic infections. Epidemiologically, Ikebe *et al.* reported that nonsense mutations in *covR* and/or *covS* are present in 46.3% of *S. pyogenes* strains isolated from severe invasive infections in Japan, but only 1.69% of isolates from non-severe ones.¹⁸ Moreover, these studies indicated that the *covR/S* mutation is not responsible for all invasive clinical strains. Thus, we hypothesize that other mechanisms are involved in the development of invasiveness.

In the present study, we aimed to explore novel hypervirulent mechanisms of *S. pyogenes* by performing a pan-genome-wide association study (pan-GWAS) on *S. pyogenes*. Our pan-GWAS focused on the *emm89* lineage of *S. pyogenes* to detect lineage-specific factors and minimize false positives due to lineage differentiation. For a comprehensive analysis, we constructed the core genome and pan-genome of *emm89* *S. pyogenes* strains and evaluated the effect of single-nucleotide polymorphisms (SNPs) on core gene alignment and accessory clusters of orthologous genes (COGs). In addition, we performed a k-mers-based GWAS to detect SNPs in the intergenic regions and multiple mutations. We collected and sequenced *emm89* clinical strains isolated in Japan during 2016–2021, in addition to public *emm89* genome sequences. Using these sequences, we investigated the bacterial factors associated with severe invasive infections in Japan and globally, using pan-GWAS. Based on the bacterial protein structural predictions, we then selected candidates with high potential relevance to the phenotype. Finally, we introduced an SNP related to non-invasiveness into a clinical strain isolated from a severe invasive infection and examined the alteration of the bacterial phenotype through an *ex vivo* infection assay.

Results

Collection of *emm89* *S. pyogenes* clinical isolates in Japan and construction of cohorts

We collected clinical *S. pyogenes* strains isolated between 2016 and 2021 from patients with non-invasive and severe invasive infections in Japan. The Ministry of Health, Labour, and Welfare of Japan has defined the clinical criteria of severe invasive β -hemolytic streptococcal infections as STSS, based on the STSS 2010 Case Definition of the Centers for Disease Control and Prevention in the US, with minor modifications, including the addition of encompassing symptoms in the central nervous system (Table S1).^{19,20}

For the *emm89* clinical isolates, we collected T serotype TB3264 and untypable strains, in addition to *emm* genotype-identified strains. T-typing is a serologically-based approach that is often used as an alternative or supplement to *emm* typing. T-antigens are trypsin-resistant surface antigens exhibiting extensive antigenic diversity.²¹ Isolates of a given *emm* type frequently share the same T serotype pattern.^{22,23} The T serotype TB3264 corresponds to the genotype *emm89* or *emm94*.^{21,24} A total of 207 clinical isolates were collected with the cooperation of the National Institute of Infectious Diseases and ten public health institutes nationwide (Tables S2 and S3). We performed draft genome sequencing of the strains and identified their *emm* types. In total, 150 of

these were determined as *emm89*, followed by 24 and 19 strains as *emm4* and *emm12*, respectively. To focus on the pathogenic mechanisms underlying severe invasive infections in the *emm89* cohort, we used 150 *emm89* strains for subsequent analyses (**Figure 1A**, Tables S2, and S3). We previously determined the draft genome sequences of 161 *emm89* strains isolated in Japan between 2011 and 2019 and determined their phenotypes using the same criteria (Table S2)¹⁴. We combined these two sets and finally considered a total of 311 *emm89* strains, including 135 severe invasive and 176 non-invasive isolates, as the Japanese cohort.

We also collected public genome sequences of *emm89 S. pyogenes* strains isolated from nine countries to further characterize the genetic properties of the Japanese cohort (Table S2)^{25–27}. In this study, the phenotypes of these strains were considered invasive if the diagnoses included severe infections, STSS, invasive infections, necrotizing fasciitis, bacteremia, or sepsis, and isolation sites were described as normally sterile sites, such as the blood, brain, kidney, muscular tissue, or brain. Consequently, we identified 666 strains in the global cohort, including 420 isolates from invasive cases and 246 from non-invasive ones (Table S2).

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

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

Next, we calculated the phylogenetic relationships based on the maximum likelihood of the core gene sequences (**Figure 1C** and S2). The tree for the global cohort branched into four clusters, with clusters A, B1, B2, and B3. Cluster B3 included 640 genetically similar strains isolated mainly from Europe, North America, and Japan, whereas cluster A comprised 19 strains isolated from Oceanian countries, Kenya, Lebanon, the US, and Japan (**Figure 1C**). The phylogenetic tree for the Japanese cohort could also be clustered as in the case of the global cohort, with no significant difference in the proportions of strains classified into each sub-cluster (chi-square test, $p=0.13$; **Figure S2** and Table S4). Thus, we concluded that the overall phylogenetic features of *emm89* strains were distributed similarly in Japan and other areas, especially Europe and North America. Within cluster B3, we identified a non-invasive strain from Japan that had no identical pattern to the reported *nga* promoter variations (**Figure 1C**)²⁸. This pattern is likely a subtype of clade 3 as it shares the haplotype A–27G–22T–18, which is distinctive of clade 3, but has a mutation in the –10 box (**Figure S2**)²⁸. Thus, we named this novel *nga* promoter variation type 3.4²⁸. Multilocus sequence typing (MLST) analysis revealed that in cluster B3, 522 strains (80.9%) were ST101, and 96 strains (14.8%) were ST646 (**Figure 1C**). Notably, ST646 was the second most dominant type and a Japan-specific lineage. Moreover, they only differed in the 295th nucleotide in the *murI* locus, one of the seven loci that determine MLST, suggesting that both lineages have a genetically close relationship. Eight novel MLSTs were determined (ST1450, 1451, 1452, 1454, 1455, 1456, 1461, and 1463) and 15 novel MLST strains were detected in the Japanese cohort (**Figure S1**). Taken together, using phylogenetic approaches, we found that most strains from Japan and countries in Europe and North America share genetically close relationships, with only one unique lineage in Japan, ST646.

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

To discover all types of genetic variants in whole genes within *emm89 S. pyogenes* associated with (severe) invasiveness, we applied pan-genome analysis and performed three types of independent analyses, including pan-GWASes targeting SNPs in core genes and the presence or absence of all

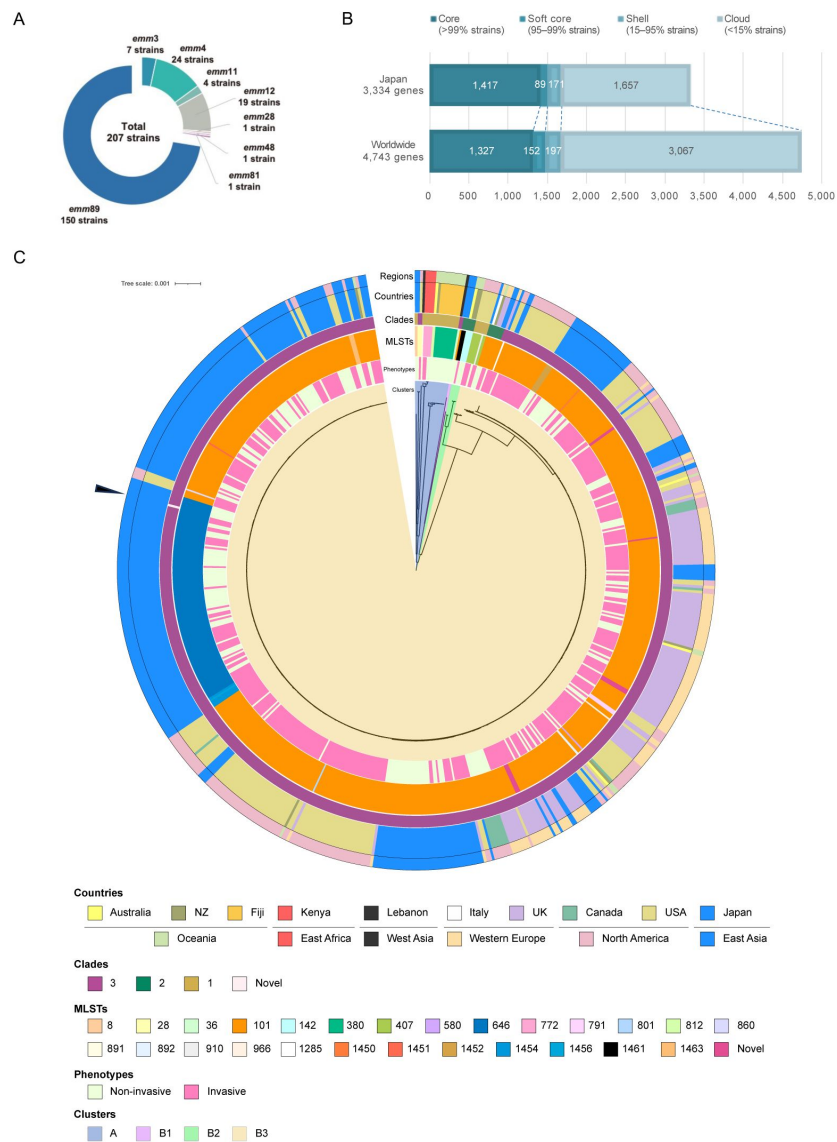


Figure 1.

Pan-genomic and phylogenetic characterization of the collected *emm89 S. pyogenes* isolates.

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

MLST, multilocus sequence typing.

genes, and a GWAS targeting other variants located in intergenic regions spanning several nucleotides. The present pan-GWAS of bacteria within a single *emm* type minimized lineage effects, thus reducing false positives.

We extracted SNPs and single-nucleotide indels from core gene alignments and detected 24,627 and 47,060 SNPs/indels in the Japanese and global cohorts, respectively. Subsequent pan-GWASes identified SNPs/indels associated with severe invasiveness in a Japanese cohort and invasiveness in a global cohort. To control for population bias, we calculated pairwise distance matrices and selected seven and three dimensions for the analyses of the Japanese and global cohorts, respectively (**Figure S3A** and S3B). For each cohort, we performed a permutation test by conducting 1,000 iterations of calculations with randomly permuted phenotypes, with the significance level set at the 5th percentile of the 1,000 minimal *p*-values ($p=5.75\times10^{-4}$ and $p=5.75\times10^{-4}$ for the Japanese and global cohorts, respectively). The pan-GWAS of the Japanese cohort detected 17 SNPs/indels in 13 core genes (**Figure 2A** and Table S5). Of the 17 significant variants, there were 7 single-nucleotide deletions (SNDs), 7 SNPs causing non-synonymous amino acid substitutions, and 3 SNPs causing synonymous substitutions. The *covS* gene (also known as *csrS*), encoding a sensor kinase of the two-component system CovR/S, contains four SNDs with the lowest *p*-values ($p=1.16\times10^{-7}$ for the 39th, 40th, and 46th nucleotides, and $p=1.15\times10^{-6}$ for the 125th nucleotide). These four deletions were associated with severe invasive infections.

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

Moreover, 19 SNPs induced nonsense mutations, whereas the effects of 12 SNPs/indels were unpredictable because of a lack of reference sequences (Table S3). Notably, 96 SNPs/indels accumulated in a single gene, *murJ*, which is involved in peptidoglycan biosynthesis, whereas 53 and 51 SNPs/indels were detected in *murE* and *group_1008*, respectively (**Figure S3C**). The SNP with the lowest *p*-value ($p=1.35\times10^{-14}$) was *lacE*, which encodes the EIICB component of the lactose-specific phosphotransferase system (**Figure 2C**). This mutation was found to be associated with an invasive phenotype and mainly observed in strains isolated in the US. Compared with the significant 17 SNPs/indels in the Japanese cohort, 10 SNPs/indels were also detected in the global cohort, including 4 SNDs in *covS* and 1 SNP each in 6 loci (**Figure 2C**). Deletions at the *covS* locus were common among strains from several countries, including Japan. In contrast, SNPs in six loci, *gataA*, *group_1102*, *group_647*, *iscS_1*, *recU*, and *fhuB*, were present exclusively in Japan (**Figure 2C** and Table S6). These results suggest that several bacterial mechanisms cause severe invasive *S. pyogenes* infections, and some prevail worldwide, such as *covS* mutations, whereas others are specific to Japan.

Pan-GWAS on COGs reveals 2 and 109 genes associated with severe invasiveness in the Japanese cohort and invasiveness in the global cohort, respectively

Next, we examined the associations of accessory COGs with severe invasiveness and global invasiveness in Japanese patients. A permutation test determined significance levels as *p*-values of 1.09×10^{-4} and 7.72×10^{-5} for the Japanese and global cohorts, respectively. Two significant genes were detected in the pan-GWAS for the Japanese cohort: *group_184*, which encodes a hypothetical protein, and *divIC*, which encodes a septum formation initiator protein ($p=8.81\times10^{-6}$ and $p=6.72\times10^{-6}$, respectively; **Figure 3A** and Table S7). Although analysis of the global cohort revealed the presence of 169 genes that were significantly related to invasiveness, no genes were identical or homologous to the two genes detected in the Japanese cohort (**Figure 3B**, 3C, and Table S8). Approximately 90% (152 of 169) of the significant COGs were found to be associated with non-invasiveness. Among the 169 genes, 37 encoded phage-related genes and 11 encoded mobile

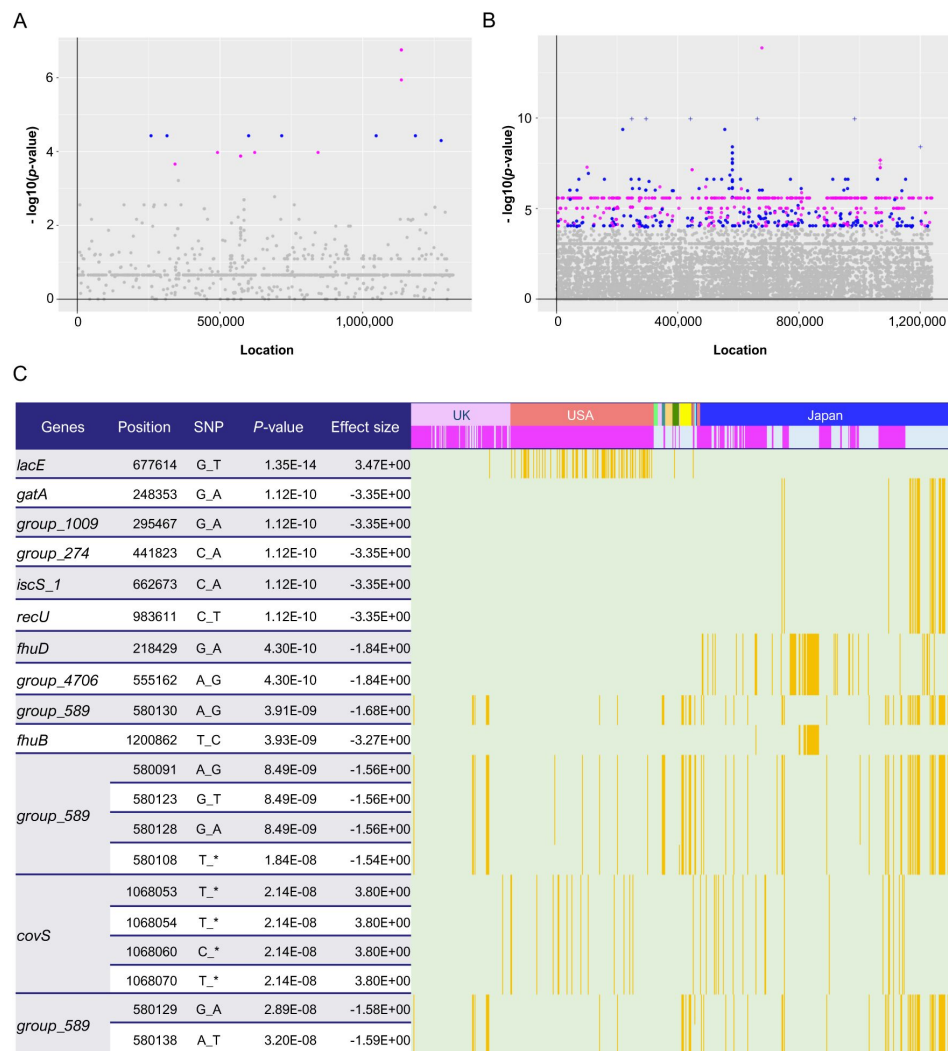


Figure 2.

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 percentile of the 1,000 minimal *p*-values ($p=5.75 \times 10^{-4}$ and $p=1.05 \times 10^{-4}$ for the Japanese and global cohorts, respectively). Plots with lower *p*-values than genome-wide significant levels are colored magenta and blue, based on the direction of their effect size (positive and negative, respectively). (C) Distribution heatmap for the global cohort, with the strains possessing the significant SNPs/indels colored orange. Only the 20 SNPs with the lowest *p*-values are shown in this heatmap. Colored bars above indicate countries and phenotypes, and magenta bars represent invasive phenotypes. Using the Roary program, gene names starting with "Group_" were automatically assigned. Position indicates the location of each SNP/indel on the core gene alignment. The full results are shown in Table S6. SNP, single-nucleotide polymorphism; indel, insertion/deletion.

genetic elements (MGEs) such as transposase, integrase, and recombinase. The gene with the lowest *p*-value was *group_829*, which encodes a transposase and is related to the invasive phenotype (**Figure 3C** and Table S8). In addition, *group_2689*, which encodes a multidrug efflux transporter permease, *rhaR_3*, which encodes a transcriptional regulator involved in rhamnose metabolism, and *group_1829*, which encodes a vitamin B12 import transporter permease, were also substantially associated with the invasive phenotype. We identified several gene distribution patterns associated with invasiveness, suggesting that multiple independent genetic factors cause invasive infections (**Figure 3C**). On the other hand, the present pan-GWAS found no genes encoding known virulence factors significantly associated with invasiveness, thus further analysis of the relationships of detected distribution patterns with prophages and MGEs was performed. For calculating the pairwise correlation of the presence of significant COGs in the 666 strains, the COGs were clustered into eight coexisting groups, seven of which contained phage- and/or MGEs-related genes (**Figure S4**). The largest group comprised 65 genes including phage proteins, while the second largest with 42 genes was found to be associated with non-invasive infections, and included *group_2689*, *group_1833*, and *ermA1*, encoding TetR/AcrR family transcriptional regulator, multidrug efflux system permease protein, and rRNA adenine N-6-methyltransferase, respectively. Taken together, genes associated with invasiveness were found to encode mobile genetic factors and transporters, whereas major virulence factors were not significantly associated with invasiveness.

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

To detect SNPs/indels and multiple mutations in the entire genome, we extracted 31-nt-length k-mers from whole genomes and performed a GWAS. The k-mers-based GWAS can handle polymorphisms spanning more than one base, such as indels, inversions, and translocations, in both the coding and non-coding regions.

In the Japanese cohort, the k-mers-based GWAS detected two regions containing causative variants associated with severe invasiveness (Table S9). As shown in the de-Bruijn graphs, overlapping k-mers were concatenated into a single node, and edges represent variability among the sequences (**Figure 4A-F**). The set of connected nodes and edges comprising each de-Bruijn graph is called a complex. Nodes were determined to be significant if their *q*-value was less than 0.05. Significant nodes are shown in red and blue, indicating an association with severe invasive or non-invasive infections, respectively. The complex comprising the nodes with the lowest *q*-value ($q=1.49 \times 10^{-2}$) was Comp_11 in the *covS* locus (Table S9). The causative mutations in *covS* were an SND and 10 nucleotide polymorphisms (**Figure 4A**). This deletion resulted in a frameshift mutation that shortened the length of *CovS* from 500 to 35 amino acids, leading to increased invasiveness, as previously reported in other *emm* types¹⁷. Another complex significantly associated with severe invasiveness is Comp_2 ($q=4.22 \times 10^{-2}$; Table S9). Comp_2 is a highly variable region containing eight hypothetical protein-coding genes, with high similarity within the first 75 bp. Significant k-mers were also mapped to the first 26 bp of *group_184* and 20 bp upstream (**Figure 4B**). These findings suggest that *group_184* possibly contributes to severe invasiveness through not only its presence but also by that of the upstream region.

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

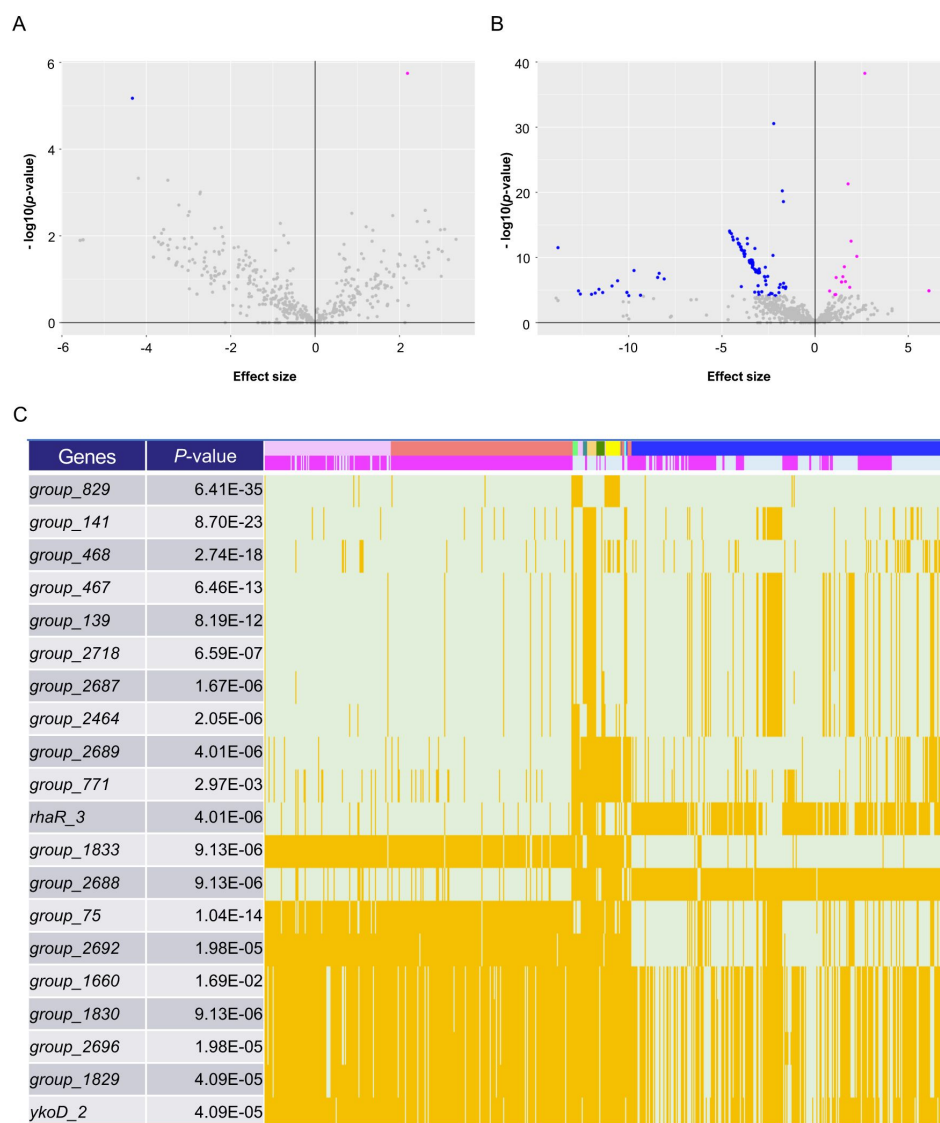


Figure 3.

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 ($p=1.09\times 10^{-4}$ and $p=7.72\times 10^{-5}$ for the Japanese and global cohorts, respectively) have been colored magenta and blue, based on the direction of their effect size (positive and negative, respectively). (C) Distribution heatmap for the global cohort, with the strains possessing the significant genes colored orange. Only the 20 genes with the lowest p -values are shown in this heatmap. Colored bars above indicate countries and phenotypes, and magenta bars represent invasive phenotypes. Using the Roary program, gene names starting with “Group_” were automatically assigned. The full results are shown in Table S8.

Two significant k-mers were present in Comp_6, which were found to be an intergenic region of 270 bp ($q=1.90\times 10^{-2}$; Table S10 and **Figure 4C**). We speculated that this region is related to regulation of gene expression. However, no promoter sequences were identified by utilizing MLDSPP, BacPP, and BLAST, thus the significance of this region remains to be clarified^{29,30}. In Comp_24, with a high sequence variation containing *group_141*, *group_142*, and *group_143*, which encode transposases, the presence of a 281-bp sequence consisting of several k-mers was significantly associated with the invasive phenotype ($q=1.90\times 10^{-2}$; Table S10 and **Figure 4D**). Furthermore, *group_141* was also identified in the COGs-based pan-GWAS as a non-invasiveness-related gene along with *group_139* and *group_467*, which encode transposase and uncharacterized protein, respectively (Table S8 and **Figure S4**). Taken together, the absence of an MGE containing *group_141*, and the presence of another MGE harboring *group_142* and *group_143* may result in an invasive phenotype. In Comp_10, n27458, the *sagG* locus encoding the ATP-binding protein of the efflux transporter of SLS, was significantly correlated with the non-invasive phenotype ($q=2.40\times 10^{-2}$; Table S10 and **Figure 4E**). However, a significant SNP in n27458, *sagG* (A882G), was found to cause a synonymous mutation. The other significant mutation was present in the *fhuB* locus, encoding a putative ferrichrome transport system permease ($q=2.40\times 10^{-2}$; Table S10). This mutation was identical to SNP T218C detected in the SNP/indels-based pan-GWAS (**Figure 4F**). Moreover, this mutation changes the 73rd residue from valine to alanine in FhuB, which is a putative ferrichrome transport system permease.

Therefore, the k-mer approach identified multiple variants, including the mutation identified in the SNPs/indels-based pan-GWAS, *fhuB* SNP T218C. In addition, while the mutation detected in *covS* differed from that detected in the SNPs/indels-based pan-GWAS, both caused frameshift mutations.

AlphaFold-based prediction of the impact of the identified mutations on function

To assess the impact of mutations on protein function, we predicted the protein structure using AlphaFold³¹. Here, we present structural predictions for three representative proteins: LacE, whose mutation was observed mainly in invasive strains from the US; CovS, whose invasion-related deletions prevailed worldwide; and FhuB, which carries a prominent mutation in the Japanese cohort and is associated with non-invasiveness.

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

Next, we predicted a homodimerized CovS model using AlphaFold, as the CovS of *S. pyogenes* forms homodimers³². SOSUI predicted that CovS has two transmembrane regions (**Figure 5A**). Mutations detected in the SNP/indels- and k-mers-based GWASes were predicted to shorten the CovS protein to 35 and 45 amino acids, respectively. As the intracellular domain of CovS is in the C-terminal region and is involved in the phosphorylation of the transcriptional regulator CovR, frameshift mutations leading to CovS truncation would inactivate the protein, and thus, CovR function (**Figure 5B**).

As described above, the SNP T218C in the *fhuB* locus substitutes the 73rd valine of FhuB with alanine. FhuB is a component of an ATP-binding cassette transporter system that utilizes ferrichrome, which is one of the carriers of Fe³⁺. FhuB is predicted to localize to the cell membrane and form a channel with FhuG (**Figure 5C**). SOSUI suggested that FhuB and FhuG are 9-transmembrane proteins. The FhuBG complex can bind to one molecule of the extracellular

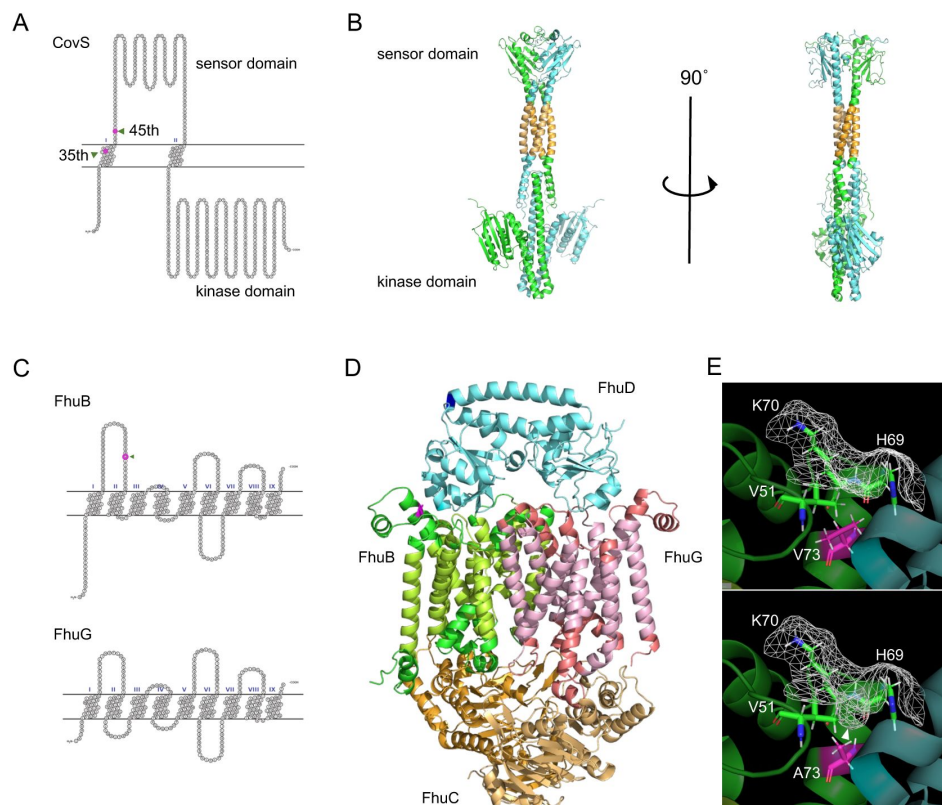


Figure 5.

Predicted protein structure models.

(A) Snake-like plot of transmembrane regions of CovS estimated using SOSUI. Frameshift mutations, detected by the SNPs/indel- and k-mer-based GWASes, cause truncation of CovS at the indicated 35th and 45th residues, respectively. (B) Structural model of the CovS homodimer (ipTM + pTM=0.614). The putative transmembrane regions are colored orange. The upper part is the sensor domain, while the lower is the C-terminal kinase domain involved in the phosphorylation of the transcriptional regulator CovR. (C) Snake-like plot depicting the transmembrane regions of FhuB and FhuG. The 73rd residue of FhuB is indicated by an arrowhead and in magenta. (D) Structural model of the FhuBDCCG complex (ipTM + pTM=0.791). The putative transmembrane regions of FhuB and FhuG are colored green-yellow and peach, respectively. The upper part of the model is located in the extracellular region, while the lower part is in the cytoplasm. (E) The 73rd valine in FhuB, shown in magenta, was substituted with alanine. The molecular surface is illustrated with a wireframe and that of the predicted indentation is shown with an arrowhead. ipTM + pTM: Weighted combination of interface-predicted TM and predicted TM scores. ipTM is used to measure structural accuracy in the protein-protein interface, while pTM is a metric for overall topological accuracy.

ferrichrome-binding lipoprotein, FhuD, and two molecules of the intracellular ATP-binding protein, FhuC. Therefore, we constructed a structural model of the FhuBCDG complex (**Figure 5D** [↗](#)), which implied that the 73rd residue of FhuB exists in a region adjacent to FhuD. The mutation was predicted to induce formation of a small indentation on the molecular surface, thus increasing the surface area accessible to the solvent, and is considered to potentially affect the stability of the hydrophobic bond between FhuB and FhuD, and thus ferrichrome transport (**Figure 5E** [↗](#)). The SNP G538A in *fhuD* was Japan-specific, significantly related to severe invasiveness, and caused the V180I mutation in FhuD (**Figure S3C** [↗](#)). The prediction suggested that the 180th residue is in an α -helix distant from the active site or the interactive sites with FhuB. Both valine and isoleucine are branched-chain amino acids, and the amino acid residue is located on a rigid structure, the α -helix (**Figure 5D** [↗](#)). Thus, we believe that this mutation is less likely to cause structural changes than the other non-synonymous mutations.

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

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

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

CovR has also been reported to regulate *sda1*, which plays a key role in invasive disease progression in *emm1* strains; however, *sda1* does not exist in the *emm89* reference strain MGAS27061 or WT strain¹⁵ [↗](#). In human blood, the mutant strain resulted in the downregulation of *mga* and *emm* expression and upregulation of the expression of genes encoding virulence factors, such as *speC*, *scpA/B*, *endoS*, *fba*, *ska*, and *sfbX*, compared to those observed in the WT strain (Table S14). Although Mga regulates the expression of surface and secreted molecules important for colonization and immune evasion³³ [↗](#), no strong expression changes were observed in the Mga regulon, except in the *emm* gene. Notably, *fhuB*, *fhuC*, and *fhuD* were upregulated in the *fhuB* mutant in both human blood and THY (Tables S11 and S14). Moreover, the transcriptomic analysis revealed that both WT and *fhuB* T218C strains displayed upregulation of the *speB*, *nga-ifs-slo* operon, and *sag* operons in human blood. In addition, the *fhuB* T218C mutation caused the upregulation of *fhuBCD* expression. To determine whether *fhuB* T218C mutation affects iron transport, we measured the intracellular free ferric ion concentration in each strain and observed no differences among the strains in either environment (**Figure 7A** [↗](#)).

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

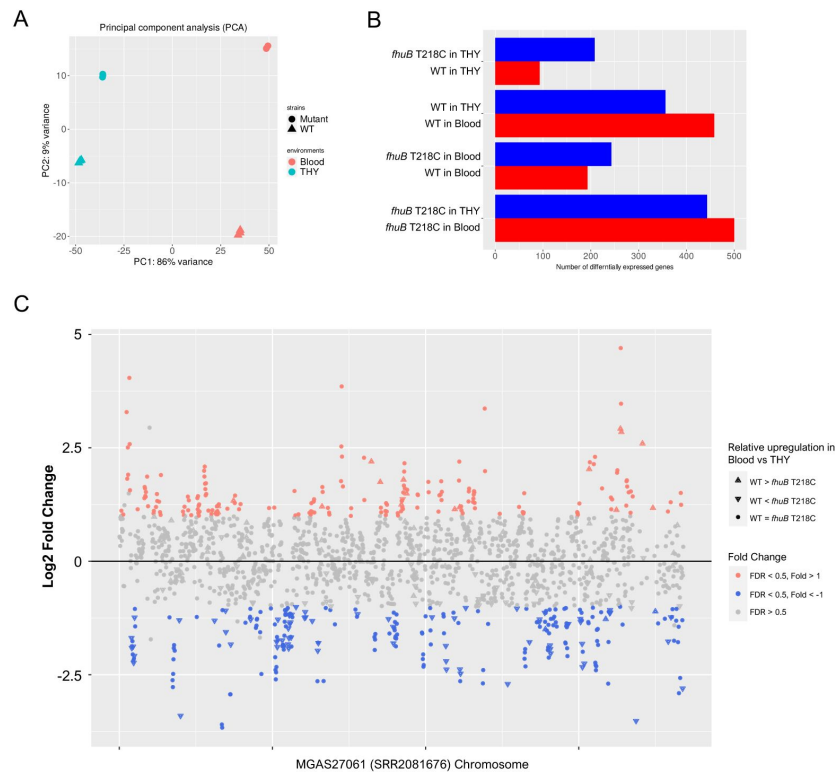


Figure 6.

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 blood vs. THY. The downward triangle plots indicate genes that are either downregulated in the WT or upregulated in the mutant strain, in blood vs. THY. WT, wild-type; THY, Todd Hewitt broth supplemented with 0.2% yeast extract.

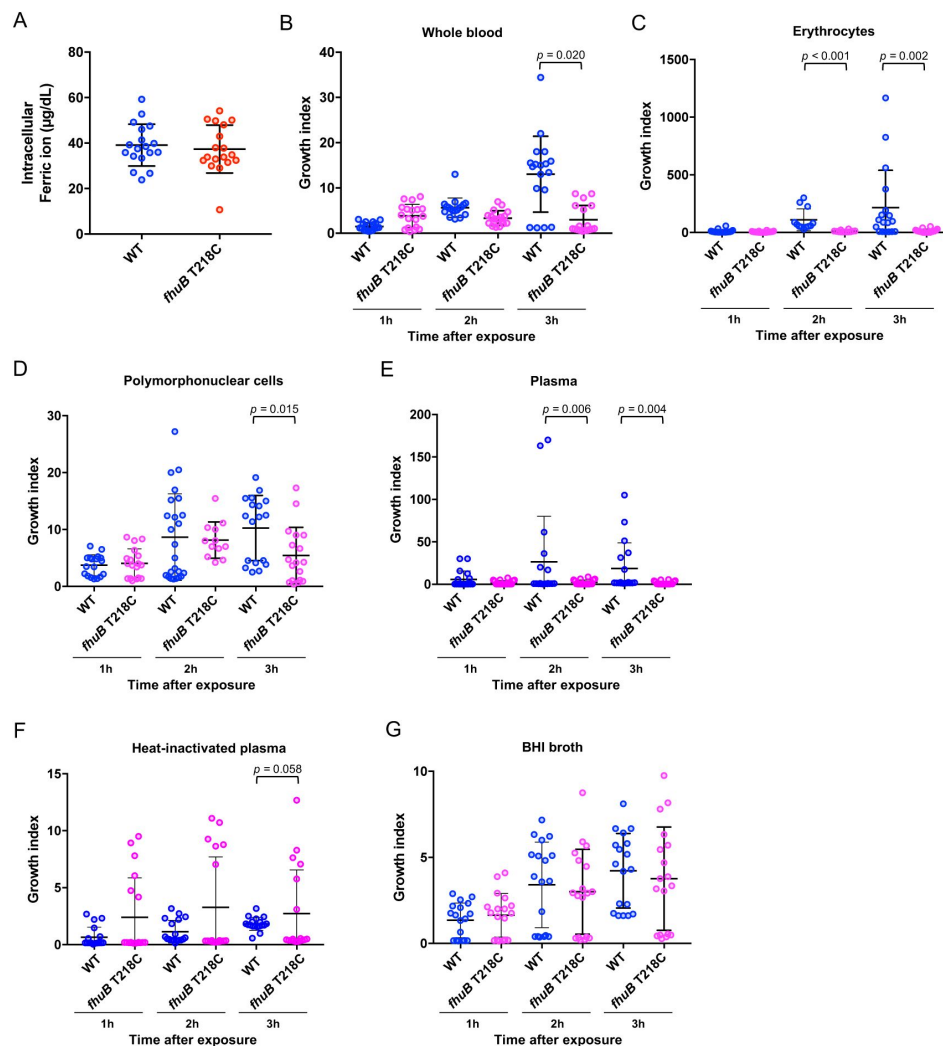


Figure 7.

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

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

Discussion

The present study focused on *emm89 S. pyogenes*, known to cause increasing rates of invasive infections worldwide, and also assessed differences between *emm89* strains causing invasive and non-invasive infections. By focusing on bacteria within closed phylogenies, false positives caused by confounding lineage effects were minimized, thus contributing to a higher level of accuracy of the pan-GWAS. We constructed a workflow for bacterial pan-GWAS and GWAS and explored the bacterial genetic factors related to severe or invasive infections to reveal the plausible mechanisms of pathogenesis. We independently performed pan-GWASes and GWASes for Japanese and global cohorts to investigate the associations of variants with strictly defined phenotypes in severe invasive *S. pyogenes* infections and with phenotypes in a broader context of invasive infections.

Several GWASes have been conducted on *S. pyogenes* to date. Davies *et al.* extracted k-mers from the genomes of 1,944 strains of *S. pyogenes* with broad *emm* genotypes and assessed their associations with invasiveness, and Kachroo *et al.* performed GWAS on k-mers using 442 *emm28* strains isolated from either invasive or non-invasive infections^{27,34}. However, no studies have described the combined analysis of SNPs, genes, and k-mers exclusively using *emm89* strains that share genetically close genomes and, thus, a large number of core genes. The investigations we performed led to the discovery of a tremendous number of causative variants, including not only SNPs in core genes but also accessory genes and insertions spanning intergenic regions.

Spontaneous mutations in *covR/S* genes potentiate the transition from localized to invasive infection by M1T1 *S. pyogenes*¹⁷. We demonstrated that some SNDs in the *covS* locus were significantly related to the phenotypes in both the SNPs/indels- and k-mers- based GWASes. CovS forms a two-component system with CovR that regulates the transcription of operons encoding virulence factors when phosphorylated by the histidine sensor kinase CovS³⁵. Sumby *et al.* suggested that systemic invasive infections are caused by the overexpression of virulence factors through the inactivation of CovS, as a result of the introduction of a depletive mutation in the *covS* locus after infection^{15,16}. Several mutation sites causing such overexpression, such as the 7 nt insertion in *covS* of *emm1 S. pyogenes* and the insertion of adenine at the 877th nucleotide, have been reported^{15,16}. Ikebe *et al.* demonstrated variations in the lengths of mutant CovS proteins by analyzing the gene sequences of 164 STSS strains in Japan¹⁸. In line with their study, we found that the shortening of CovS that potentially occurred owing to frameshifts caused by SNDs is significantly associated with invasive infections, potentially contributing to the upregulation of virulence factors and invasiveness.

In contrast, we observed several factors associated with invasiveness to be independent of *covR/S* mutations. Based on conformational predictions, we selected SNPs with non-synonymous substitutions that were likely to affect protein function. Transcriptomic analysis findings suggested that the Japan-specific *fhuB* mutation associated with non-severe invasive infections contributes to the growth rate of *S. pyogenes* in human blood by adapting to the environment. In addition, *emm89* clade 3 carries the identical promoter region pattern of the *nga* operon as *emm1* strains, and the pattern conferred similarly high expression of *nga* and *slo*^{11,12,28}. Our RNA-seq data demonstrated that a severe invasive Japanese strain without *covS* mutations increased the

expression of *speB*, *nga*, *slo*, and genes in the *sag* operon in the blood. Although *covR/S* mutations downregulate the expression of *SpeB* and *SLS*, *SpeB* and *SLS* act as virulence factors, allowing *S. pyogenes* to invade host tissues^{36–39}. These data suggest that severe invasive infections have multiple gene expression profiles in addition to *covR/S* mutation-induced profiles within a single lineage, *emm89*, and that the synergy between optimizing bacterial survival in human blood and upregulating multiple virulence factors contributes to the development of severe invasive infections.

Two possible roles of the *FhuB* mutation in the pathogenesis of severe invasive infections are thus proposed. First, the mutation in *FhuB* could increase bacterial susceptibility to free radicals generated in the presence of ferric ions in the blood. Generally, ferric ions are essential for the survival of almost all living organisms. Catalyzed by ferric ions, the Fenton and Haber–Weiss reactions generate hydroxyl radicals in erythrocytes⁴⁰. Previously, we provided evidence that the iron in erythrocytes partially inhibits pneumococcal growth via a free radical-based mechanism⁷. Although no significant difference was observed in the intracellular ferric ion concentration in the present study, the structural changes in *FhuB* may have caused an increased generation of free radicals in bacterial cells and the prevention of bacterial survival in human blood. Second, the *FhuB* mutation could affect bacterial transcriptional profiles and result in reduced fitness for survival in human blood. We also observed a lower survival rate of the mutant strain with polymorphonuclear cells (**Figure 7D**). Lower *emm* transcription in the mutant may lead to attenuated immune evasion. The M protein can bind C4-binding protein and factor H and inactivate the deposited C4b and C3b, leading to limited surface opsonization⁴¹. Furthermore, complements could have inhibited the proliferation of the mutant strain, as suggested by the activation of the membrane attack complex (**Figure 7D** and **7E**). Although the mechanisms of the *FhuB* V73A mutation in the interaction with each blood component must be further elucidated, our results indicated that the 73rd residue of *FhuB* is a key factor for bacterial survival in human blood during systemic infection.

Recently, He *et al.* reported that antibodies against the ferrichrome-binding protein FtsB, also known as *FhuD*, decrease blood bacterial burden and skin abscess formation in murine models infected with *emm1 S. pyogenes*⁴². *FhuD* is a lipoprotein that has been studied as a potential vaccine candidate against *Staphylococcus aureus*⁴³. The detailed function and effect of the *FhuBCCDG* complex in *S. pyogenes* remain unknown; however, at least in *emm1* strains, the complex can be expressed and localized abundantly on the cell membrane. Hence, this complex may be a promising target for vaccines against *S. pyogenes*.

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

In the global cohort, 169 genes were related to the phenotype, including 37 phage- and 11 MGE-related genes. The genome of *S. pyogenes* is rich in prophages, phage-like chromosomal islands, and MGEs⁴⁴. Prophages and MGEs sometimes function as vectors of virulence factors and antibiotic-resistance genes through interspecies and intraspecific transmission. Common superantigen-coding genes, such as *speA*, *speC*, *speG*, *speH*, *speI*, *speJ*, *speK*, and *speL*, and DNase-coding genes, such as *spd1*, are exogenous genes derived from prophages⁴⁵. In addition, in the genome of *S. pyogenes*, there are broad MGEs where *erm* and *mefA* related to macrolide resistance and *tetM* involved in resistance to tetracyclines are located⁴⁶. In the present study, we observed no direct association between invasiveness and virulence of genes located in the prophages and MGEs. On the other hand, a cluster comprising 49 non-invasiveness-associated genes including

antibiotic-resistance genes was identified. Furthermore, among the genes showing a significant correlation with the infectious phenotype, approximately 90% (152 of 169) were associated with non-invasiveness. One possible explanation is that significantly related genes reflect the process of not only gain of factors but also loss of those affecting fitness cost. Given that we also identified hypothetical genes associated with this phenotype, we assumed that there is an unknown mechanism contributing to the pathogenesis of invasive infections.

In the present study, MLST analysis revealed that 522 of the 666 strains belonged to ST101. Following ST101, ST646 was the second most prevalent lineage in the Japanese cohort and was not detected in any other country. In addition, the difference between ST101 and ST646 was found only at the 295th nucleotide in *mull*, among the seven genes subjected to MLST typing. Our phylogenetic trees also suggested that ST101 and ST646 have a close phylogenetic relationship. Five previous studies have described ST646 in *S. pyogenes*, and all of them reported *emm89* ST646 strains isolated in Japan⁴⁷⁴⁸⁴⁹⁵⁰⁵¹. Furthermore, all ST646 strains were isolated from the 4,249 *emm89* *S. pyogenes* strains in the MLST database PubMLST⁵². Therefore, ST646 appears to be specific to *emm89* strains and a unique lineage in Japan. Notably, Ubukata *et al.* also reported that ST646 strains began increasing after 2012, whereas the ST101 lineage was dominant until 2008⁵¹. Taken together, we conclude that ST646 is possibly a relatively new and Japan-specific lineage of *emm89* *S. pyogenes*.

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

In this study, we revealed the genotype-phenotype associations found in not only the known factors represented by *covS* but also factors that are related to invasiveness, including *fhuB*. Moreover, we experimentally validated the contribution of the *fhuB* mutation to bacterial survival in human blood. This study demonstrates the potential of our genomic statistical approach for elucidating the pathogenesis of invasive infections. Further analyses of the invasiveness-related factors identified in this study could provide a platform for establishing novel treatments and preventive strategies against invasive infections.

Materials and methods

Clinical isolates in Japan

Clinical isolates were collected from public health institutions in Tokyo, Osaka, Yamaguchi, Fukushima, Kobe, Kyoto, Amagasaki, Sapporo, and Niigata, Japan. We defined the strains collected as STSS according to the Infectious Diseases Control Law in Japan. Non-invasive strains were defined based on diagnostic names, including asymptomatic, pharyngitis, tonsillitis, or non-invasive infections. Strains with no diagnostic names for the non-STSS strains were defined as non-invasive based on the isolate sites. Information on all the strains included in this study is presented in Table S3. The collected *S. pyogenes* strains were cultured at 37°C in an atmosphere containing 5% CO₂, in Todd Hewitt broth supplemented with 0.2% yeast extract (THY; both from BD Biosciences, Franklin Lakes, NJ, USA) and stored in THY broth with 30 % glycerol (Nacalai Tesque, Kyoto, Japan), at -80°C.

Genomic DNA sequencing of the clinical isolates

The *S. pyogenes* strains were cultured until the exponential growth phase (OD₆₀₀=0.3–0.4). Bacterial cells were lysed with T10E1N100 buffer (10 mM Tris-HCl buffer, 100 mM sodium chloride, and 1 mM EDTA), 10 units/mL mutanolysin (Sigma-Aldrich, St. Louis, MO, USA), 10 mg/mL lysozyme (Fujifilm Wako Pure Chemical Co., Osaka, Japan), 0.5 mg/mL achromopeptidase (Fujifilm Wako Pure Chemical Co.), and 0.3 mg/mL RNase A (Promega, Madison, WI, USA). Next, genomic DNA was extracted from each lysate using a Maxwell[®] RSC instrument (Promega), according to the manufacturer's instructions, and 250 bp paired-end libraries were then generated from the extracted DNA using a Nextera XT DNA Kit (Illumina, San Diego, CA, USA). Libraries were sequenced using a NovaSeq 6000 system (Illumina) at the Genome Information Research Center, Research Institute for Microbial Diseases, The University of Osaka, Osaka, Japan. On average, the number of reads was 5,433,301 (range 3,437,124–9,117,301).

Collection of published genome sequences

We previously sequenced the draft genomes of 161 *emm89* clinical isolates collected in Japan between 2011 and 2019¹⁴. We defined strains derived from STSS as “severe invasive,” and those obtained from pharyngitis, tonsillitis, and superficial skin lesions as “non-invasive” phenotypes.

To obtain public genome sequences of *emm89* strains isolated from other countries, we downloaded draft genome sequences in FASTA format from the National Center for Biotechnology Information (NCBI) database, using Fasterq-dump v.2.9.6. The phenotype of each strain, whether invasive or non-invasive, was defined according to the definitions in the respective references that reported the strains^{25–27}.

Genomic data processing and pan-genome analysis

All processes and analyses were performed using the National Institute of Genetics (NIG) supercomputer and SQUID at the D3 center of The University of Osaka (Osaka, Japan). We constructed a workflow for bacterial pan-GWAS and other bioinformatic processes (Figure S6). All collected sequences were subjected to quality checks using Fastp v.0.20.1, with a default cutoff value of Q>15⁵³. For newly isolated strains in Japan, *emm* typing was performed using the *emm*-typing-tool v.0.0.1, and only sequences of strains determined as *emm89* were used for the following analyses⁵⁴. All *emm89* sequence data were then subjected to *de novo* assembly using SKESA v.2.4.0, with default parameters⁵⁵.

Next, the MLST of each sequence was performed using MLST v.2.19.0^{56,57}. Clade typing was performed using BLAST v.2.13.0, with reference to the three *nga* promoter region sequences¹³. After the genes were annotated with Prokka v.1.14.5, the pan-genome of all sequences was calculated using Roary v.3.12.0, with the parameters “-e -mafft -r -qc -cd 99”^{58,59}. Roary generated a core gene alignment and the distribution of all genes among the strains. To extract SNPs/indels from core genes, including single-nucleotide indels, snp-sites v.2.5.1 with the option “-v” was used. The output files were further processed using BCFtools v.1.9, with the parameter “norm -m -”, enabling analysis of multiple alleles in the pan-GWAS. In parallel, k-mers were extracted using DBGWAS v.0.5.3, and the length of k-mers was set as 31 nt using the “-k 31” parameter of DBGWAS⁶⁰.

Phylogenetic analysis

Phylogenetic relationships were calculated from the core gene alignment, using IQ-tree v.1.16.12⁶¹, based on maximum likelihood. The substitution model was automatically selected considering the Akaike's and Bayesian information criteria by setting the “-m MFP” parameter of IQ-tree⁶². Phylogenetic trees were constructed using iTOL v.6.6⁶³. The similarity of clustering

in the two phylogenetic trees was statistically examined using Pearson's chi-square test with R v.4.0.3⁶⁴, followed by *post-hoc* analysis using residual analysis adjusted with the Holm's method, if $p < 0.05$.

GWAS

To investigate the associations between phenotypes and genotypes, including SNPs/indels and genes, we performed a pan-GWAS using pyseer v.1.3.4⁶⁵. The VCF file of SNPs/indels or the gene distribution matrix was designated as the genotype. To remove biases derived from lineages, we added information on the genetic distances between all pairs of strains as covariates using mash v.2.3⁶⁶. Briefly, *de novo* assemblies were compressed through conversion into minimum hash values using the command “mash sketch -s 10000.” Subsequently, the commands “mash dist” and “square_mash” were utilized to generate a genetic distance matrix expressed with Jaccard coefficients⁶⁶. The obtained matrix underwent eigenvalue decomposition, and the number of eigenvalues used for multidimensional scaling was visually determined using the plot of the relationships between the eigenvalues and contribution ratios. The number of eigenvalues and the distance matrix were then added as pyseer parameters. The pyseer calculation was iterated 1,000 times with randomized phenotypes, and the 5-percentile value of the minimal p -value in each calculation was set at the significance level. Using R and the package ggplot2, the results were visualized as a Manhattan plot for the SNPs/indels-based pan-GWAS and a volcano plot for the genes-based pan-GWAS, respectively⁶⁷. Heatmaps of the strains possessing significant variants were generated using Excel (v.16.66.1; Microsoft, Redmond, WA, USA). The correlation of the presence of significant COGs was calculated and visualized using the R program.

The k-mers-based GWAS was carried out using DBGWAS⁶⁰. K-mers were considered significant at a false discovery rate (q -value) of < 0.05 . DBGWAS-calculated complexes, which are regions encompassing the k-mers, were significantly related to pathology, and de-Bruijn graphs were generated based on these complexes. The sequences of the k-mers were outputted and mapped to a reference sequence using Geneious Prime v.2022.0.1 (Biomatters, Auckland, New Zealand) to identify the mutations indicated by the k-mers. For the reference strain, we adopted MGAS27061, which was isolated from an invasive case in the USA and whose complete chromosomal sequence has been used as the reference sequence of *emm89* clade 3⁶⁸. Promoter sequences in intergenic regions were predicted using web-based tools, MLDSPP and BacPP^{29,30}. Additionally, BLAST was employed to search the promoter sequences of *S. pyogenes* strain SF370 registered in the CDBProm database (<https://aw.iimas.unam.mx/cdbprom/>)⁶⁹.

Protein structure prediction

Significant variants found in the pan-GWAS, resulting in non-synonymous substitutions in proteins, were searched by converting nucleotide sequences into amino acid sequences using EMBOSS Transeq v.6.6.0.0⁷⁰. To assess whether these mutations affected the protein function, protein structure prediction models were constructed using AlphaFold v.2.2.2³¹. The calculations were performed five times for each model, by setting the option `multimer_predictions_per_model=5`. We predicted multimer models using the option `--model_preset="multimer"` if a protein is reported or anticipated to form a multimer. For each monomer, we selected the model with the best predicted local difference distance test (an indicator of local structural accuracy) score⁷¹. For each multimer, AlphaFold was calculated and expressed as a weighted combination of the interface-predicted TM and predicted TM scores ($ipTM + pTM$). pTM is a metric for overall topological accuracy and $ipTM$ is used to measure the structural accuracy of the protein-protein interface⁷². The transmembrane regions of the proteins were predicted using SOSUI (<https://harrier.nagahama-i-bio.ac.jp/sosui/mobile/>)⁷³. The structures of the obtained model were visualized using PyMOL v.2.5 (Schrödinger, LLC., New York, NY, USA).

Construction of the *fhuB* T218C mutant strains

We used the several times-passaged *S. pyogenes* TK02 strain, 'TK02', as the WT strain. The TK02 strain was originally isolated from a patient with severe invasive infection¹⁴. The whole genome of TK02' was sequenced, and the generated fasta file is available in Document S1. A point mutation, *fhuB* T218C, was introduced using the temperature-sensitive shuttle vector, pSET4s, as reported previously⁷⁴. Sanger sequencing confirmed the presence of the point mutation. In addition, we resequenced the draft genome to confirm that there were no differences apart from the point mutation, as described above. The bacterial strains, primers, and plasmids used in this study are listed in Tables S15 and S16. *Escherichia coli* strain DH5α (Takara Bio, Shiga, Japan) was used as a host for the plasmid derivatives. All *E. coli* strains were cultured in Luria Bertani broth, at 37°C, with agitation. For selection and maintenance of mutants, antibiotics were added to the media at the following concentrations: carbenicillin (Nacalai Tesque), 100 µg/mL for *E. coli*; and spectinomycin (Fujifilm Wako Pure Chemical Corporation), 100 µg/mL for *E. coli* and *S. pyogenes*.

Transcriptomic analysis

The *fhuB* WT and mutant strains were harvested in 30 mL of THY broth, of which 1 mL was dispensed into 10 mL of THY and the remainder was centrifuged and resuspended in 2 mL of heparinized human blood. Bacterial mixtures with THY or blood were dispensed into three aliquots (n=3 for each condition) and incubated at 37°C for 3 h. THY samples were centrifuged and resuspended in RNA Shield (Zymo Research, Irvine, CA, USA). For each blood sample, 2 volumes (1 mL) of RNA protection bacteria reagent (Qiagen, Hilden, Germany) were added. L5, included in the PureLink™ Total RNA Blood Purification Kit (Thermo Fisher Scientific, Waltham, MA, USA), was added to remove erythrocytes. The bacterial cell wall was mechanically lysed in Lysing Matrix B using a MagNA Lyser (Roche, Basel, Switzerland). After centrifugation, the total bacterial RNA was extracted using a Quick-RNA™ Miniprep Kit (Zymo Research), according to the manufacturer's instructions. Full-length cDNA was generated using the SMART-Seq® HT Kit (Takara Bio), according to the manufacturer's instructions. Pair-end libraries were generated using a Nextera XT DNA Kit and sequenced using a NovaSeq 6000 system (both from Illumina, San Diego, CA, USA). Sequenced data were preprocessed using Trimmomatic v.0.33 and FastQC v.0.12.1. The reads were mapped to the complete MGAS27061 genome (NCBI reference sequence: NZ_CP013840.1) using STAR v.2.7.0a. After a second quality check using FastQC, read counting was performed using featureCounts v.1.5.2⁷⁵. Differentially expressed genes were identified using iDEP v.0.96 and gene annotations from NCBI and Prokka were combined⁷⁶. Plots were created using iDEP and the R package ggplot2.

Intracellular ferric ion assay

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

Bactericidal assay

The bactericidal assay was performed as described previously, with minor modifications^{36,77-79}. Briefly, whole blood was collected from healthy adults. Human neutrophils and erythrocytes were prepared using PolymorphPrep™ (Serumwerk Bernburg, Bernburg, Germany), according to the manufacturer's instructions. Heparinized human blood was centrifuged at $500 \times g$ for 30 min, to isolate erythrocytes and polymorphonuclear cells, which were then suspended in Roswell Park Memorial Institute (RPMI)-1640 medium containing L-glutamine and Phenol Red (Fujifilm Wako Pure Chemical Corporation). Heat-inactivated plasma was prepared at 56°C for 30 min. Subsequently, 195 μ L of heparinized human whole blood, erythrocytes in RPMI-1640, polymorphonuclear leukocytes in RPMI-1640, plasma, heat-inactivated plasma, or brain heart infusion broth (BD Biosciences), and 5 μ L of early exponential phase bacteria with $0.9\text{--}2.0 \times 10^4$ CFUs/well were mixed in 96-well plates and incubated at 37°C, in an atmosphere containing 5% CO₂, for 1, 2, and 3 h. Viable bacterial cells were counted as CFUs by plating the diluted samples onto THY agar plates. The growth index was calculated as the number of CFUs at a specified time point divided by the number of CFUs in the initial inoculum. The assay was repeated three times, and the results of the respective experiments were combined. Statistical analyses were performed using the Mann–Whitney U test. Differences were considered statistically significant at $p < 0.05$, using Prism v.7.0c (GraphPad, La Jolla, CA, USA).

Data availability

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

Acknowledgements

We would like to thank the NGS core facility of the Genome Information Research Center at the Research Institute for Microbial Diseases of The University of Osaka for their support in the DNA sequencing and data analysis and the Bioinformatic Research Unit of Graduate School of Dentistry, The University of Osaka for their support in the bioinformatics analysis. This study was partially performed on the National Institute of Genetics (NIG) supercomputer at the Research Organization of Information and Systems National Institute of Genetics. This study was partly completed using SQUID at the D3 center, The University of Osaka, Japan, under the “Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures (JHPCN)” (Project ID: EX22701, jh230035, jh240003, and jh250016). Masayuki Ono and Kotaro Higashi were recipients of the Iwadare Scholarship from the Iwadare Scholarship Foundation. We wish to express our gratitude to Mami Tateshita (Sapporo City Institute of Public Health) as well as the medical institutions that participated in the collection of clinical strains.

This study was partly supported by AMED (JP17fk0108044, JP20fk0108130, JP20wm0325001, and JP243fa727001h), the Japan Society for the Promotion of Science KAKENHI (grant numbers 20KK0210, 22H03262, 22K19618, 22K19619, 23H03073, 23K19687, and 24K19854), the Takeda Science Foundation, Naito Foundation, and Joint Research Program of the Research Center for GLOBAL and LOCAL Infectious Diseases, Oita University (2022B05, 2025B01, and 2025B13). This study was conducted as part of “The Nippon Foundation - Osaka University Project for Infectious Disease Prevention.” This study was supported by JST SPRING (grant number JPMJSP2138). The funders had no role in the study design, data collection or analysis, decision to publish, or preparation of the manuscript.

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

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

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

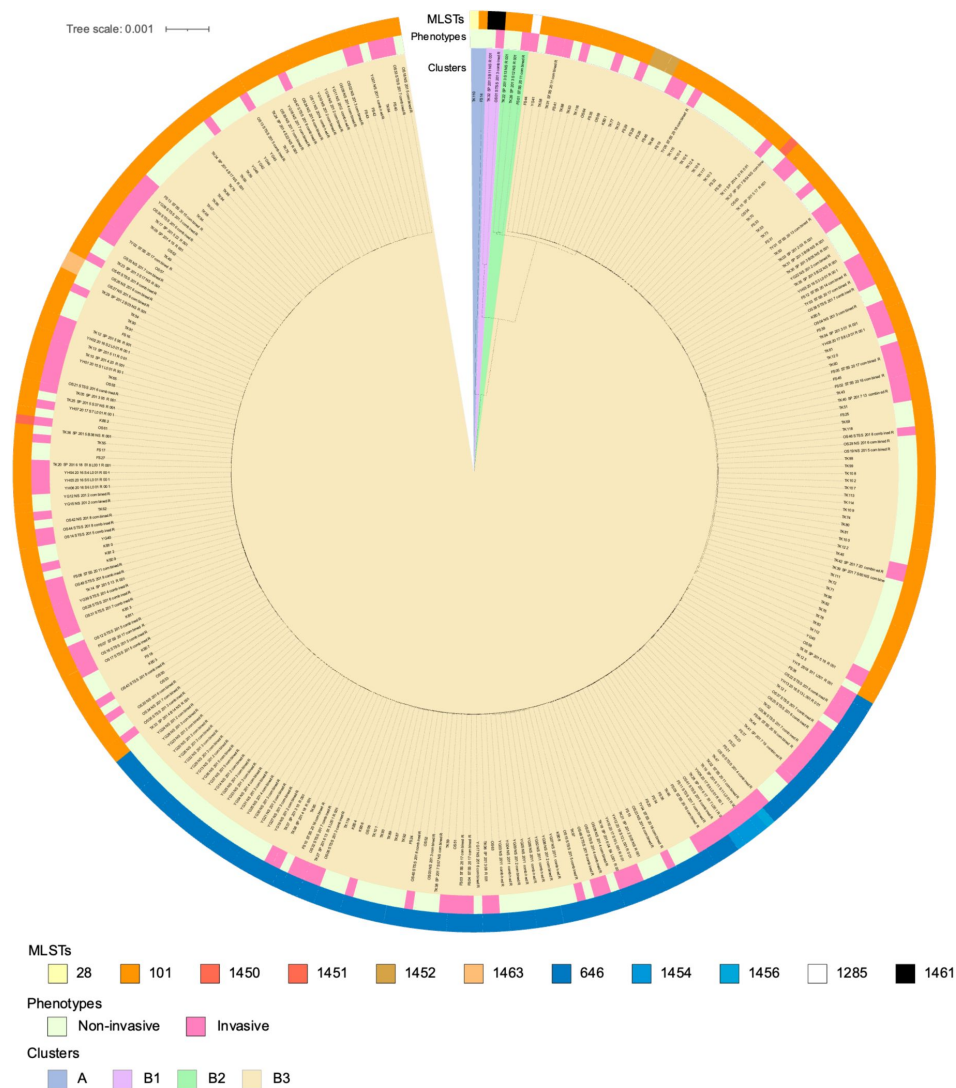


Figure S1.

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 promoter activity. Turner *et al.* further determined subtypes using mutations other than the key residues. FS24 strain has the same allele A-27G- 22T-18 as the type 3 or clade 3 but has an SNP highlighted in yellow in the -10 box.

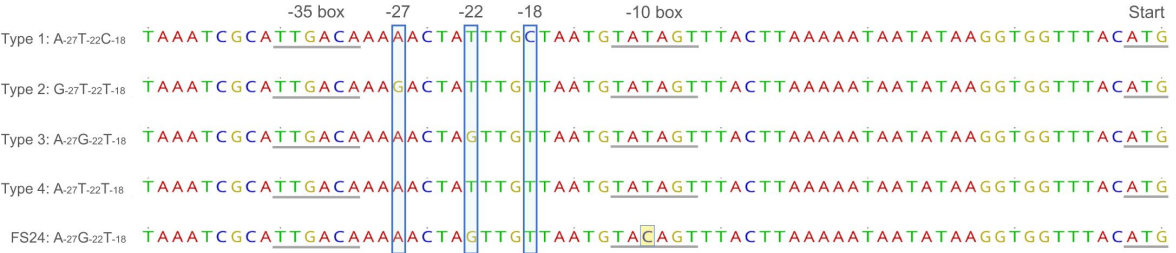
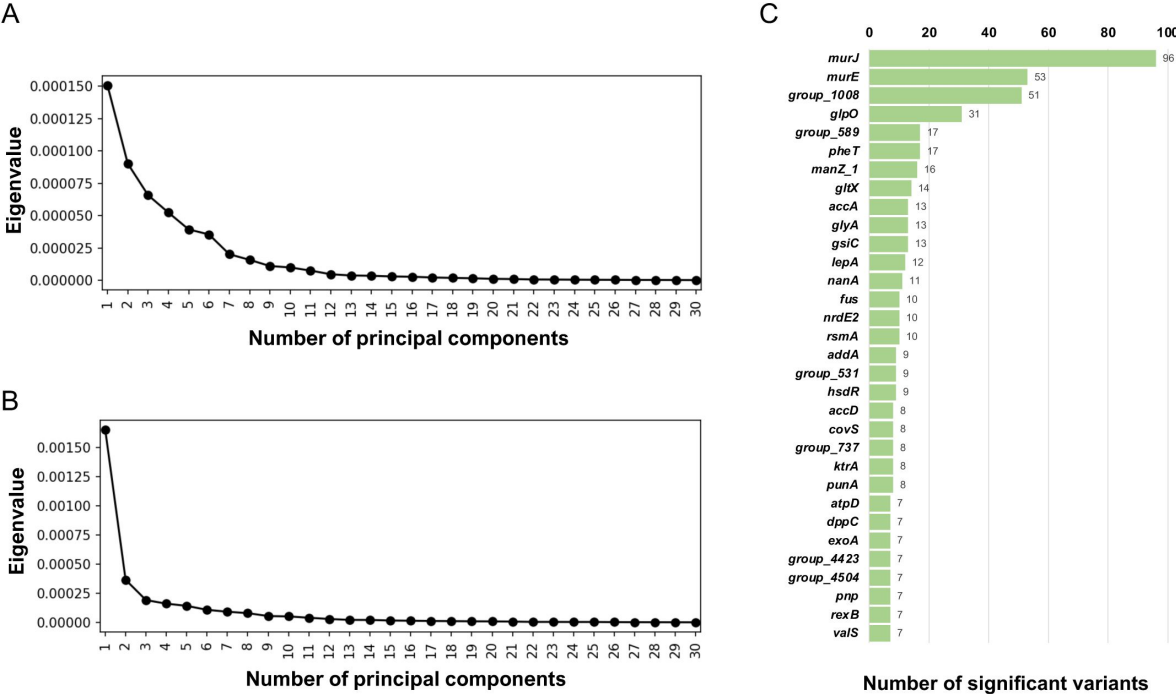


Figure S3.

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.



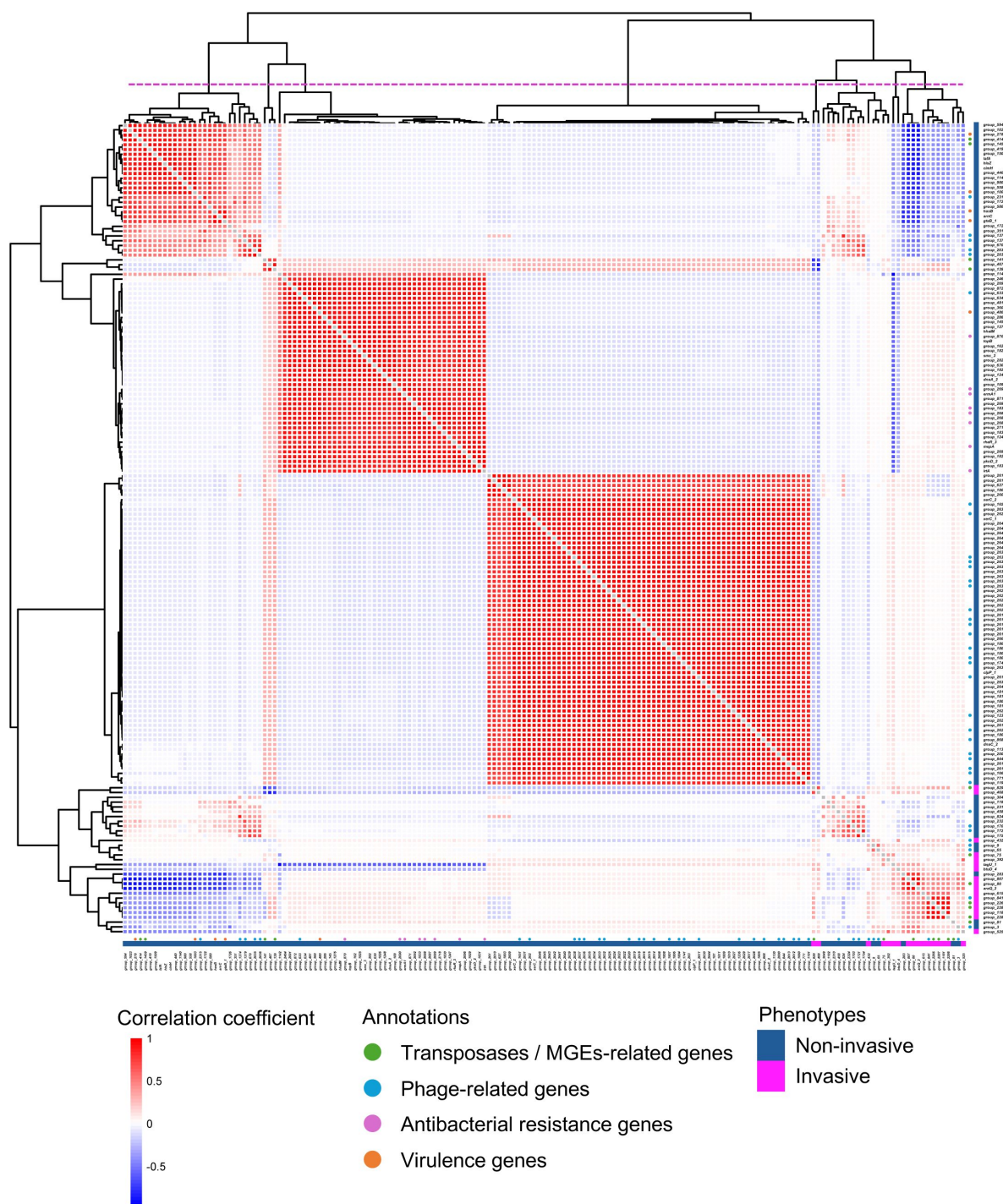


Figure S4.

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.

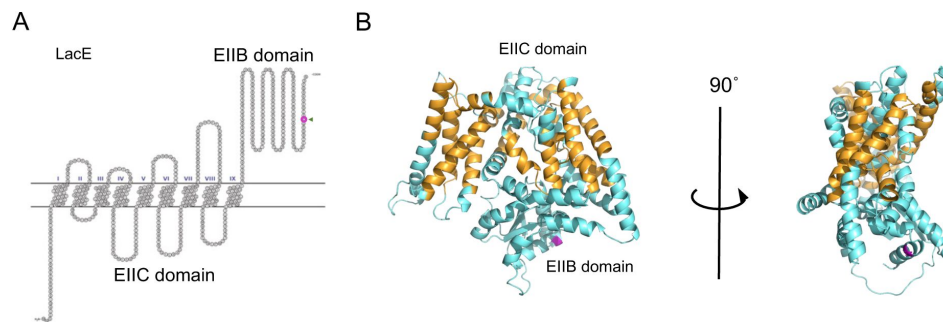


Figure S5.

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 EIIIC 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.

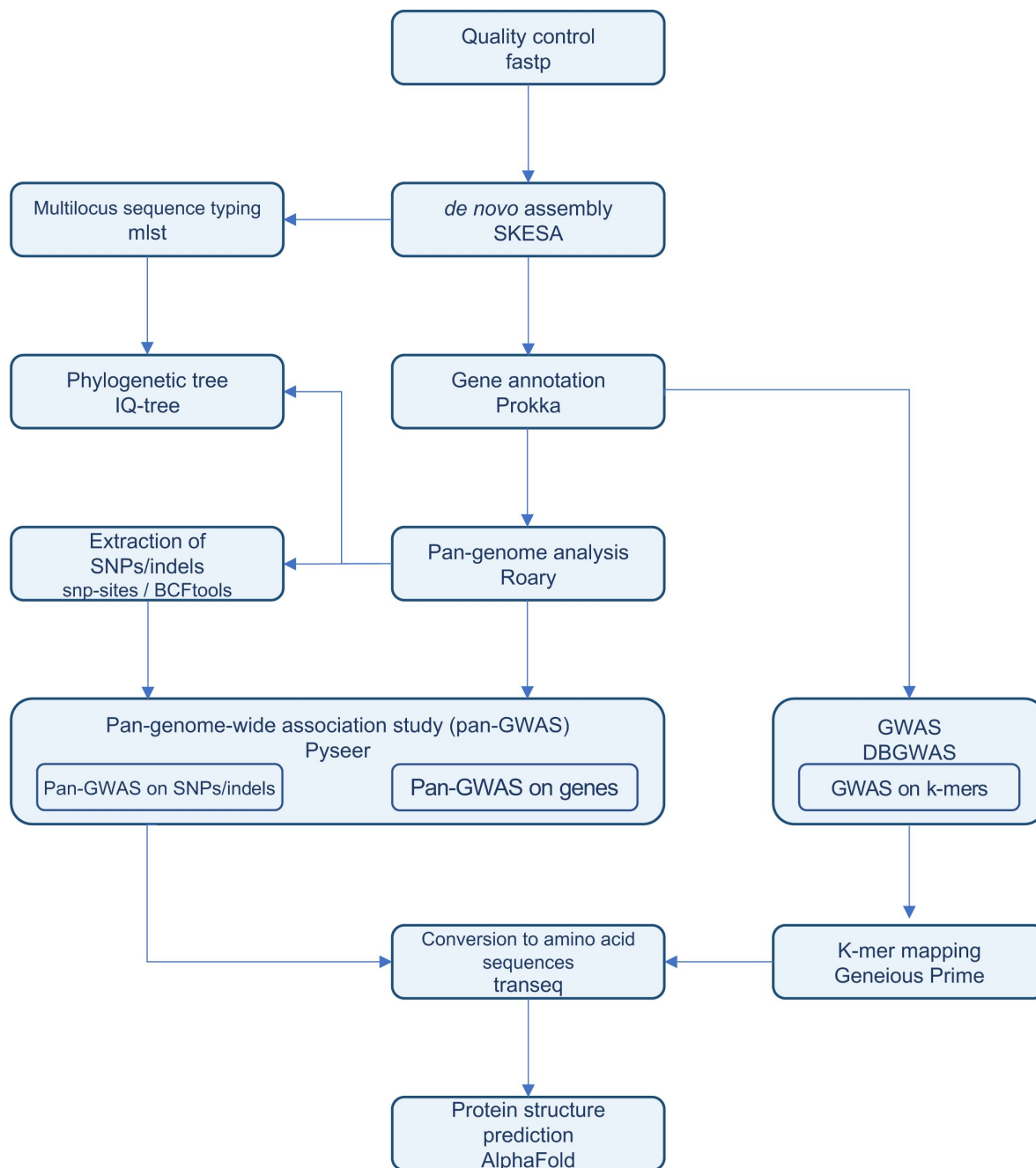


Figure S6.

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.

Additional information

Ethical approval

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

Author contributions

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

Additional files

Supplementary Tables [↗](#)

Document S1 [↗](#)

References

1. Craik N., Hla T., Cannon J., Moore H., Carapetis J.R., Sanyahumbi A (2022) **Global Disease Burden of Streptococcus pyogenes** In: Ferretti J.J., Stevens D.L., Fischetti V.A., editors. *Streptococcus pyogenes: Basic Biology to Clinical Manifestations* University of Oklahoma Health Sciences Center [Google Scholar](#)
2. Brouwer S., Rivera-Hernandez T., Curren B.F., Harbison-Price N., De Oliveira D.M.P., Jespersen M.G., Davies M.R., Walker M.J. (2023) **Pathogenesis, epidemiology and control of Group A Streptococcus infection** *Nat. Rev. Microbiol* **21**:431–447 <https://doi.org/10.1038/s41579-023-00865-7>[Google Scholar](#)
3. Carapetis J.R., Steer A.C., Mulholland E.K., Weber M (2005) **The global burden of group A streptococcal diseases** *Lancet Infect. Dis* **5**:685–694 [https://doi.org/10.1016/S1473-3099\(05\)70267-X](https://doi.org/10.1016/S1473-3099(05)70267-X)[Google Scholar](#)
4. Stevens D.L., Bisno A.L., Chambers H.F., Dellinger E.P., Goldstein E.J., Gorbach S.L., Hirschmann J.V., Kaplan S.L., Montoya J.G., Wade J.C., et al. (2014) **Practice guidelines for the diagnosis and management of skin and soft tissue infections: 2014 update by the Infectious Diseases Society of America** *Clin. Infect. Dis* **59**:e10–e52 <https://doi.org/10.1093/cid/ciu444>[Google Scholar](#)
5. Stevens D.L., Bryant A.E. (2016) **Severe Group A Streptococcal Infections** In: Ferretti J.J., Stevens D.L., Fischetti V.A., editors. *Streptococcus pyogenes: Basic Biology to Clinical Manifestations* University of Oklahoma Health Sciences Center [Google Scholar](#)
6. Walker M.J., Barnett T.C., McArthur J.D., Cole J.N., Gillen C.M., Henningham A., Sriprakash K.S., Sanderson-Smith M.L., Nizet V (2014) **Disease manifestations and pathogenic mechanisms of Group A Streptococcus** *Clin. Microbiol. Rev* **27**:264–301 <https://doi.org/10.1128/CMR.00101-13>[Google Scholar](#)
7. Yamaguchi M., Terao Y., Mori-Yamaguchi Y., Domon H., Sakaue Y., Yagi T., Nishino K., Yamaguchi A., Nizet V., Kawabata S (2013) **Streptococcus pneumoniae invades erythrocytes and utilizes them to evade human innate immunity** *PLOS One* **8**:e77282 <https://doi.org/10.1371/JOURNAL.PONE.0077282>[Google Scholar](#)
8. Bessen D.E., Smeesters P.R., Beall B.W (2018) **Molecular Epidemiology, Ecology, and Evolution of Group A Streptococci** *Microbiol. Spectr* **6** <https://doi.org/10.1128/microbiolspec.CPP3-0009-2018>[Google Scholar](#)
9. Ikebe T., Okuno R., Kanda Y., Sasaki M., Yamaguchi T., Otsuka H., Kazawa Y., Suzuki M., Ohya H., Uchida K., Ohnishi M (2021) **Molecular characterization and antimicrobial resistance of group A streptococcus isolates in streptococcal toxic shock syndrome cases in Japan from 2013 to 2018** *Int. J. Med. Microbiol* **311** <https://doi.org/10.1016/j.ijmm.2021.151496>[Google Scholar](#)
10. Fay K., Onukwube J., Chochua S., Schaffner W., Cieslak P., Lynfield R., Muse A., Smelser C., Harrison L.H., Farley M., et al. (2021) **Patterns of Antibiotic Nonsusceptibility Among Invasive Group A Streptococcus Infections-United States, 2006–2017** *Clin. Infect. Dis* **73**:1957–1964 <https://doi.org/10.1093/cid/ciab575>[Google Scholar](#)

11. Turner C.E., Abbott J., Lamagni T., Holden M.T.G., David S., Jones M.D., Game L., Efstratiou A., Sriskandan S (2015) **Emergence of a new highly successful acapsular group A *Streptococcus* clade of genotype emm89 in the United Kingdom** *mBio* **6**:e00622 <https://doi.org/10.1128/mBio.00622-15>Google Scholar
12. Zhu L., Olsen R.J., Nasser W., Beres S.B., Vuopio J., Kristinsson K.G., Gottfredsson M., Porter A.R., DeLeo F.R., Musser J.M (2015) **A molecular trigger for intercontinental epidemics of group A *Streptococcus*** *J. Clin. Invest* **125**:3545–3559 <https://doi.org/10.1172/JCI82478>Google Scholar
13. Zhu L., Olsen R.J., Nasser W., De La Riva Morales I., Musser J.M. (2015) **Trading Capsule for Increased Cytotoxin Production: Contribution to Virulence of a Newly Emerged Clade of emm89 *Streptococcus pyogenes*** *mBio* **6**:e01378–e01315 <https://doi.org/10.1128/mbio.01378-15>Google Scholar
14. Hirose Y., Yamaguchi M., Takemoto N., Miyoshi-Akiyama T., Sumitomo T., Nakata M., Ikebe T., Hanada T., Yamaguchi T., Kawahara R., et al. (2020) **Genetic Characterization of *Streptococcus pyogenes* emm89 Strains Isolated in Japan From 2011 to 2019** *Infect. Microbes Dis* **2**:160–166 <https://doi.org/10.1097/IM9.000000000000038>Google Scholar
15. Walker M.J., Hollands A., Sanderson-Smith M.L., Cole J.N., Kirk J.K., Henningham A., McArthur J.D., Dinkla K., Aziz R.K., Kansal R.G., et al. (2007) **DNase Sda1 provides selection pressure for a switch to invasive group A streptococcal infection** *Nat. Med* **13**:981–985 <https://doi.org/10.1038/nm1612>Google Scholar
16. Sumbly P., Whitney A.R., Graviss E.A., DeLeo F.R., Musser J.M (2006) **Genome-wide analysis of group A streptococci reveals a mutation that modulates global phenotype and disease specificity** *PLoS Pathog* **2**:e5 <https://doi.org/10.1371/journal.ppat.0020005>Google Scholar
17. Cole J.N., Barnett T.C., Nizet V., Walker M.J (2011) **Molecular insight into invasive group A streptococcal disease** *Nat. Rev. Microbiol* **9**:724–736 <https://doi.org/10.1038/nrmicro2648>Google Scholar
18. Ikebe T., Ato M., Matsumura T., Hasegawa H., Sata T (2010) **Highly Frequent Mutations in Negative Regulators of Multiple Virulence Genes in Group A Streptococcal Toxic Shock Syndrome Isolates** *PLOS Pathog* **6**:e1000832 <https://doi.org/10.1371/journal.ppat.1000832>Google Scholar
19. Ministry of Health, Labour and Welfare (no date) **Severe Invasive Group A Streptococcal Infections** <https://www.mhlw.go.jp/bunya/kenkou/kekkaku-kansenshou11/01-05-06.html>
20. Centers for Disease Control and Prevention, C.f.D.C.a (no date) **Streptococcal Toxic Shock Syndrome (STSS) (*Streptococcus pyogenes*) 2010 Case Definition** <https://ndc.services.cdc.gov/case-definitions/streptococcal-toxic-shock-syndrome-2010/>
21. Ikebe T., Murai N., Endo M., Okuno R., Murayama S., Saitoh K., Yamai S., Suzuki R., Isobe J., Tanaka D., et al. (2003) **Changing prevalent T serotypes and emm genotypes of *Streptococcus pyogenes* isolates from streptococcal toxic shock-like syndrome (TSLs) patients in Japan** *Epidemiol. Infect* **130**:569–572 [Google Scholar](#)
22. Beall B., Facklam R.R., Elliott J.A., Franklin A.R., Hoenes T., Jackson D., Laclaire L., Thompson T., Viswanathan R (1998) **Streptococcal emm types associated with T-agglutination types and the use of conserved emm gene restriction fragment patterns for subtyping group A streptococci** *J. Med. Microbiol* **47**:893–898 <https://doi.org/10.1099/00222615-47-10-893>Google Scholar

23. Johnson D.R., Kaplan E.L (1993) **A review of the correlation of T- agglutination patterns and M-protein typing and opacity factor production in the identification of group A streptococci** *J. Med. Microbiol* **38**:311–315 <https://doi.org/10.1099/00222615-38-5-311>Google Scholar
24. Katsukawa C., Tamaru A., Morikawa Y., Oda K (2002) **M protein gene (emm) typing of Streptococcus pyogenes** *Kansenshogaku Zasshi* **76**:238–245 <https://doi.org/10.11150/kansenshogakuzasshi1970.76.238>Google Scholar
25. Beres S.B., Olsen R.J., Saavedra M.O., Ure R., Reynolds A., Lindsay D.S.J., Smith A.J., Musser J.M (2017) **Genome sequence analysis of emm89 Streptococcus pyogenes strains causing infections in Scotland, 2010–2016** *J. Med. Microbiol* **66**:1765–1773 <https://doi.org/10.1099/jmm.0.000622>Google Scholar
26. Chochua S., Metcalf B.J., Li Z., Rivers J., Mathis S., Jackson D., Gertz R.E., Srinivasan V., Lynfield R., Van Beneden C., et al. (2017) **Population and Whole Genome Sequence Based Characterization of Invasive Group A Streptococci Recovered in the United States during 2015** *mBio* **8**:e01422–17 <https://doi.org/10.1128/mBio.01422-17>Google Scholar
27. Davies M.R., McIntyre L., Mutreja A., Lacey J.A., Lees J.A., Towers R.J., Smeesters P.R., Frost H.R., Price D.J., G Holden M.T., et al. (2019) **Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics** *Nat. Genet* **51**:1035–1043 <https://doi.org/10.1038/s41588-019-0417-8>Google Scholar
28. Turner C.E., Holden M.T.G., Blane B., Horner C., Peacock S.J., Sriskandan S (2019) **The Emergence of Successful Streptococcus pyogenes Lineages through Convergent Pathways of Capsule Loss and Recombination Directing High Toxin Expression** *mBio* **10**:1–20 <https://doi.org/10.1128/mBio.02521-19>Google Scholar
29. Paul S., Olymon K., Martinez G.S., Sarkar S., Yella V.R., Kumar A (2024) **MLDSPP: Bacterial Promoter Prediction Tool Using DNA Structural Properties with Machine Learning and Explainable AI** *J Chem Inf Model* **64**:2705–2719 <https://doi.org/10.1021/acs.jcim.3c02017>Google Scholar
30. de Avila e Silva S., Echeverrigaray S., Gerhardt G.J. (2011) **BacPP: Bacterial promoter prediction - A tool for accurate sigma-factor specific assignment in enterobacteria** *J Theor Biol* **287**:92–99 [Google Scholar](#)
31. Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., Žídek A., Potapenko A., et al. (2021) **Highly accurate protein structure prediction with AlphaFold** *Nature* :583–589 <https://doi.org/10.1038/s41586-021-03819-2>Google Scholar
32. Jain I., Danger J.L., Burgess C., Uppal T., Sumby P (2020) **The group A Streptococcus accessory protein RocA: regulatory activity, interacting partners and influence on disease potential** *Mol. Microbiol* **113**:190–207 <https://doi.org/10.1111/mmi.14410>Google Scholar
33. Ribardo D.A., McIver K.S (2006) **Defining the Mga regulon: Comparative transcriptome analysis reveals both direct and indirect regulation by Mga in the group A streptococcus** *Mol. Microbiol* **62**:491–508 <https://doi.org/10.1111/j.1365-2958.2006.05381.x>Google Scholar
34. Kachroo P., Eraso J.M., Beres S.B., Olsen R.J., Zhu L., Nasser W., Bernard P.E., Cantu C.C., Saavedra M.O., Arredondo M.J., et al. (2019) **Integrated analysis of population genomics, transcriptomics and virulence provides novel insights into Streptococcus pyogenes**

- [pathogenesis](#) *Nat. Genet* **51**:548–559 <https://doi.org/10.1038/S41588-018-0343-1>Google Scholar
35. Vega L.A., Malke H., McIver K.S. (2016) **Virulence-Related Transcriptional Regulators of *Streptococcus pyogenes*** In: Ferretti J.J., Stevens D.L., Fischetti V.A., editors. *Streptococcus pyogenes : Basic Biology to Clinical Manifestations* University of Oklahoma Health Sciences Center [Google Scholar](#)
36. Terao Y., Mori Y., Yamaguchi M., Shimizu Y., Ooe K., Hamada S., Kawabata S (2008) **Group A streptococcal cysteine protease degrades C3 (C3b) and contributes to evasion of innate immunity** *J. Biol. Chem* **283**:6253–6260 <https://doi.org/10.1074/jbc.M704821200>Google Scholar
37. Sumitomo T., Nakata M., Higashino M., Jin Y., Terao Y., Fujinaga Y., Kawabata S (2011) **Streptolysin S contributes to group A streptococcal translocation across an epithelial barrier** *J. Biol. Chem* **286**:2750–2761 <https://doi.org/10.1074/jbc.M110.171504>Google Scholar
38. Honda-Ogawa M., Ogawa T., Terao Y., Sumitomo T., Nakata M., Ikebe K., Maeda Y., Kawabata S (2013) **Cysteine proteinase from *Streptococcus pyogenes* enables evasion of innate immunity via degradation of complement factors** *J. Biol. Chem* **288**:15854–15864 <https://doi.org/10.1074/jbc.M113.469106>Google Scholar
39. Lin A., Loughman J.A., Zinselmeyer B.H., Miller M.J., Caparon M.G (2009) **Streptolysin S inhibits neutrophil recruitment during the early stages of *Streptococcus pyogenes* infection** *Infect. Immun* **77**:5190–5201 <https://doi.org/10.1128/IAI.00420-09>Google Scholar
40. Cimen M.Y (2008) **Free radical metabolism in human erythrocytes** *Clin. Chim. Acta* **390**:1–11 <https://doi.org/10.1016/j.cca.2007.12.025>Google Scholar
41. Laabei M., Ermert D (2019) **Catch Me if You Can: *Streptococcus pyogenes* Complement Evasion Strategies** *J. Innate. Immun* **11**:3–12 <https://doi.org/10.1159/000492944>Google Scholar
42. He L.Y., Yu Y.B., Liu Y., Le Y.J., Li S., Yang X.Y (2022) **Immunization with the lipoprotein FtsB stimulates protective immunity against *Streptococcus pyogenes* infection in mice** *Front. Microbiol* **13** <https://doi.org/10.3389/fmicb.2022.969490>Google Scholar
43. Nguyen M.T., Gotz F (2016) **Lipoproteins of Gram-Positive Bacteria: Key Players in the Immune Response and Virulence** *Microbiol. Mol. Biol. Rev* **80**:891–903 <https://doi.org/10.1128/MMBR.00028-16>Google Scholar
44. Bessen D.E., McShan W.M., Nguyen S.V., Shetty A., Agrawal S., Tettelin H (2015) **Molecular epidemiology and genomics of group A *Streptococcus*** *Infect. Genet. Evol* **33**:393–418 <https://doi.org/10.1016/j.meegid.2014.10.011>Google Scholar
45. McShan W.M., McCullor K.A., Nguyen S.V (2019) **The Bacteriophages of *Streptococcus pyogenes*** *Microbiol. Spectr* **7** <https://doi.org/10.1128/microbiolspec.GPP3-0059-2018>Google Scholar
46. Berbel D., Camara J., Gonzalez-Diaz A., Cubero M., Lopez de Egea G., Marti S., Tubau F., Dominguez M.A., Ardanuy C. (2021) **Deciphering mobile genetic elements disseminating macrolide resistance in *Streptococcus pyogenes* over a 21 year period in Barcelona, Spain** *J. Antimicrob. Chemother* **76**:1991–2003 <https://doi.org/10.1093/jac/dkab130>Google Scholar
47. Wajima T., Chiba N., Morozumi M., Shouji M., Sunaoshi K., Sugita K., Tajima T., Ubukata K., Group A *Streptococcus* Surveillance Study Group (2014) **Prevalence of macrolide resistance**

among group A streptococci isolated from pharyngotonsillitis *Microb. Drug. Resist* **20**:431–435 <https://doi.org/10.1089/mdr.2013.0213>Google Scholar

48. Hasegawa T., Hata N., Matsui H., Isaka M., Tatsuno I (2017) **Characterisation of clinically isolated *Streptococcus pyogenes* from balanoposthitis patients, with special emphasis on emm89 isolates** *J. Med. Microbiol* **66**:511–516 <https://doi.org/10.1099/jmm.0.000460>Google Scholar
49. Tatara K., Gotoh K., Okumiya K., Teramachi M., Ishimoto K., Tanaka Y., Iwahashi J., Shindou S., Yamashita Y., Watanabe H (2020) **Molecular epidemiology, antimicrobial susceptibility, and characterization of fluoroquinolone non-susceptible *Streptococcus pyogenes* in Japan** *J. Infect. Chemother* **26**:280–284 <https://doi.org/10.1016/j.jiac.2019.10.004>Google Scholar
50. Tanaka Y., Gotoh K., Teramachi M., Ishimoto K., Tsumura N., Shindou S., Yamashita Y (2016) **Molecular epidemiology, antimicrobial susceptibility, and characterization of macrolide-resistant *Streptococcus pyogenes* in Japan** *J. Infect. Chemother* **22**:727–732 <https://doi.org/10.1016/j.jiac.2016.06.013>Google Scholar
51. Ubukata K., Wajima T., Morozumi M., Sakuma M., Tajima T., Matsubara K., Itahashi K., Iwata S (2020) **Changes in epidemiologic characteristics and antimicrobial resistance of *Streptococcus pyogenes* isolated over 10 years from Japanese children with pharyngotonsillitis** *J. Med. Microbiol* **69**:443–450 <https://doi.org/10.1099/jmm.0.001158>Google Scholar
52. Jolley K.A., Bray J.E., Maiden M.C.J (2018) **Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications** *Wellcome Open Res* **3** <https://doi.org/10.12688/wellcomeopenres.14826.1>Google Scholar
53. Chen S., Zhou Y., Chen Y., Gu J (2018) **fastp: an ultra-fast all-in-one FASTQ preprocessor** *Bioinformatics* **34**:i884–i890 <https://doi.org/10.1093/bioinformatics/bty560>Google Scholar
54. Kapatai G., Coelho J., Platt S., Chalker V.J (2017) **Whole genome sequencing of group A *Streptococcus*: development and evaluation of an automated pipeline for emm gene typing** *PeerJ* **5**:e3226 <https://doi.org/10.7717/peerj.3226>Google Scholar
55. Souvorov A., Agarwala R., Lipman D.J (2018) **SKESA: strategic k-mer extension for scrupulous assemblies** *Genome Biol* **19** <https://doi.org/10.1186/s13059-018-1540-z>Google Scholar
56. Jolley K.A., Maiden M.C (2010) **BIGSdb: scalable analysis of bacterial genome variation at the population level** *BMC Bioinformatics* **11** <https://doi.org/10.1186/1471-2105-11-595>Google Scholar
57. Maiden M.C.J., Jansen Van Rensburg M.J., Bray J.E., Earle S.G., Ford S.A., Jolley K.A., McCarthy N.D (2013) **MLST revisited: the gene-by-gene approach to bacterial genomics** *Europe PMC Funders Group Nat. Rev. Microbiol* **11**:728–736 <https://doi.org/10.1038/nrmicro3093>Google Scholar
58. Seemann T (2014) **Prokka: rapid prokaryotic genome annotation** *Bioinformatics* **30**:2068–2069 <https://doi.org/10.1093/bioinformatics/btu153>Google Scholar

59. Page A.J., Cummins C.A., Hunt M., Wong V.K., Reuter S., Holden M.T., Fookes M., Falush D., Keane J.A., Parkhill J (2015) **Roary: rapid large- scale prokaryote pan genome analysis** *Bioinformatics* **31**:3691–3693 <https://doi.org/10.1093/bioinformatics/btv421>Google Scholar
60. Jaillard M., Lima L., Tournoud M., Mahé P., van Belkum A., Lacroix V., Jacob L (2018) **A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events** *PLOS Genet* **14**:e1007758 <https://doi.org/10.1371/journal.pgen.1007758>Google Scholar
61. Nguyen L.T., Schmidt H.A., von Haeseler A., Minh B.Q. (2015) **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies** *Mol. Biol. Evol* **32**:268–274 <https://doi.org/10.1093/molbev/msu300>Google Scholar
62. Kalyanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermin L.S. (2017) **ModelFinder: fast model selection for accurate phylogenetic estimates** *Nat. Methods* **14**:587–589 <https://doi.org/10.1038/nmeth.4285>Google Scholar
63. Letunic I., Bork P (2021) **Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation** *Nucleic Acids Res* **49**:W293–W296 <https://doi.org/10.1093/nar/gkab301>Google Scholar
64. R Core Team (2022) **A language and environment for statistical computing** <https://www.R-project.org/>
65. Lees J.A., Galardini M., Bentley S.D., Weiser J.N., Corander J (2018) **pyseer: A comprehensive tool for microbial pangenome-wide association studies** *Bioinformatics* **34**:4310–4312 <https://doi.org/10.1093/bioinformatics/bty539>Google Scholar
66. Ondov B.D., Treangen T.J., Melsted P., Mallonee A.B., Bergman N.H., Koren S., Phillippy A.M (2016) **Mash: fast genome and metagenome distance estimation using MinHash** *Genome Biol* **17** <https://doi.org/10.1186/s13059-016-0997-x>Google Scholar
67. Wickham H (2016) **ggplot2: Elegant Graphics for Data Analysis** New York: Springer-Verlag <https://doi.org/10.1007/978-3-319-24277-4>Google Scholar
68. Beres S.B., Kachroo P., Nasser W., Olsen R.J., Zhu L., Flores A.R., de la Riva I., Paez-Mayorga J., Jimenez F.E., Cantu C., et al. (2016) **Transcriptome remodeling contributes to epidemic disease caused by the human pathogen *Streptococcus pyogenes*** *mBio* **7**:1–14 <https://doi.org/10.1128/mBio.00403-16>Google Scholar
69. Martinez G.S., Perez-Rueda E., Kumar A., Dutt M., Maya C.R., Ledesma-Dominguez L., Casa P.L., Kumar A., Avila E Silva S., Kelvin D.J. (2024) **CDBProm: the Comprehensive Directory of Bacterial Promoters** *NAR Genom Bioinform* **6**:lqae018 <https://doi.org/10.1093/nargab/lqae018>Google Scholar
70. Rice P., Longden I., Bleasby A (2000) **EMBOSS: the European Molecular Biology Open Software Suite** *Trends Genet* **16**:276–277 [https://doi.org/10.1016/s0168-9525\(00\)002024-2](https://doi.org/10.1016/s0168-9525(00)002024-2)Google Scholar
71. Mariani V., Biasini M., Barbato A., Schwede T (2013) **IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests** *Bioinformatics* **29**:2722–2728 <https://doi.org/10.1093/bioinformatics/btt473>Google Scholar

72. Yin R., Feng B.Y., Varshney A., Pierce B.G (2022) **Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants** *Protein Sci* **31**:e4379 <https://doi.org/10.1002/pro.4379>Google Scholar
73. Hirokawa T., Boon-Chieng S., Mitaku S (1998) **SOSUI: classification and secondary structure prediction system for membrane proteins** *Bioinformatics* **14**:378–379 <https://doi.org/10.1093/bioinformatics/14.4.378>Google Scholar
74. Takamatsu D., Osaki M., Sekizaki T (2001) **Thermosensitive suicide vectors for gene replacement in *Streptococcus suis*** *Plasmid* **46**:140–148 <https://doi.org/10.1006/PLAS.2001.1532>Google Scholar
75. Liao Y., Smyth G.K., Shi W (2014) **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features** *Bioinformatics* **30**:923–930 <https://doi.org/10.1093/bioinformatics/btt656>Google Scholar
76. Ge S.X., Son E.W., Yao R (2018) **iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data** *BMC Bioinformatics* **19** <https://doi.org/10.1186/s12859-018-2486-6>Google Scholar
77. Lancefield R.C (1957) **Differentiation of group A streptococci with a common R antigen into three serological types, with special reference to the bactericidal test** *J. Exp. Med* **106**:525–544 <https://doi.org/10.1084/jem.106.4.525>Google Scholar
78. Yamaguchi M., Goto K., Hirose Y., Yamaguchi Y., Sumitomo T., Nakata M., Nakano K., Kawabata S (2019) **Identification of evolutionarily conserved virulence factor by selective pressure analysis of *Streptococcus pneumoniae*** *Commun Biol* **2** <https://doi.org/10.1038/s42003-019-0340-7>Google Scholar
79. Takemura M., Yamaguchi M., Kobayashi M., Sumitomo T., Hirose Y., Okuzaki D., Ono M., Motooka D., Goto K., Nakata M., et al. (2022) **Pneumococcal BgaA Promotes Host Organ Bleeding and Coagulation in a Mouse Sepsis Model** *Front. Cell. Infect. Microbiol* **12** <https://doi.org/10.3389/fcimb.2022.844000>Google Scholar

Author information

Masayuki Ono

Department of Microbiology, Graduate School of Dentistry, The University of Osaka, Osaka, Japan, Bioinformatics Research Unit, Graduate School of Dentistry, The University of Osaka, Osaka, Japan
ORCID iD: [0009-0006-9557-9669](https://orcid.org/0009-0006-9557-9669)

Masaya Yamaguchi

Department of Microbiology, Graduate School of Dentistry, The University of Osaka, Osaka, Japan, Bioinformatics Research Unit, Graduate School of Dentistry, The University of Osaka, Osaka, Japan, Bioinformatics Center, Research Institute for Microbial Diseases, The University of Osaka, Osaka, Japan, Center for Infectious Diseases Education and Research, The University of Osaka, Osaka, Japan, Laboratory of Microbial Informatics, Microbial Research Center for Health and Medicine, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan

ORCID iD: [0000-0001-6218-7112](https://orcid.org/0000-0001-6218-7112)

For correspondence: yamaguchi.masaya@nibn.go.jp

Daisuke Motooka

Bioinformatics Center, Research Institute for Microbial Diseases, The University of Osaka, Osaka, Japan, NGS Core Facility, Research Institute for Microbial Diseases, The University of Osaka, Osaka, Japan, Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives (OTRI), The University of Osaka, Osaka, Japan

Yujiro Hirose

Department of Microbiology, Graduate School of Dentistry, The University of Osaka, Osaka, Japan

Kotaro Higashi

Department of Microbiology, Graduate School of Dentistry, The University of Osaka, Osaka, Japan

Tomoko Sumitomo

Department of Microbiology, Graduate School of Dentistry, The University of Osaka, Osaka, Japan, Department of Oral Microbiology, Graduate School of Biomedical Sciences, Tokushima University, Tokushima, Japan

Tohru Miyoshi-Akiyama

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

Rumi Okuno

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

Takahiro Yamaguchi

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

Ryuji Kawahara

Department of Bacteriology, Osaka Institute of Public Health, Osaka, Japan
ORCID iD: [0000-0002-6864-3788](https://orcid.org/0000-0002-6864-3788)

Hitoshi Otsuka

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

Noriko Nakanishi

Department of Infectious Diseases, Kobe Institute of Health, Hyogo, Japan

Yu Kazawa

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

Chikara Nakagawa

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

Ryo Yamaguchi

Sapporo Public Health Office, Hokkaido, Japan

Hiroo Sakai

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

Yuko Matsumoto

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

Tadayoshi Ikebe

Department of Bacteriology I, National Institute of Infectious Diseases, Tokyo, Japan
ORCID iD: [0000-0003-1021-1728](https://orcid.org/0000-0003-1021-1728)

Shigetada Kawabata

Department of Microbiology, Graduate School of Dentistry, The University of Osaka, Osaka, Japan, Center for Infectious Diseases Education and Research, The University of Osaka, Osaka, Japan

Editors

Reviewing Editor

Dominique Soldati-Favre

University of Geneva, Geneva, Switzerland

Senior Editor

Dominique Soldati-Favre

University of Geneva, Geneva, Switzerland

Reviewer #1 (Public review):

Summary:

In this study, the authors sequenced emm89 serotype genomes of clinical isolates from patients in Japan, where the number of invasive Group A Streptococcus (GAS), especially those of the emm89 serotype, has drastically increased over the past 10-15 years. The sequences from this cohort were compared against a large collection of publicly available global isolates, yielding a total of almost 1000 genomes in the analysis. Because the researchers focused on the emm89 serotype, they could construct a common core genome, with subsequent ability to analyze genomic differences in accessory genes and intergenic regions that contributed to the invasive phenotype using multiple types of GWAS analysis (SNP, k-mer). Their analysis demonstrates some mutations responsible for invasiveness are specific to the Japanese strains, and that multiple independent virulence factors can contribute to invasiveness. None of the invasive phenotypes were correlated with new gene acquisition. Together, the data support that synergy between bacterial survival and upregulation of virulence factors contribute to the development of severe infection.

Strengths:

- The authors verify their analysis by confirming that *covS* is one of the more frequently mutated genes in invasive strains of GAS, as has been shown in other publications.

- A mutation in one of the SNPs attributed to invasiveness (SNP fhuB) was introduced into an invasive strain. The authors demonstrate that this mutant strain survives less well in human blood. Therefore, the authors have experimental data to support their claims that their analysis uncovered a new mutation/SNP that contributed to invasiveness.

Weaknesses:

- It would be helpful for the authors to highlight why their technique (large scale analysis of one emm type) can yield more information than a typical GWAS analysis of invasive vs. non-invasive strains. Are SNPs easier to identify using a large-scale core genome? Is it more likely evolutionarily to find mutations in non-coding regions as opposed to the core genome and accessory genes, and this is what this technique allows? Did the analysis yield unexpected genes or new genes that had not been previously identified in other GWAS analyses? These points may need to be made more apparent in the results and deserves some thought in the discussion section.
- The Alpha-fold data does not demonstrate why the mutations the authors identified could contribute to the invasive phenotype. It would be helpful to show an overlay of the predicted structures containing the different SNPs to demonstrate the potential structural differences that can occur due to the SNP. This would make the data more convincing that the SNP has a potential impact on the function of the protein. Similarly, the authors discuss modification of the hydrophobicity of the side chain in the ferrichrome transporter (lines 317-318) due to a SNP, but this is not immediately obvious in the figure (Fig. 5).

Comments on revisions:

The authors have addressed the concerns from reviewers. The implemented revisions have improved the manuscript's clarity.

<https://doi.org/10.7554/eLife.101938.2.sa1>

Author response:

The following is the authors' response to the original reviews

Reviewer #1 (Public review):

Weaknesses:

It would be helpful for the authors to highlight why their technique (large scale analysis of one emm type) can yield more information than a typical GWAS analysis of invasive vs. non-invasive strains. Are SNPs easier to identify using a large-scale core genome? Is it more likely evolutionarily to find mutations in non-coding regions as opposed to the core genome and accessory genes, and this is what this technique allows? Did the analysis yield unexpected genes or new genes that had not been previously identified in other GWAS analyses? These points may need to be made more apparent in the results and deserve some thought in the discussion section.

We thank the reviewer for pointing out the importance of this study. By focusing on bacteria within a single emm type, false positives caused by confounding lineage effects can be minimized, which contributes to greater accuracy of the pan-GWAS. We have added relevant text describing the strong points of our pan-GWAS approach to the Results and Discussion sections, as shown following:

“The present pan-GWAS of bacteria within a single *emm* type minimized lineage effects, thus reducing false positives.” (lines 204–205)

The present study focused on *emm89 S. pyogenes*, known to cause increasing rates of invasive infections worldwide, and also assessed differences between *emm89* strains causing invasive and non-invasive infections. By focusing on bacteria within closed phylogenies, false positives caused by confounding lineage effects were minimized, thus contributing to a higher level of accuracy of the pan-GWAS.” (lines 420–424)

In addition, we would like to comment more regarding the reviewer’s question, “Is it more likely evolutionarily to find mutations in non-coding regions as opposed to the core genome and accessory genes, and this is what this technique allows?”. Mutations are generally considered to be more frequent in non-coding than coding regions. However, the actual mutation frequencies in both types of regions were not assessed in this study. Nevertheless, exploring non-coding regions using the k-mer method is of considerable importance, as variations significantly associated with infectious phenotypes may contribute to alterations in gene expression and other regulatory mechanisms.

The Alpha-fold data does not demonstrate why the mutations the authors identified could contribute to the invasive phenotype. It would be helpful to show an overlay of the predicted structures containing the different SNPs to demonstrate the potential structural differences that can occur due to the SNP. This would make the data more convincing that the SNP has a potential impact on the function of the protein. Similarly, the authors discuss modification of the hydrophobicity of the side chain in the ferrichrome transporter (lines 317-318) due to a SNP, but this is not immediately obvious in the figure (Fig. 5).

As the reviewer suggested, we have substituted Figure 5E in the previous version with a figure illustrating the molecular surface within proximity of the mutation. We speculated that the mutation may induce a small indentation on the surface, and thus attenuate the stability of the hydrophobic bound between FhuB and FhuD by invasion of solvent into the indentation. Additionally, images showing the wild-type and mutated models have been separated for better visibility instead of as an overlay of the predicted models suggested by the reviewer. Relevant text in the Results section and legend of Figure 5E have been accordingly revised, as shown following:

“The mutation was predicted to induce formation of a small indentation on the molecular surface, thus increasing the surface area accessible to the solvent, and is considered to potentially affect the stability of the hydrophobic bond between FhuB and FhuD, and thus ferrichrome transport (Figure 5E).” (lines 360–363)

“The 73rd valine in FhuB, shown in magenta, was substituted with alanine. The molecular surface is illustrated with a wireframe and that of the predicted indentation is shown with an arrowhead.” (lines 1162–1164)

Reviewer #1 (Recommendations for the author):

The figure legend for Fig. 3C needs to be explained so that it is similarly laid out as in Fig. 2C. Fig. 2C should indicate that the magenta color represents the invasive phenotype.

Based on this helpful suggestion, more detailed information about the magenta color representing the invasive phenotype has been added to the legends of Fig. 2C and 3C, with relevant text also included in the revised legends, as shown following:

“Colored bars above indicate countries and phenotypes, and magenta bars represent invasive phenotypes. Using the Roary program, gene names starting with “Group_” were automatically assigned. Position indicates the location of each SNP/indel on the core gene alignment. The full results are shown in Table S6.” (lines 1116–1120)

“Colored bars above indicate countries and phenotypes, and magenta bars represent invasive phenotypes. Using the Roary program, gene names starting with “Group_” were automatically assigned. The full results are shown in Table S8. (lines 1130–1133)

The wording and organization of results in the k-mer section started to get confusing around lines 270-271. It begins to be a list of results and would be better served by some interpretation or explanation of the significance (why it is important to find such mutations). For example, for mutations you find in non-coding regions, do you expect them to have a detrimental effects on gene expression/regulation?

As the reviewer kindly suggested, we have added interpretation or explanation of the significance of Comp_6 and Comp_24 to the Results section. We analyzed the function of the non-coding region of Comp_6 by employing web-based *in silico* tools, including MLDSPP and BacPP, though no promoter sequences could be identified. Next, using BLAST, a search for known promoter sequences of *S. pyogenes* M1 strain SF370 of the CDBProm database was attempted, because the web-based *in silico* promoter prediction tools are not suitable for *S. pyogenes*. However, neither identical nor homologous sequences were detected. Thus, the significance of this region remains unknown. In Comp_24, *group_141* was also identified in the COGs-based pan-GWAS as a non-invasiveness related gene. Furthermore, *group_141* showed high levels of correlation with *group_139* and *group_467*, encoding transposase and uncharacterized protein, respectively, which suggests that the presence of an MGE is associated with a non-invasive phenotype.

Relevant text has been added to the Materials and Methods (lines 653–657) and Results (lines 308–311 and 314–319) sections, as shown following:

“Promoter sequences in intergenic regions were predicted using web-based tools, MLDSPP and BacPP[29,30]. Additionally, BLAST was employed to search the promoter sequences of *S. pyogenes* strain SF370 registered in the CDBProm database (<https://aw.iimas.unam.mx/cdbprom/>)[69]” (lines 653–657)

“We speculated that this region is related to regulation of gene expression. However, no promoter sequences were identified by utilizing MLDSPP, BacPP, and BLAST, thus the significance of this region remains to be clarified[29,30].” (lines 308–311)

“Furthermore, *group_141* was also identified in the COGs-based pan-GWAS as a non-invasiveness-related gene along with *group_139* and *group_467*, which encode transposase and uncharacterized protein, respectively (Table S8 and Figure S4). Taken together, the absence of an MGE containing *group_141*, and the presence of another MGE harboring *group_142* and *group_143* may result in an invasive phenotype.” (lines 314–319)

Additionally, new references (#29, 30, and 69) concerning bacterial promoter prediction have been included in the revised version of the manuscript.

Because there is no difference in intracellular free ferric ions in the fhuB mutant compared with the wild-type, the authors speculate that the upregulation of the fhuBCD operon can compensate for the loss of function of the fhuB gene, but there is insufficient data to support this claim.

As the reviewer indicate, the data presented in the previous version were insufficient to support our speculation. Therefore, the following sentence has been deleted from the manuscript (previous version line 367):

“Therefore, the upregulation of *fhuBCD* may compensate for the impaired function mediated by SNP T218C.”

The authors mention that there was no direct association between invasiveness and acquisition of genes (lines 451-455), including antibiotic resistance genes from prophages and MGEs (lines 467-469). These data should be moved to the results section to focus the results on the correlation between invasiveness and mutation of existing DNA vs acquisition of new DNA.

Accordingly, we have added relevant text to the Results section, as shown following:

“On the other hand, the present pan-GWAS found no genes encoding known virulence factors significantly associated with invasiveness, thus further analysis of the relationships of detected distribution patterns with prophages and MGEs was performed.” (lines 264–267)

Minor spelling error at line 210 ("waws" instead of "was").

As the reviewer kindly pointed out, the spelling has been corrected. (line 233)

Reviewer #2 (Recommendations for the authors):

Minor comments:

Line 55: Does this rate apply to all types of infections?

The authors appreciate this question from the reviewer. We checked what types of infections the mortality rate is applied to and confirmed that it only represents STSS. Therefore, relevant text has been revised, as shown following:

However, even with proper treatment, the mortality rate of patients with STSS remains high, ranging from 23–81%[6]”. (lines 72–73)

Line 58: Could you explain the protein encoded by the emm gene and the role of the hypervariable region in pathogenesis?

As requested, relevant text regarding the pathogenic role of the hypervariable region of M protein has been added, as shown following:

“*S. pyogenes* has been classified into at least 240 *emm* types based on a hypervariable region sequence of the *emm* gene, which encodes the M protein. This hypervariable region of the M protein is responsible for type-specific antigenicity and binds with high affinity to C4b-binding protein, a major fluid phase inhibitor of the classical and lectin pathways of the complement system that confers resistance to opsonophagocytosis[8].” (lines 76–81)

Line 161: Figure 1C does not show the strain with the different pattern.

The authors apologize for the lack of clarity. In Fig. 1C, the strain is shown by a pale pink color bar used to indicate the related clade. For clarity, an arrowhead pointing to the strain from outside of the tree has been added along with the following text in the legend:

“Arrowhead indicates strain belonging to the novel clade.” (lines 1102–1103)

Line 239: It could be interesting to examine the genes in the region between the mobile elements found in the global cohort, as the result profile was very different from the Japanese group, which revealed more specific genes. Consider adding this to the results section.

Based on the reviewer's insightful suggestion, we attempted to find regions between the mobile genetic element-related genes. However, contigs generated from short reads were not adequate to identify such a genome structure. Therefore, calculations to analyze the pairwise correlation of the presence of significant COGs in the 666 strains to predict genes on prophages and MGEs were performed, and the results added to Figure S4. Eight clusters were detected as coexisting COG groups, seven of which comprised phage- or MGE-related genes. Furthermore, a cluster with antimicrobial-resistant genes was shown to be correlated with non-invasive infections. It is thus speculated that gain or loss of gene sets via phages and MGEs rather than acquisition of virulence genes may lead to changes in fitness to the environment and bacterial phenotypes. Relevant text has been added to the revised versions of the Results, Discussion, and Materials and Methods sections, as shown following:

"On the other hand, the present pan-GWAS found no genes encoding known virulence factors significantly associated with invasiveness, thus further analysis of the relationships of detected distribution patterns with prophages and MGEs was performed. For calculating the pairwise correlation of the presence of significant COGs in the 666 strains, the COGs were clustered into eight coexisting groups, seven of which contained phage- and/or MGEs-related genes (Figure S4). The largest group comprised 65 genes including phage proteins, while the second largest with 42 genes was found to be associated with non-invasive infections, and included *group_2689*, *group_1833*, and *ermA1*, encoding TetR/AcrR family transcriptional regulator, multidrug efflux system permease protein, and rRNA adenine N-6-methyltransferase, respectively." (lines 264–273)

"On the other hand, a cluster comprising 49 non-invasiveness-associated genes including antibiotic-resistance genes was identified. Furthermore, among the genes showing a significant correlation with the infectious phenotype, approximately 90% (152 of 169) were associated with non-invasiveness. One possible explanation is that significantly related genes reflect the process of not only gain of factors but also loss of those affecting fitness cost." (lines 517–522)

"The correlation of the presence of significant COGs was calculated and visualized using the R program." (lines 643–644)

Line 548: What cutoff values were used in Fastp?

The default cutoff value for Fastp ($Q > 15$) was used, and relevant text has been added to the Materials and Methods section in the revised version, as shown following:

"All collected sequences were subjected to quality checks using Fastp v.0.20.1, with a default cutoff value of $Q > 15$ [53]." (lines 600–601)

Line 635: Were the transcriptome experiments performed in triplicate?

We apologize for the confusion. The transcriptome experiment was performed only once with three samples for each condition. The notation "(n=3 for each condition)" has been added to the relevant text portion in the Materials and Methods section (line 696).

Discussion section: I believe the authors should place more emphasis on the fact that FhuB is associated with non-invasiveness, to provide clearer context in the discussion.

Based on this helpful suggestion, we have revised relevant text in the Discussion section, as shown following:

“Transcriptomic analysis findings suggested that the Japan-specific *fhuB* mutation associated with non-severe invasive infections contributes to the growth rate of *S. pyogenes* in human blood by adapting to the environment.” (lines 457–459)

Also, “V73A” has been removed from the relevant text in the Discussion section to provide a more clear and precise context, with the revised sentence shown following:

“Two possible roles of the FhuB mutation in the pathogenesis of severe invasive infections are thus proposed.” (lines 470–471)

<https://doi.org/10.7554/eLife.101938.2.sa0>