












Emergence of distinct *Streptococcus pyogenes emm1* and *emm12* lineages in China

Received: 29 January 2025

Accepted: 13 March 2026

Published online: 23 April 2026

 Check for updates

Yuanhai You ^{1,17}✉, Dingle Yu^{2,17}, Chao Yang ^{3,17}, Xiaomin Peng^{4,17}, Ouli Xie ^{5,6,17}, Hesheng Chang^{7,17}, Chunzhen Hua^{8,17}, Fei Zhao¹, Xiaomei Yan¹, Menghan Zhang⁹, Xinwei Ruan ⁵, Jasmine E. J. Wells¹⁰, Stephan Brouwer ¹⁰, Camila Duitama González ¹¹, Ming Fang¹², Xiaojie Yu¹³, Lu Sun¹, Xiaoyue Wei¹, Jie Liu¹, Daitao Zhang⁴, Lihua He¹, Jiazheng Wang¹⁴, Chuyang Sun¹³, Yuejie Zheng², Sebastian Duchene^{5,11}, Mingming Zhou⁸, Lifang Sun², Mark R. Davies ⁵✉, Mark J. Walker ¹⁰✉, Quanyi Wang ⁴✉, Jianzhong Zhang ¹✉ & Yonghong Yang ^{15,16}✉

Cases of scarlet fever have increased since 2011 across China. However, genomic epidemiological knowledge of *Streptococcus pyogenes*, the causative agent, is limited. Here we present a longitudinal analysis of *S. pyogenes* isolates ($n = 1,029$) across *emm1* and *emm12* genotypes collected from eight provinces across China between 1993 and 2024. Genomic data integrated with national scarlet fever incidence data confirmed *emm12* and *emm1* as dominant genotypes underlying five incidence peaks and disease resurgence in 2024. Phylogenetic analysis showed independent evolution of these genotypes in China compared to global epidemic lineages. Four *emm12* clades were present in China before 2011 but were replaced by a single clade, Clade II, by 2020. A dominant *emm1* clade, M1_{china}, distinct from global lineages and the M1_{UK} lineage, represents >98% of *emm1* cases in China. Sub-clade expansion coincides with carriage of integrative conjugative elements containing macrolide and tetracycline resistance genes and virulence gene-encoding prophage. Ongoing maintenance of these elements in *emm1* and *emm12* populations likely underlies the resurgence of scarlet fever in China.

Streptococcus pyogenes (group A *Streptococcus*; GAS) causes a variety of human diseases ranging from mild infections to deadly diseases. The most common diseases include streptococcal pharyngitis, impetigo and scarlet fever¹. Globally, the incidence of GAS diseases such as scarlet fever had been in decline since the widespread introduction of antibiotics^{2,3}. In the 1980s (ref. 4,5) and again after the COVID-19 pandemic (ref. 6), invasive GAS disease has resurged in Western countries. During the 2010s, the incidence of scarlet fever witnessed a sharp increase in both Asian and European countries^{6–10}.

A new pandemic *emm1* GAS lineage, designated M1_{UK}, associated with increased incidence of scarlet fever and invasive disease, was

reported in 2019 with retrospective analysis indicating first emergence in 2008 (ref. 6). This lineage has disseminated to other high-income countries, including the USA and Australia^{6,11–14}. In Europe, Australia and the USA^{15–18}, there was a sharp rebound in the incidence of invasive GAS disease in 2022 after lifting of COVID-19 social distancing measures^{19,20}. In a period of 12 weeks to 7 December 2022, a total of 6,600 cases of scarlet fever were reported in the UK, with 652 cases of invasive GAS infection also reported and 60 deaths¹⁵.

In China, scarlet fever resurgence began in 2011, with fluctuations at a high incidence until 2019. Scarlet fever cases reduced by 80% during the COVID-19 pandemic²¹. In 2023, 25,819 cases were reported to

A full list of affiliations appears at the end of the paper. ✉e-mail: yuyuanhai@icdc.cn; mark.davies1@unimelb.edu.au; mark.walker@uq.edu.au; bjcdcxm@126.com; zhanngjianzhong@icdc.cn; yuh628628@sina.com

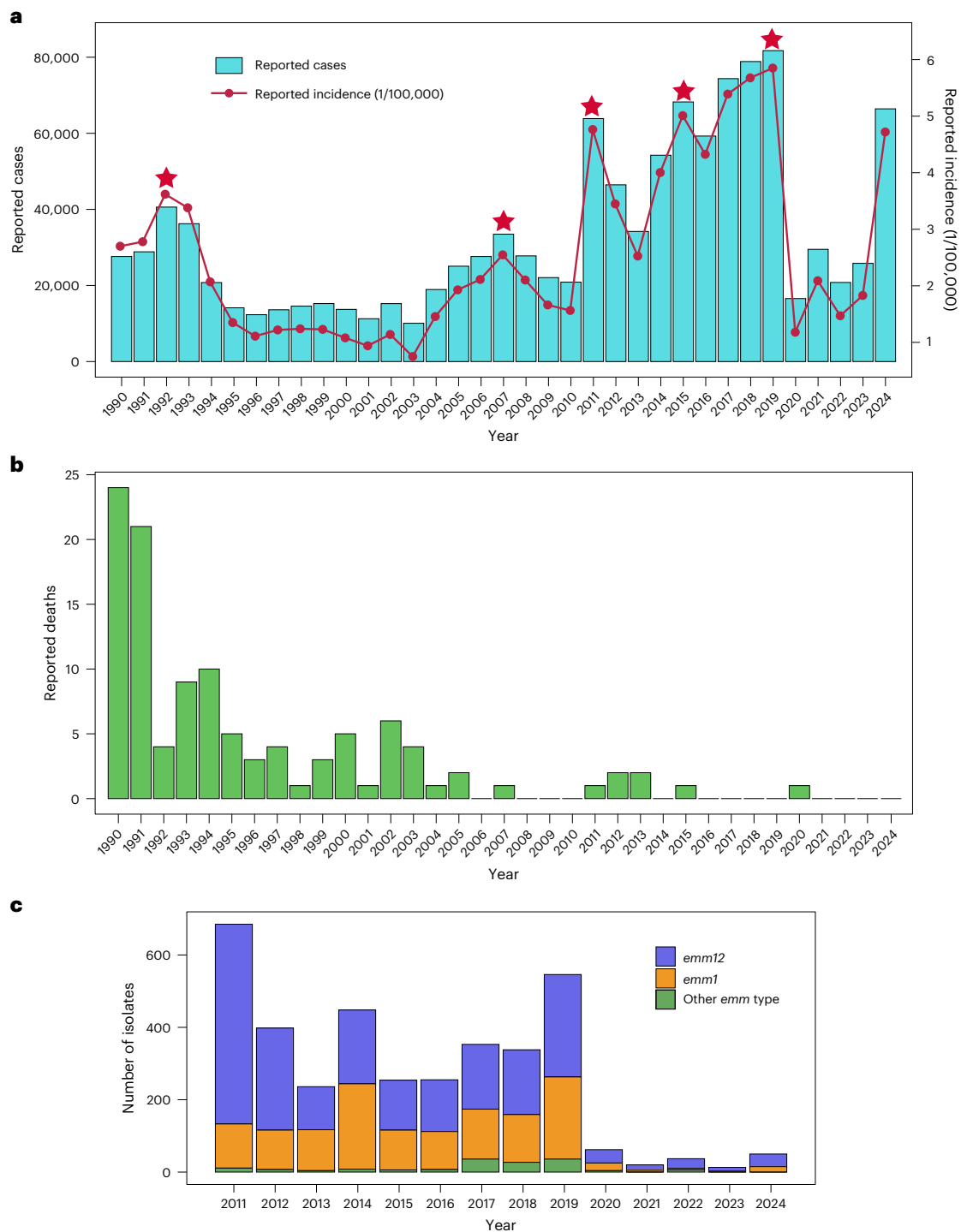


Fig. 1 | Scarlet fever notifications and *emm* type trends in China. a, b, Scarlet fever number of cases and incidence (a) and reported deaths from scarlet fever (b) between 1990 and 2024 in China. The five scarlet fever incidence peaks are

indicated (stars). Scarlet fever cases and deaths were reported to the Chinese NNIDSS. c, Genome-derived GAS *emm* types from scarlet fever and tonsillitis infection from China during 2011–2024 (ref. 23).

the Chinese Center for Disease Control and Prevention (CDC), and the incidence increased by 24% compared with that of 2022 (ref. 21). With a rebound in case numbers also expected following the post-COVID-19 lifting of social distancing restrictions²², pathogen surveillance will be critical for scarlet fever control. In previous work, scarlet fever in China was caused predominantly by both *emm12* and *emm1* strains^{10,23}. These *emm* types had acquired an extended bacteriophage-encoded toxin gene repertoire of superantigens streptococcal superantigen (SSA) and streptococcal pyrogenic exotoxin C (SpeC), and the streptococcal phage

DNase 1 (Spd1). The presence of ICE-*emm12* and ICE-HKU397 elements harbouring macrolide and tetracycline resistance genes suggested a role for antibiotic use in selection and expansion of scarlet fever lineages in China. The available genomic data for mainland China strains was limited in these studies^{7,8,10,24,25}. The detailed population structure of GAS strains from mainland China is largely unexplored, and more broadly there are limited GAS surveillance and molecular epidemiology data from Asia.

In this Article, we used a combination of epidemiological and pathogen genomic approaches to comprehensively analyse the

changes in clonality of *emm1* and *emm12* populations between 1993 and 2024, which underpin elevated rates of scarlet fever in mainland China since 2011, and define genetic factors driving emergence of scarlet fever-causing GAS lineages.

Results

Scarlet fever incidence between 1990 and 2024

Over the course of 1990–2024, there were five scarlet fever incidence peaks evident in the National Notifiable Infectious Disease Surveillance System (NNIDSS) incidence data, occurring in 1992, 2007, 2011, 2015 and 2019. Between 1990 and 2010, the incidence remained relatively low at 0.75–3.62 per 100,000 population per year. Resurgence of scarlet fever occurred in 2011 and fluctuated at a higher incidence until 2019, before declining across 2020–2023. The lowest incidence was observed in 2003 (0.75 cases per 100,000 population, 95% confidence interval (CI) 0.60–0.93) and the highest in 2019 (5.85 per 100,000 95% CI 5.82–5.88). Scarlet fever has resurged again in 2024 (Fig. 1a). We report 111 deaths in China caused by scarlet fever across this period. Scarlet fever cases resulting in death were highest in 1990 and 1991, with only 0–2 deaths per year reported across 2011–2024 (Fig. 1b). A consistent pattern of molecular types was observed with *emm12* and *emm1* molecular types being identified in high numbers of nationally reported cases from the sampled provinces in this study during 2011–2024 (Fig. 1c).

Population structure and clonal shift within *emm12* causing scarlet fever

Expanding on our earlier *emm12* population framework^{7,8}, five predominant global *emm12* lineages, designated as *emm12*-Clade I through to *emm12*-Clade V, were evident before the 2011 scarlet fever outbreak (Fig. 2a). To further refine the evolutionary associations between Clade I–V lineage evolution and mobile genetic elements (MGEs), we systematically extracted 3,275 MGEs from 904 *emm12* genomes. Of these, 11 distinct MGE clusters (as defined by Jaccard similarity 0.9) were carried in a minimum of 10 genomes.

From a global view of the *emm12* population structure, *emm12*-Clade IV consisted of strains primarily from Western Europe and North America, with only a few isolates from China belonging to this clade. The most differential MGE feature of this clade was carriage of the prophage Φ MGAS9429.1 (Fig. 2a). No lineage-defining integrative conjugative elements (ICEs) were detected in this clade. *emm12*-Clade III is the oldest lineage within the China *emm12* population in this study, estimated to have diverged before 1940 (95% credible interval (CrI): 1934–1946). Within *emm12*-Clade III, 57 out of 67 (85%) of strains were from China. This clade was detected between 1993 and 2015 and with low frequency after 2011 (7 out of 282, 2%) which was mirrored by a decline in effective population size in the late 2000s (Fig. 2b,c). The majority of this clade harboured the prophage Φ HKU16.vir (50 out of 57 (88%) of genomes from China, 0 out of 10 of international genomes) carrying the superantigen genes *ssa*, *speC* and DNase gene *spd1*. *emm12*-Clade III also harboured ICE-HKU397 (30 out of 57 (53%) of genomes from China, 0 out of 10 international genomes) or ICE-1993BJGAS10 (10 out of 57 (18%) of genomes from China, 0 out of 10 international genomes), which both carried the multidrug resistance genes *tetM* and *ermB* via a module which was highly homologous across different ICE backbones⁷ (Extended Data Fig. 3). In China, *emm12*-Clade I and *emm12*-Clade III were the dominant lineages before 2011 (63 out of 82, 77%) before the surge in scarlet fever notifications. Isolates from these subclades carried the superantigen gene *ssa* on the previously described prophages Φ HKU.ssa and Φ HKU16.vir and antimicrobial resistance genes *tetM* and *ermB* on integrative conjugative elements ICE-HKU397 and ICE-1993BJGAS10 (ref. 7). The ICE-HKU397 element was not present in the selected international *emm12*-Clade I genomes (Fig. 2a). Several genetically distinct prophage carrying the *speC* and *spd1* virulence factors were variably carried in the globally disseminated *emm12*-Clade I, *emm12*-Clade III and *emm12*-Clade IV

populations (Fig. 2a and Extended Data Fig. 4). *emm12*-Clade V mainly included strains from earlier in the study period in 1993 and 1994, two isolates from Guizhou in 2005 and one isolate from Hong Kong in 2005. This clade was infrequently isolated before 2011 (6 out of 82, 7%) and carried *tetM* and *ermA* on ICE-HKU165 (Extended Data Fig. 3).

In our dataset, from 2011 there was a rapid expansion of the *emm12*-Clade II in China (339 out of 404, 84%), which displaced the preexisting *emm12*-Clades I, III, IV and V by 2018 (Fig. 2b). The expansion of this clade (Fig. 2b,c) coincided with the surge in scarlet fever notifications (Fig. 1). The progenitor of the *emm12*-Clade II population in China is also evident in publicly available genomes from North America, Western Europe and Oceania. However, the mainland China strains could be divided into two subclades distinct from global strains and are characterized by both the presence of the previously described prophage Φ HKU16.vir and the presence of two different ICEs, ICE-*emm12* and ICE-HKU397, which carry the multidrug resistance genes *tetM* and *ermB*⁷ (Fig. 2a). The two subclades were predicted to have a common ancestor in 1998 (95% CrI: 1994–2001) and showed a dramatic increase in effective population size after 2010 coinciding with the peak in cases in 2015. An ancestral strain isolated in 1993 in China (1993GAS18) and other international *emm12*-Clade II genomes did not carry antimicrobial resistance genes. ICE-*emm12* was unique to strains isolated after 2011. The carriage of ICE carrying macrolide and tetracycline resistance genes in recently expanded modern *emm12* clades in China support acquisition of *tetM* and *ermB* on ICE as associated with the success of these clades in China. ICE-*emm12*, ICE-HKU397 and ICE-1993BJGAS10 show >98% nucleotide identity in an ~20 kb (kilobase) Tn916-like genetic module encompassing the *ermB* and *tetM* genes (Extended Data Fig. 3), supportive of dispersal into different ICEs and clades in China.

Toxin gene carriage and expression in representative strains from *emm12*-Clade I (SP1177), *emm12*-Clade II (SP1203) and *emm12*-Clade III (NS488) was assessed (Extended Data Fig. 6). SpeC, Spd1 and streptolysin O (SLO) toxin expression was equivalent across strains representative of each clade, while Clade II and III strains expressed SSA.

Population structure and clonal shift of lineages within the *emm1* genotype

The population structure of the *emm1* population was investigated based on 385 isolates collected between 1993 and 2024 from China and 467 from other geographic locations, containing 3,217 MGEs, of which 9 distinct MGEs (Jaccard similarity of 0.9) were carried in a minimum of 10 genomes. Analysis of *emm1* clinical strains from China collected after 2011 showed that 302 out of 307 (98%) belong to a distinct lineage within the M1_{global} genetic population. This lineage was not evident in other geographical locations sampled over the same period. We therefore refer to this population as the 'M1_{china}' lineage (Fig. 3a). M1_{china} shares a common ancestor with the M1_{global} population as determined by the hallmark evolutionary features of the M1_{global} population (alternatively designated the MIT1 clone), namely, the presence of the *speA2* containing prophage Φ MGAS5005.1, in addition to the *spd3* prophage Φ MGAS5005.2 and *sdaD2* Φ MGAS5005.3, and the 36 kb *purA-nadC* recombination region. M1_{china} likely emerged from a single ancestor shortly after the M1_{global} expansion and has since diverged into three dominant subclades that we have defined as subclades 1, 2a and 2b, which remain clinically relevant in our study period (Fig. 3a). The relative clinical proportion of these M1_{china} subclades has fluctuated over time, with subclade 2a and subclade 2b emerging as the most common subclades since 2010 (Fig. 3b). The inferred effective population size of all three subclades rapidly expanded in the early 2000s coinciding with the resurgence of scarlet fever in China. The three M1_{china} subclades are defined by the presence of two multidrug-resistant ICEs conferring macrolide and tetracycline resistance, termed ICE-HLJGAS2022 (subclade 1) and ICE-HKU397 (subclades 2a and 2b). ICE-HKU397 has also been in circulation in the *emm12* population in China and shares over 99% similarity to ICE-HKU488 (Extended Data Fig. 3). While the

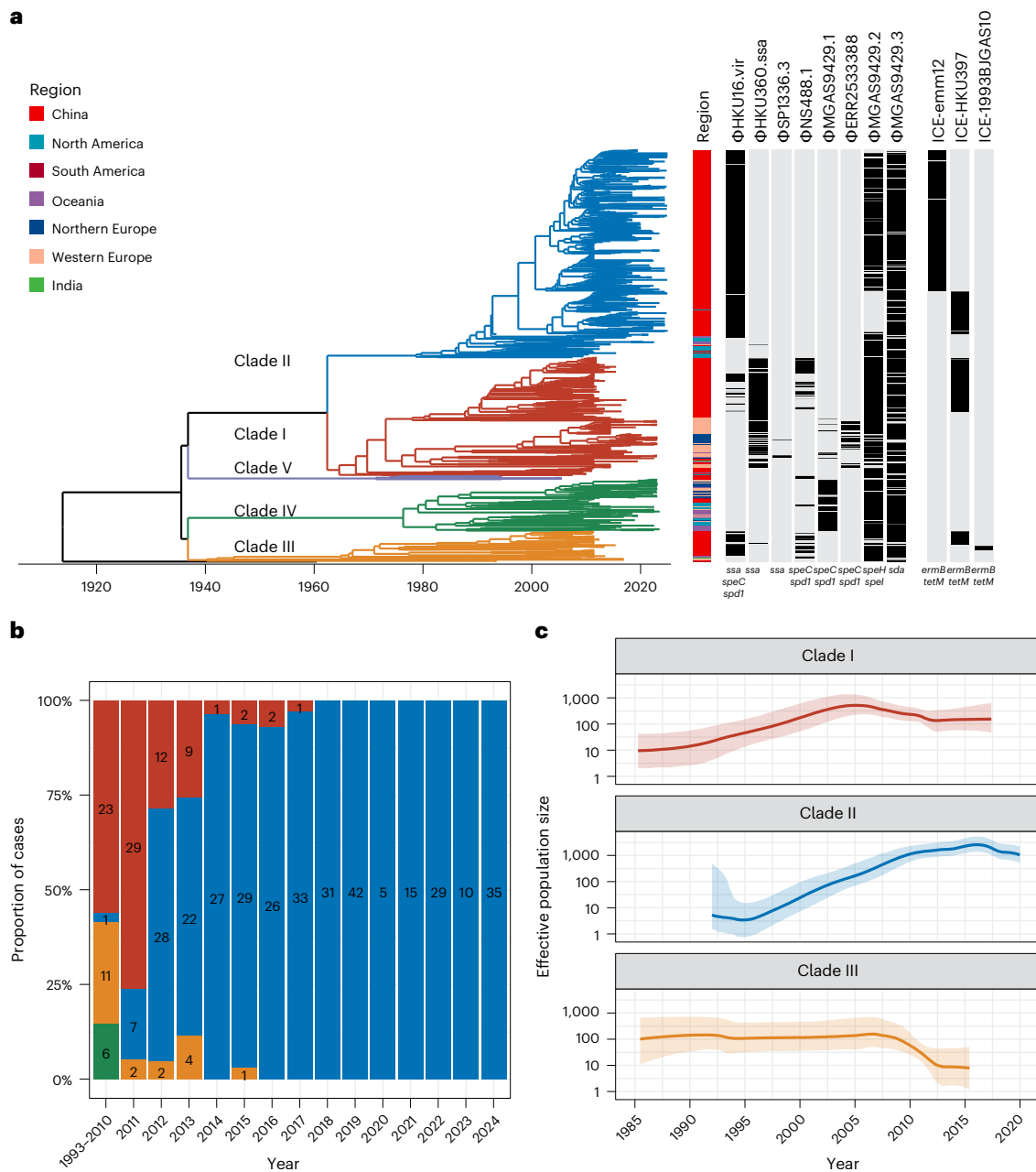


Fig. 2 | Population structure and shifts in lineages within *S. pyogenes emm12* from China. a, Timescaled phylogeny of *emm12* genomes from this study ($n = 444$) and previously published global genomes ($n = 460$). The branches of the tree are coloured by designated Clades I–V. The region of isolation and the presence of MGEs of interest are indicated by the heat map (black blocks indicating presence). Only MGE present in ≥ 10 genomes are represented. **b**, Proportion of genomes from *emm12* clades between 1993 and 2024 in China. The number of genomes in each clade per year are labelled in the respective

bars. Genomes from 1993–2010 are grouped as a single category owing to comparatively limited sampling during this period. **c**, Median effective population size of *emm12* Clades I–III using only genomes isolated from China (from this and previously published studies). The 95% highest probability distribution is represented by the shading. The effective population size is inferred from coalescent events to the most recent sample resulting in different time ranges between the clades.

backbone of ICE-HLJGAS2022 is different from that of ICE-HKU397, both share the 20 kb Tn916-like module encompassing *ermB* and *tetM*, similar to the ICE found in the modern *emm12* population in China (Extended Data Fig. 3). Nearly all modern clinically relevant clones of these subclades also carry the *ssa*, *speC* and *spd1* genes harboured on prophage Φ HKU488.vir (351 out of 379 (93%) of $M1_{\text{china}}$ genomes), which likely entered the population during the late 1980s (Fig. 3a). This prophage is a genetic homologue of Φ HKU16.vir in *emm12* (ref. 24) (Extended Data Fig. 5). In a similar context to *emm12* scarlet-fever-associated isolates in China, these data show the ongoing

expansion of multiple subclades is associated with virulence encoding and multidrug-resistant MGEs.

While $M1_{\text{UK}}$ has rapidly emerged as the dominant *emm12* clone in many continents and countries including Western Europe, North America and Australia, only a single $M1_{\text{UK}}$ isolate from China was reported in this dataset (GAS1916 (SF*emm12*-2018))²⁶. We report no additional $M1_{\text{UK}}$ strains in our surveillance study, which suggests that the $M1_{\text{UK}}$ lineage has yet to establish and spread widely in China.

Toxin repertoire of a representative $M1_{\text{china}}$ strain (SP1165) was compared to $M1_{\text{global}}$ (reference strain 5448)²⁷ and $M1_{\text{ancestral}}$ (reference

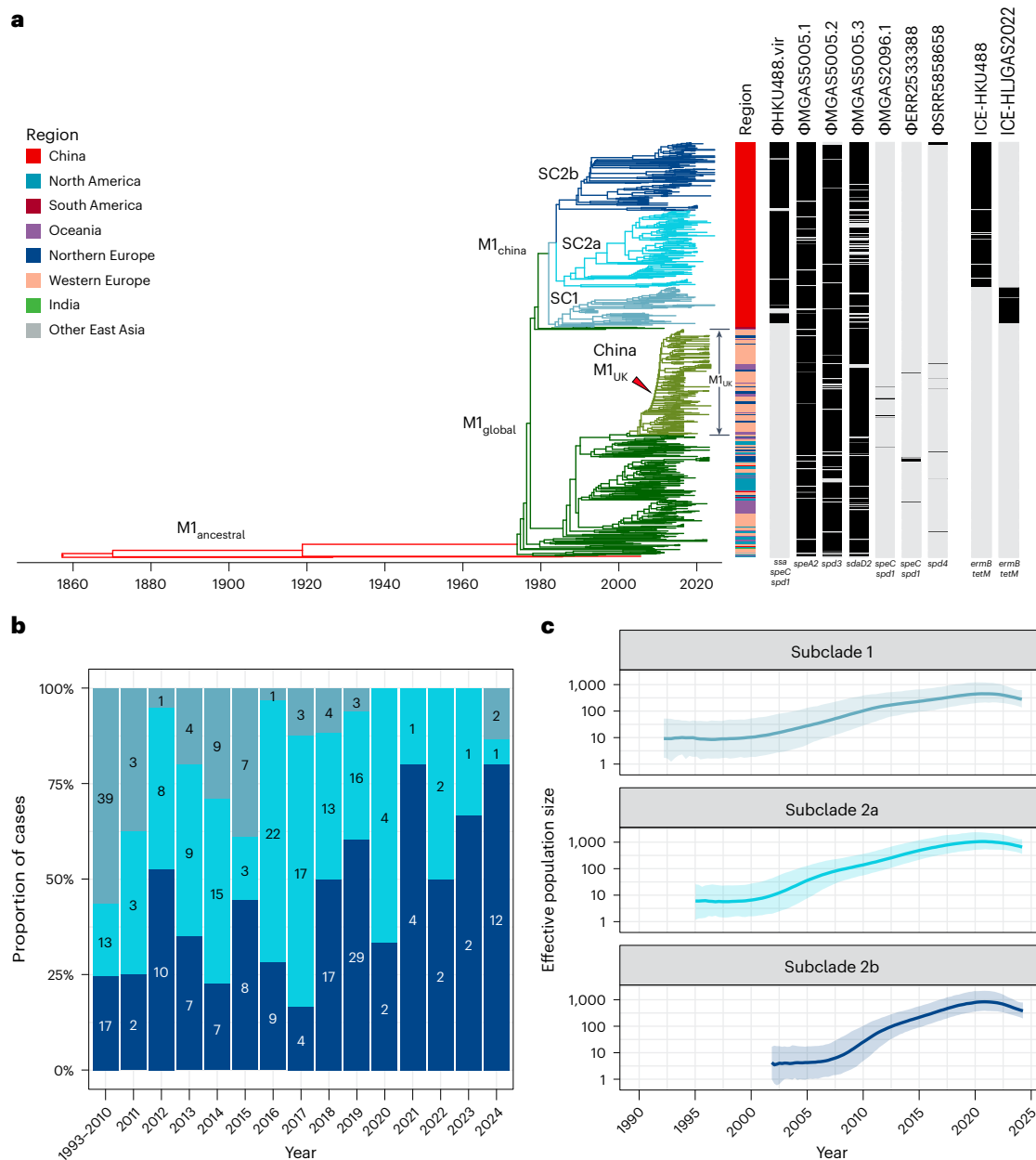


Fig. 3 | Population structure and clonal shift in lineages within *emm1* GAS from China. **a**, Timescaled phylogeny of *emm1* genomes from this study ($n = 337$) and previously published global genomes ($n = 515$). The branches of the tree are coloured as designated by temporally relevant lineages. The emergence of M1_{china} lineages (blue) are colour coded by subclades (SC1–3). The region of isolation and the presence of MGEs of interest are indicated by the heat map (black blocks indicating presence). Only MGE present in ≥ 10 genomes are represented. **b**, Proportion of *emm1* clinical cases in China as colour coded by genomic

subclade between 2011 and 2020. The absolute number of isolates are provided within the bars. Genomes from 1993–2010 are grouped as a single category owing to comparatively limited sampling during this period. The red arrowhead refers to the position of the single M1_{UK} isolate from China. **c**, Median effective population size of *emm1* subclades 1, 2a and 2b. The 95% highest probability distribution is represented by the shading. The effective population size is inferred from coalescent events to the most recent sample resulting in different time ranges between the clades.

strain SF370)²⁸ (Extended Data Fig. 7). SLO toxin expression was equivalent across each representative strain. SSA was only expressed by the M1_{china} strain SP1165, which also expressed SpeA levels equivalent with 5448, and SpeC and Spd1 equivalent to SF370.

Discussion

Scarlet fever surged in China in 2011 and continued on an upward trajectory until 2020, when COVID-19 social distancing measures²² introduced in 2020 correlated with reduced scarlet fever disease burden. The relaxation of social distancing in a number of Western countries, exemplified by the UK, coincided with a significant rebound of scarlet

fever and invasive disease caused by GAS, driven in large part by the *emm1* GAS strain M1_{UK}²⁹. Although M1_{UK} has been identified as circulating in China²⁶, significant disease causation attributed to M1_{UK} has yet to be detected.

The most common *emm* type circulating in China associated with scarlet fever is *emm12*. Initial phylogenetic analyses on 2011 scarlet fever resurgence strains described the multiclinality of *emm12* outbreak strains which were associated with the acquisition of multidrug resistance and toxin-carrying prophage^{7,8,10}. The incorporation of clinical isolates from 1993–2024 has now revealed a dramatic increase in *emm12*-Clade II, which over a short period has emerged as the sole

dominant *emm12* clone throughout mainland China. Analysis of *emm12*-Clade II supports the observation that carriage of Φ HKU.vir (carrying the toxins genes *ssa*, *speC* and *spd1*) and ICE carrying *ermB* and *tetM* have played a key role in the ongoing expansion of these subclades. Analysis of *emm12*-Clade III suggests that although this clade acquired ICE before the 2011 scarlet fever resurgence, it did not become a dominant clone during the years 1993–2024. Instead, a conserved genetic Tn916-like module containing *ermB* and *tetM* has disseminated across multiple clades and both *emm12* and *emm1* populations on different ICE backgrounds. Thus, the 2011 scarlet fever outbreak was mainly characterized by the co-existence of three *emm12* lineages and emergence of *emm12*-Clade II as a major outbreak driver. Identification of the population structure and clonal shifts of GAS lineages causing scarlet fever in China is crucial for future domestic and international surveillance.

Globally, *emm1* is the most widespread and common genotype in Western countries²⁹ and the second most common cause of scarlet fever in China²³. Since the 1980s, the M1_{global} outbreak lineage (also known as the MIT1 clone) became the worldwide dominant clone for about 30 years. From 2008 (ref. 6), M1_{global} has been gradually replaced by the emergent M1_{UK} clone in Western countries, hypothetically owing to higher pathogenicity and transmission ability^{1,6,11}. Of the 385 *emm1* genomes from China represented in this study, only a single strain was found to belong to M1_{UK}, suggesting that M1_{UK} has yet to clonally expand in China. There has been a lag phase between the first identification of M1_{UK} in a national jurisdiction and when M1_{UK} became the dominant *emm1* strain. For instance, M1_{UK} was identified in the UK in 2008, where it became the dominant *emm1* lineage in 2013 (ref. 6). In Australia, M1_{UK} was identified in 2013 but only became the dominant *emm1* lineage in 2017 (ref. 11). In China, M1_{UK} was identified in 2015 (ref. 10,26) but has yet to become the dominant *emm1* lineage. This period includes the reduction in the number of cases of scarlet fever during the COVID-19 pandemic period of 2020–2023 (Fig. 1a). As GAS isolates and disease rebound in China after the COVID-19 pandemic (as has happened in all other national jurisdictions) we hypothesize that M1_{UK} numbers will increase. China CDC is closely monitoring the situation. Nonetheless, it is an open question as to why M1_{UK} has yet to replace dominant *emm1* lineages in China and whether environmental, host genetic or immune factors may also have a role in such an expansion.

Although evolutionarily related to the archetypical M1_{global} lineage that emerged during the 1980s, we identified expansion of an endemic M1_{global} lineage in China designated M1_{china}, which is evolutionarily distinct from lineages reported in other countries. The delineating features of M1_{china} compared to other M1_{global} lineages is the acquisition of multidrug-resistant ICE in addition to the virulence-associated *ssa*, *speC* and *spd1* encoding prophage (Φ HKU488.vir)²⁴. We hypothesize that the maintenance of these elements within the population indicates that they have a key role in the diversification and expansion of *emm1* genotypes in China. This suggests a different pathogen population dynamic between China and Western countries. There are at least two historical lineages within the GAS *emm1* population in China, yet these have been replaced by the M1_{china} clone following its emergence in the 1980s. Φ HKU488.vir and ICE have both been circulating in the historical lineages of *emm1*, yet these lineages have not experienced large-scale clonal expansion.

For the five scarlet fever incidence peaks observed during 1993–2020, we can now report the associated epidemic lineages. By combining analyses of the shifts in *emm12* and *emm1* clonal patterns, we propose that the 1993 peak was mainly caused by M1_{china} subclade 1. The 2008 peak was caused by M1_{china} subclade 2a, M1_{china} subclade 2b and *emm12*-Clade I. The 2011 peak was mainly driven by the emergence and expansion of *emm12*-Clade II. Across 2014–2019, several lineages within *emm12* and *emm1* contributed to the high incidence during this period, the major lineages being *emm12*-Clade II, M1_{china} subclade 2a and M1_{china} subclade 3a. During the period of COVID-19 restrictions in China

between 2020 and 2023, there was decline in incidence of scarlet fever including *emm1* and *emm12* incidence. In 2024, there was a rebound in incidence to levels comparable to pre-pandemic levels. We did not observe a substantial shift in population structure of *emm1* and *emm12* strains causing disease in 2024. However, continued surveillance is required given the unprecedented rebound in GAS disease incidence in other countries after the lifting of COVID-19 restrictions and emergence of M1_{UK} in many geographical regions.

Following 2011, the increased number of cases of scarlet fever in China was largely caused by *emm12*-Clade II. Several factors that may be associated with this resurgence have been hypothesized previously, including genomic variation within the GAS population, changing herd immunity, increased surveillance capability, climate change and the relaxation of the China second-child policy³⁰. Some of these hypotheses lack strong supporting evidence, such as the lifting of the second-child policy, where we have demonstrated there is no linkage between scarlet fever increase and the newborn population size³¹. We hypothesize that the likely factors responsible for the 2011 scarlet fever upsurge are changing bacterial genetic features and human herd immunity. Limitations of this study include the lack of broader monitoring of the scarlet fever-causing GAS population in China across all geographic jurisdictions and the lack of prospective cohort sampling of serum samples to assess herd immunity of scarlet fever patients before and after infection. The lack of carriage or other clinical isolate sampling means we cannot definitively extrapolate these observations to the entire GAS population circulating in China. Further monitoring of the GAS population in China is warranted, given the risk the M1_{UK} clone has posed in other countries and the maintenance of multidrug-resistant GAS clones in circulation.

Methods

Bacterial clinical isolates and ethical considerations

A total of 781 GAS isolates were collected from patients at the Beijing CDC, Beijing Children's Hospital and six other institutions across China between 1993 and 2024. The majority of isolates (85%, $n = 664$) were recovered from throat swabs, with the remainder obtained from a range of other clinical specimens, including sputum, vaginal secretions, pus and bronchoalveolar lavage fluid (a full list is provided in Supplementary Table 1). This study was conducted in accordance with the Declaration of Helsinki. The protocol was approved by the Ethics Committee of the National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control (approval no. ICDC-2023003). The requirement for written informed consent was waived as all anonymized data were available from routine clinical care.

Notifiable GAS disease cases

In China, scarlet fever is the only GAS disease reported to the NNIDSS, listed as a class B notifiable disease since 1950. Cases of suspected scarlet fever are notified by clinicians to NNIDSS on the basis of symptoms consistent with scarlet fever, with or without laboratory confirmation of GAS infection. Here we report incidence data from NNIDSS from 1990 to 2024.

Collection of strains during 1993–2024

GAS isolates were collected from eight provinces across China from 1993 to 2024, with 516 out of 781 isolates collected in the Beijing region. GAS isolates were cultured and stored frozen in tryptone soy broth. Isolates were sent to the China CDC GAS laboratory and cultured on Columbia blood agar (Oxoid) at 37 °C with 5% CO₂. PCR-derived *emm* type data of GAS from eight provinces was extracted from a previous systematic review which revealed the dominance of *emm12* and *emm1* GAS between 1993 and 2020 in China, and thus we focused on these *emm* types for further whole-genome analysis²³. In total, we retrieved 337 *emm1* and 444 *emm12* clinical isolates between 1993 and 2024. These were combined with 248 previously published genomes from

China (48 *emm1* and 200 *emm12*). Of the 1,029 total genomes from China, 160 strains were collected before the 2011 resurgence, allowing us to trace the origin of epidemic clones and analyse the dynamic shift in molecular patterns during this long-term investigation. These strains geographically represented the high-incidence regions of China: the Northern region, the Eastern region and the Southern region. In total, 85% of newly sequenced strains were isolated from throat swabs (661 out of 781 isolates; Supplementary Table 1). The 1,029 genomes from China were contextualized with 727 previously published global sequences (260 *emm12* and 467 *emm1*)^{6–8,24,31,32} (Supplementary Table 1).

emm typing and genome sequencing

GAS isolates were *emm* typed using standard protocols³³. In brief, purified GAS DNA (QIAamp DNA Mini Kit, QIAGEN) was PCR amplified using forward primer (5'-TATTAGCTTAGAAAATTA-3') and reverse primer (5'-GCAAGTCTTCAGCTTGTTT-3'). The following PCR cycling conditions were used: denaturation at 94 °C for 1 min, 30 cycles of 94 °C for 15 s, annealing at 47 °C for 30 s, extension at 72 °C for 1 min 25 s and elongation at 72 °C for 7 min. Sequencing of PCR products was undertaken using standard procedures (bigdye V3.1, Applied Biosystems). Sequences were analysed with the Blast2.0 server hosted on the US CDC website (<https://www2a.cdc.gov/ncidod/biotech/strepblast.asp>). *emm12* and *emm1* isolates were subject to genome sequencing using BGI MGISEQ-2000 with 150 bp paired end reads (Supplementary Table 1). Four representative isolates from dominant lineages in China were selected for sequencing using the Oxford Nanopore long-read platform to enable MGE characterization: M1_{china} isolate SP1165 (BJCYGAS59), SP1177 (HLJGAS2012) from *emm12*-Clade I, SP1203 (QD62) from *emm12*-Clade II and NS488 from *emm12*-Clade III. DNA was extracted using the Monarch Spin gDNA Extraction Kit (New England Biolabs). Sequencing libraries were prepared with the Native barcoding kit (SQK-NBD114.96) and sequenced using the Nanopore platform with R10.4.1 flow cells to an approximate depth of >100 and base-called using Dorado v0.4.2 (dna_r10.4.1_e8.2_400bps_sup@v5.0.0). The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive³⁴, in the National Genomics Data Center³⁵, China National Center for Bioinformatics/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA033131), which are publicly accessible at <https://ngdc.cncb.ac.cn/gsa>. Accession numbers for individual isolates and publicly available Illumina reads are listed in Supplementary Table 1 (refs. 36,37). Long-read Oxford Nanopore Technologies (ONT) data are available at the National Center for Biotechnology Information (NCBI) under the BioProject PRJEB103775.

Genome analysis

Illumina genome sequences were quality controlled using a previously described pipeline. In brief, species assignment was confirmed from reads using Kraken2 v2.1.2³⁸. Reads with >5% reads assigned to another species were excluded owing to suspected contamination. Genomes were assembled using Shovill v1.1.0 (<https://github.com/tseemann/shovill>) with SPAdes assembler v3.14.0 (ref. 39). To expand the genome collection from China, previously published genomes from China where only assemblies were available were also included^{6–8,24,31,32}. Only genomes with <200 contigs (mean 54, range 19–172), total assembly size 1.75–2 Mb (mean, 1.87 Mb, range 1.78–1.97 Mb), and guanine-cytosine content (GC%) between 38% and 39% (mean 38.3%, range 38.0–38.4%) were included. The mean N50 was 111,681 bp (range 64,455–327,809 bp). Oxford Nanopore sequence reads were filtered by Filtlong v0.2.1 (<https://github.com/rrwick/Filtlong>) to remove the reads ≤2,000 bp and retain the top 90% of reads based on quality scores. The filtered sequences were assembled with Flye v2.9.5 (ref. 40). Assemblies were annotated using Prokav1.14.6 (ref. 41). The *emm* type was inferred from assemblies using emmtyper v0.2.0 (<https://github.com/MDU-PHL/emmtyper>). Multilocus sequencing typing (MLST) was inferred using MLST v2.23 (<https://github.com/tseemann/mlst>), using alleles from the

seven loci: *gki*, *gtr*, *murl*, *mutS*, *recP*, *xpt* and *yqil*. The sequence types were determined by PubMLST database for *S. pyogenes*⁴².

Single-nucleotide polymorphism alignments were generated by alignment of BGI short reads against reference genomes MGAS5005 (NC_007297.2) and HKU16 (NZ_AFRYO1000001.1) for the *emm1* and *emm12* populations, respectively, using Snippy v4.6.0 (<https://github.com/tseemann/snippy>) with MGE regions masked. For previously published genomes from China where only assemblies were available, the ‘-contigs’ option in Snippy was used to map the contigs to the reference. Regions in the alignment predicted to be affected by recombination were masked using Gubbins v3.4 with IQ-tree v3.0.1 for phylogeny inference using a general time reversible model and a discrete Gamma model⁴³ with 4 rate categories (G4) for rate heterogeneity. The minimum single-nucleotide polymorphisms to identify a recombination block was set at 5 with a minimum window size of 100 bp and maximum window size of 10,000 bp.

Timescaled phylogenies and inference of effective population size

Bayesian timescaled phylogenies were inferred by BactDating v1.1.4 using recombination-mased phylogenies inferred by Gubbins^{44–46}. Both the *emm1* and *emm12* timescaled phylogenies were inferred using an additive relaxed clock model⁴⁷, coalescent prior and 10⁷ iterations to ensure Markov chain Monte Carlo convergence with parameter effective sample size (ESS) > 200 after 50% burn-in (Extended Data Figs. 1 and 2). Effective population size (a measure of genetic diversity) of *emm1* and *emm12* clades in China was inferred by fitting a Bayesian coalescent skyride model using BEAST (v1.10.5)⁴⁸. To assess the expansion of these clades in China, the analysis was conducted using only genomes isolated from China (from this and previously published studies listed in Supplementary Table 1). The full Bayesian model consisted of a relaxed molecular clock model with an underlying lognormal distribution and Hasegawa–Kishino–Yano nucleotide substitution model⁴⁹ with a discrete gamma model (G) using 4 categories for rate heterogeneity. The prior for the substitution model was a continuous-time Markov chain-rate reference prior, which is a Gamma distribution with mean 0.5/tree length^{50,51}, while the remaining parameters had the default prior distribution in BEAST. The posterior distribution was sampled using Markov chain Monte Carlo with chain length of 3.5 × 10⁸ steps, sampling every 5 × 10⁴. Sufficient sampling was determined by conducting one replicate of each analysis and by verifying that the ESS of key parameters was at least 200.

Detection of virulence and antibiotic-resistance-related MGEs

MGEs were systematically extracted adapting a previously described pangenome pipeline^{36,37}. In brief, a combined pangenome including all *emm1* and *emm12* genomes was inferred using Panaroo v1.5.2 (<https://github.com/gtonkinhill/panaroo>)⁵² in ‘strict’ mode with initial clustering at 98% length and sequence identity followed by a family threshold of 70%. Pangenome synteny was then mapped using Corekarrav0.0.5 (ref. 52,53). Segments of accessory genes (those which were present in <95% of all genomes) were then examined for integrase/recombinase subfamilies using HMMer v3.4 and a scheme designed by ref. 54 with an additional ICE-associated DDE recombinase (PF06782)⁵⁵. Accessory segments were classified into MGE categories based on recombinase subfamily and the presence of prophage and/or ICE structural genes as inferred by eggNOG-mapper v2.1.7 (ref. 56) or hidden Markov models (HMMs) from TXSScan (accessed August 2025)⁵⁷ with ‘hmmsearch’ using an *E* value threshold of 0.001. Segments of accessory genes with multiple prophage and ICE-associated integrases/structural genes (for example, nested MGEs) or with insufficient contextual information (for example, due to assembly fragmentation) were binned as mobility elements (MEs). Segments with a prophage-associated integrase but without at least two annotated phage structural genes were classified as phage-like^{54,58}.

Sequences of accessory segments categorized as ‘Phage’, ‘ICE’, ‘ME’ and ‘Phage-like’ were extracted for downstream analyses and grouped according to their insertion sites, defined by adjacent core gene pairs. To manage the effects of genome re-arrangements, only segments extracted from insertion sites with the most common arrangement between the core genes were retained⁵⁹. Specifically, uncommon insertion sites with an occurrence <10 across genomes and containing a conflicting core gene within other arrangements were excluded. Both full-length segments without contig breaks and segments with contig breaks but located adjacent to one of these core genes were extracted. In total, 2,165 full-length MGEs and 4,485 fragmented MGEs were extracted from the combined *emm1* and *emm12* pangenome.

Full-length segments (without contig breaks) extracted from the common insertion sites were first clustered using CD-HIT v4.8.1 (ref. 60), with thresholds of 95% sequence identity and 95% length identity, yielding 126 full-length representative MGEs. The representative sequences generated by CD-HIT were further clustered by ‘sourmash compare’ using Sourmash v4.9.3 (ref. 61), applying MinHash-based sketching with a scaled parameter of 30. After aligning the representatives to visualize differences, a Jaccard similarity threshold of $\geq 90\%$ was determined to define MGE clusters. Thus, full-length MGE segments with $\geq 90\%$ Jaccard similarity were grouped into the same cluster, resulting in 81 unique MGEs (Supplementary Table 1). Of the 81 unique MGEs, 16 MGE carrying known virulence factors or antimicrobial resistance genes are present in ≥ 10 isolates. To classify MGE segments containing contig breaks, whole genome assemblies were also converted into sketches by ‘sourmash sketch’ and mapped to the new representatives by ‘sourmash gather’. Indexed RocksDB databases were built from representatives at each insertion site to identify the best-matching MGE per assembly and insertion site. If $>90\%$ of the *k*-mer content from an MGE sketch matched with a genome assembly sketch, the corresponding genome was considered to contain that MGE. This procedure clustered 4,349 of the 4,485 fragmented MGE segments into the 81 MGE representatives.

Antimicrobial resistance genes and virulence factors in both assemblies and the MGE segment were identified by ABRicate v1.0.1 (<https://github.com/tseemann/abricate>) with the NCBI AMRFinder Plus and virulence factor databases, using cut-offs of 70% nucleotide identity and 70% coverage^{62,63}. For each MGE, the presence or absence in a genome was confirmed only when the ABRicate results for the MGE classification were consistent with those from the corresponding assembly. Representative clustering of common ICEs as well as prophage carrying *speC/spd1*, and *ssa* phage variants across the *emm12* and *emm1* populations are represented in Extended Data Figs. 3–5. Corresponding pairwise alignments were visualized by Genofig v1.1 (ref. 64), with nucleotide sequence identity defined by BlastN v2.16.0.

PCR

All GAS strains were routinely grown in Todd–Hewitt broth (Thermo Fisher Scientific) supplemented with 1% (*w/v*) yeast extract (Merck; THY) at 37 °C to late-exponential growth phase at an optical density at 600 nm (OD_{600}) of 0.8. Virulence genes *speA*, *speC*, *ssa*, *spd1*, *speB* and *slo* were PCR-amplified from purified GAS genomic DNA (QIAamp DNA Mini Kit, QIAGEN) using MangoTaq DNA polymerase (Meridian Bioscience) according to the manufacturer’s instructions. The following primers were used: *speA* forward primer (5′-GTGACATTTCTTGGACTACAATCTCG-3′) and reverse primer (5′-TATCATAAGATATTTAGATTGAGTAAATTTCTGGTTTCAG-3′), *speC* forward primer (5′-GAATGTTAAAGTGATTTACTTTATGCATACACTATAACTC-3′) and reverse primer (5′-CTCATAAGACATTTCCGGAATAATATAATAGTC-3′), *ssa* forward primer (5′-CAGAACAATTAACAATCTAGCCAATTTACTG-3′) and reverse primer (5′-GAAAAATCAATCATGCTGTAAAAGCTGAC-3′), *spd1* forward primer (5′-CGTGGC ATCTAGTCGGATA-3′) and reverse primer (5′-GGTGAAGTGCAAGCCAA GAA-3′), and *slo* forward primer (5′-GCTGGCTAATAAAGTTTACCG-3′)

and reverse primer (5′-CGGTAACCTTTATTAGCCAGC-3′)¹¹. PCR products were analysed on 1% agarose gels and visualized using a GelDoc XR+ imaging system (Bio-Rad).

Western blotting

Total protein from bacterial culture supernatants isolated from GAS cultures grown in THY to an OD_{600} of 0.8 were centrifuged, and the culture supernatants were filtered through a 0.22 μm membrane (Millex-GV) before the addition of trichloroacetic acid (final concentration, 10%; Sigma-Aldrich). Proteins were precipitated overnight at 4 °C, pelleted by centrifugation and washed with ice-cold ethanol. Pellets were air-dried and resuspended in LDS loading buffer (Thermo Fisher) containing 100 mM dithiothreitol (Sigma-Aldrich), then boiled for 15 min with intermittent vortexing. Samples were separated by SDS–PAGE, and proteins were transferred onto a methanol-activated PVDF membrane (Merck) using a wet transfer system (Bio-Rad). Membranes were blocked in Intercept Blocking Buffer (LI-COR) for 1 h at room temperature, followed by overnight incubation at 4 °C with primary antibodies against SpeA (PAI111, Toxin Technology; 1:1,000 dilution), SpeC (PCI333, Toxin Technology; 1:1,000 dilution) and SSA (Mimotopes; 1:500 dilution)¹¹. Spd1 and Slo were detected using murine primary antibodies at 1:1,000 and 1:2,000 dilutions, respectively⁶⁵. Fluorescent secondary antibodies (DyLight 800 anti-mouse or anti-rabbit IgG; NEB; 1:10,000) were applied for 1 h at room temperature, and membranes imaged using an Odyssey Imaging System (LI-COR).

Statistics and reproducibility

Population incidence of scarlet fever in China was calculated per 100,000 population per year using NNIDSS data from 1990 to 2024; 95% CI of the proportion was calculated using the method embedded in an online calculator (<http://vassarstats.net/prop1.html>), according to a method described by Newcombe⁶⁶, derived from a procedure outlined by Wilson⁶⁷, using the Wilson procedure with a correction for continuity.

For phylogenetic contextualization, global sequences from the UK, Australia and Denmark were randomized based on global temporal and geographic distribution. To minimize over-representation of these datasets, isolates were randomly subsampled by R function sample based on temporal and geographic distribution ($n = 10$ isolates per year per study).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive³⁴ in the National Genomics Data Center³⁵, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA CRA033131), which are publicly accessible at <https://ngdc.cncb.ac.cn/gsa>. Accession numbers for individual isolates and publicly available Illumina reads are listed in Supplementary Table 1. Long-read ONT data are available at the NCBI under BioProject PRJEB103775. Additional metadata are available upon request from Y. You as some of these include patient-level information; a response should be received within 14 days. A formal data transfer agreement would need to be concluded with the China CDC within parameters as set out in ethics Approval No. ICDC-2023003. Source data are provided with this paper.

References

1. Brouwer, S. et al. Pathogenesis, epidemiology and control of group A *Streptococcus* infection. *Nat. Rev. Microbiol.* **21**, 431–447 (2023).

2. Quinn, R. W. Comprehensive review of morbidity and mortality trends for rheumatic fever, streptococcal disease, and scarlet fever: the decline of rheumatic fever. *Rev. Infect. Dis.* **11**, 928–953 (1989).
3. Ralph, A. P. & Carapetis, J. R. Group A streptococcal diseases and their global burden. *Curr. Top. Microbiol. Immunol.* **368**, 1–27 (2013).
4. Cole, J. N., Barnett, T. C., Nizet, V. & Walker, M. J. Molecular insight into invasive group A streptococcal disease. *Nat. Rev. Microbiol.* **9**, 724–736 (2011).
5. Aziz, R. K. & Kotb, M. Rise and persistence of global MIT1 clone of *Streptococcus pyogenes*. *Emerg. Infect. Dis.* **14**, 1511–1517 (2008).
6. Vieira, A. et al. Rapid expansion and international spread of M1_{UK} in the post-pandemic UK upsurge of *Streptococcus pyogenes*. *Nat. Commun.* **15**, 3916 (2024).
7. Davies, M. R. et al. Emergence of scarlet fever *Streptococcus pyogenes* emm12 clones in Hong Kong is associated with toxin acquisition and multidrug resistance. *Nat. Genet.* **47**, 84–87 (2015).
8. You, Y. et al. Scarlet fever epidemic in China caused by *Streptococcus pyogenes* serotype M12: epidemiologic and molecular analysis. *EBioMedicine* **28**, 128–135 (2018).
9. Lamagni, T. et al. Resurgence of scarlet fever in England, 2014–16: a population-based surveillance study. *Lancet Infect. Dis.* **18**, 180–187 (2018).
10. Cai, J. et al. Ongoing epidemic of scarlet fever in Shanghai and the emergence of M1_{UK} lineage group A *Streptococcus*: a 14-year surveillance study across the COVID-19 pandemic period. *Lancet Reg. Health West. Pac.* **58**, 101576 (2025).
11. Davies, M. R. et al. Detection of *Streptococcus pyogenes* M1_{UK} in Australia and characterization of the mutation driving enhanced expression of superantigen SpeA. *Nat. Commun.* **14**, 1051 (2023).
12. Demczuk, W., Martin, I., Domingo, F. R., MacDonald, D. & Mulvey, M. R. Identification of *Streptococcus pyogenes* M1_{UK} clone in Canada. *Lancet Infect. Dis.* **19**, 1284–1285 (2019).
13. Rumke, L. W. et al. Dominance of M1_{UK} clade among Dutch M1 *Streptococcus pyogenes*. *Lancet Infect. Dis.* **20**, 539–540 (2020).
14. Li, Y., Nanduri, S. A., Van Beneden, C. A. & Beall, B. W. M1_{UK} lineage in invasive group A *Streptococcus* isolates from the USA. *Lancet Infect. Dis.* **20**, 538–539 (2020).
15. Jain, N., Lansiaux, E. & Reinis, A. Group A streptococcal (GAS) infections amongst children in Europe: taming the rising tide. *New Microbes New Infect.* **51**, 101071 (2023).
16. Ho, E. C. et al. Outbreak of invasive group A *Streptococcus* in children-Colorado, October 2022-April 2023. *J. Pediatric Infect. Dis. Soc.* **12**, 540–548 (2023).
17. Xie, O. et al. Temporal and geographical lineage dynamics of invasive *Streptococcus pyogenes* in Australia from 2011 to 2023: a retrospective, multicentre, clinical and genomic epidemiology study. *Lancet Microbe* **6**, 101053 (2025).
18. Gouveia, C. et al. Sustained increase of paediatric invasive *Streptococcus pyogenes* infections dominated by M1_{UK} and diverse emm12 isolates, Portugal, September 2022 to May 2023. *Euro Surveill.* **28**, 2300427 (2023).
19. Cunningham, C. et al. Incidence and treatment of group A streptococcal infections during COVID-19 pandemic and 2022 outbreak: retrospective cohort study in England using OpenSAFELY-TPP. *BMJ Med.* **3**, e000791 (2024).
20. Guy, R. et al. Increase in invasive group A streptococcal infection notifications, England, 2022. *Euro Surveill.* **28**, 2200942 (2023).
21. You, Y., Xiaojuan, Z., Jie, L. & Bike, Z. Analysis of public health risks and countermeasures for group A streptococcal diseases. *Dis. Surveill.* **39**, 1–7 (2024).
22. Wang, Z. & Bao, S. The impact of social distancing measures (quarantine) policy on tertiary education and medical consultations in China during the COVID-19 pandemic. *Front. Public Health* **12**, 1365805 (2024).
23. Xiang, C., Zhang, J., Zhao, F. & You, Y. *Emm* type distribution of group A *Streptococcus* in China during 1990 and 2020: a systematic review and implications for vaccine coverage. *Front. Public Health* **11**, 1157289 (2023).
24. Ben Zakour, N. L. et al. Transfer of scarlet fever-associated elements into the group A *Streptococcus* MIT1 clone. *Sci. Rep.* **5**, 15877 (2015).
25. You, Y. et al. Complete genome sequence of a *Streptococcus pyogenes* serotype M12 scarlet fever outbreak isolate from China, compiled using Oxford Nanopore and Illumina Sequencing. *Genome Announc.* **6**, e00389–18 (2018).
26. Yu, D., Zheng, Y., Chen, Y., You, Y. & Yang, Y. *Streptococcus pyogenes* M1_{UK} presence in China in 2018. *J. Glob. Antimicrob. Resist.* **42**, 175–176 (2025).
27. Kansal, R. G., McGeer, A., Low, D. E., Norrby-Teglund, A. & Kotb, M. Inverse relation between disease severity and expression of the streptococcal cysteine protease, SpeB, among clonal MIT1 isolates recovered from invasive group A streptococcal infection cases. *Infect. Immun.* **68**, 6362–6369 (2000).
28. Ferretti, J. J. et al. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl Acad. Sci. USA* **98**, 4658–4663 (2001).
29. Smeesters, P. R. et al. Global *Streptococcus pyogenes* strain diversity, disease associations, and implications for vaccine development: a systematic review. *Lancet Microbe* **5**, e181–e193 (2024).
30. You, Y. Research progress on scarlet fever epidemic and associated factors. *Chin. J. Appl. Clin. Pediatr.* **37**, 1626–1629 (2022).
31. Yu, D. et al. Molecular characteristics of *Streptococcus pyogenes* isolated from Chinese children with different diseases. *Front. Microbiol.* **12**, 722225 (2021).
32. Lynskey, N. N. et al. Emergence of dominant toxigenic M1T1 *Streptococcus pyogenes* clone during increased scarlet fever activity in England: a population-based molecular epidemiological study. *Lancet Infect. Dis.* **19**, 1209–1218 (2019).
33. Beall, B., Facklam, R. & Thompson, T. Sequencing emm-specific PCR products for routine and accurate typing of group A streptococci. *J. Clin. Microbiol.* **34**, 953–958 (1996).
34. Zhang, S. et al. The GSA family in 2025: a broadened sharing platform for multi-omics and multimodal data. *Genom. Proteom. Bioinform.* **23**, qzaf072 (2025).
35. CNCB-NGDC Members and Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformatics in 2025. *Nucleic Acids Res.* **53**, D30–D44 (2025).
36. Xie, O. et al. Inter-species gene flow drives ongoing evolution of *Streptococcus pyogenes* and *Streptococcus dysgalactiae* subsp. *equisimilis*. *Nat. Commun.* **15**, 2286 (2024).
37. Xie, O. et al. Overlapping *Streptococcus pyogenes* and *Streptococcus dysgalactiae* subspecies *equisimilis* household transmission and mobile genetic element exchange. *Nat. Commun.* **15**, 3477 (2024).
38. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
39. Pribelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinformatics* **70**, e102 (2020).
40. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
41. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

42. Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* **3**, 124 (2018).
43. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. In *Some Mathematical Questions in Biology: DNA Sequence Analysis, Lectures on Mathematics in the Life Sciences* **17**, 57–86 (American Mathematical Society, 1986).
44. Croucher, N. J. et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
45. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
46. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134 (2018).
47. Didelot, X., Siveroni, I. & Volz, E. M. Additive uncorrelated relaxed clock models for the dating of genomic epidemiology phylogenies. *Mol. Biol. Evol.* **38**, 307–317 (2021).
48. Minin, V. N., Bloomquist, E. W. & Suchard, M. A. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25**, 1459–1471 (2008).
49. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).
50. Gao, J., May, M. R., Rannala, B. & Moore, B. R. Model misspecification misleads inference of the spatial dynamics of disease outbreaks. *Proc. Natl Acad. Sci. USA* **120**, e2213913120 (2023).
51. Barone, R. & Tancredi, A. Bayesian inference for discretely observed continuous time multi-state models. *Stat. Med.* **41**, 3789–3803 (2022).
52. Tonkin-Hill, G. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
53. Jespersen, M. G., Hayes, A. J. & Davies, M. R. Corekiburra: pan-genome post-processing using core gene synteny. *J. Open Source Softw.* **7**, 4910 (2022).
54. Khedkar, S. et al. Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Res.* **50**, 3155–3168 (2022).
55. Ambroset, C. et al. New insights into the classification and integration specificity of *Streptococcus* integrative conjugative elements through extensive genome exploration. *Front. Microbiol.* **6**, 1483 (2015).
56. Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
57. Abby, S. S. et al. Identification of protein secretion systems in bacterial genomes. *Sci. Rep.* **6**, 23080 (2016).
58. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
59. Jespersen, M. G., Hayes, A. J., Tong, S. Y. C. & Davies, M. R. Insertion sequence elements and unique symmetrical genomic regions mediate chromosomal inversions in *Streptococcus pyogenes*. *Nucleic Acids Res.* **52**, 13128–13137 (2024).
60. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
61. Irber, L. et al. sourmash v4: A multitool to quickly search, compare, and analyze genomic and metagenomic data sets. *J. Open Source Softw.* **9**, 6830 (2024).
62. Liu, B., Zheng, D., Zhou, S., Chen, L. & Yang, J. VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.* **50**, D912–D917 (2022).
63. Feldgarden, M. et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.* **11**, 12728 (2021).
64. Branger, M. & Leclercq, S. O. GenoFig: a user-friendly application for the visualization and comparison of genomic regions. *Bioinformatics* **40**, btac372 (2024).
65. Brouwer, S. et al. Prophage exotoxins enhance colonization fitness in epidemic scarlet fever-causing *Streptococcus pyogenes*. *Nat. Commun.* **11**, 5018 (2020).
66. Newcombe, R. G. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat. Med.* **17**, 857–872 (1998).
67. Wilson, E. B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **22**, 209–212 (1927).

Acknowledgements

We thank China CDC for providing scarlet fever surveillance data and the Centre for Pathogen Genomics – Innovation Hub (Melbourne) for providing genome sequencing support. This work was supported by grants from the Beijing Natural Science Foundation (7242190) (to Y. You and J.Z.), National Natural Science Foundation of China (82304198), Shenzhen Clinical Research Center (20220819113341005) (to D.Y.), Hainan Natural Science Foundation Youth Fund Project (821QN421) (to X. Yu), and both grants (to S.B., M.J.W. and M.R.D.) and postgraduate scholarship (O.X.) from the National Health and Medical Research Council of Australia (GNT2019767, GNT2040856, GNT2043549, GNT2013831).

Author contributions

Y. You conceptualized this study. Y. You, Y. Yang, O.X., M.R.D. and M.J.W. generated hypotheses. Y. You, D.Y., X.P., H.C., C.H., F.Z., X. Yan, M. Zhang, M.F., X. Yu, Lu Sun, X.W., L.H., J.L., D.Z., J.W., C.S., Y.Z., M. Zhou, Lifang Sun, Q.W., J.Z. and Y. Yang contributed isolates for this study. Y. You, J.E.J.W. and S.B. generated data for this study. Y. You, C.Y., O.X., X.R., C.D.G., S.D. and MRD undertook data analysis. Y. You, C.Y., X.R., O.X., M.R.D. and M.J.W. wrote the paper. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interest.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41564-026-02324-4>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41564-026-02324-4>.

Correspondence and requests for materials should be addressed to Yuanhai You, Mark R. Davies, Mark J. Walker, Quanyi Wang, Jianzhong Zhang or Yonghong Yang.

Peer review information *Nature Microbiology* thanks Cheryl Andam, Dennis Nurjadi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

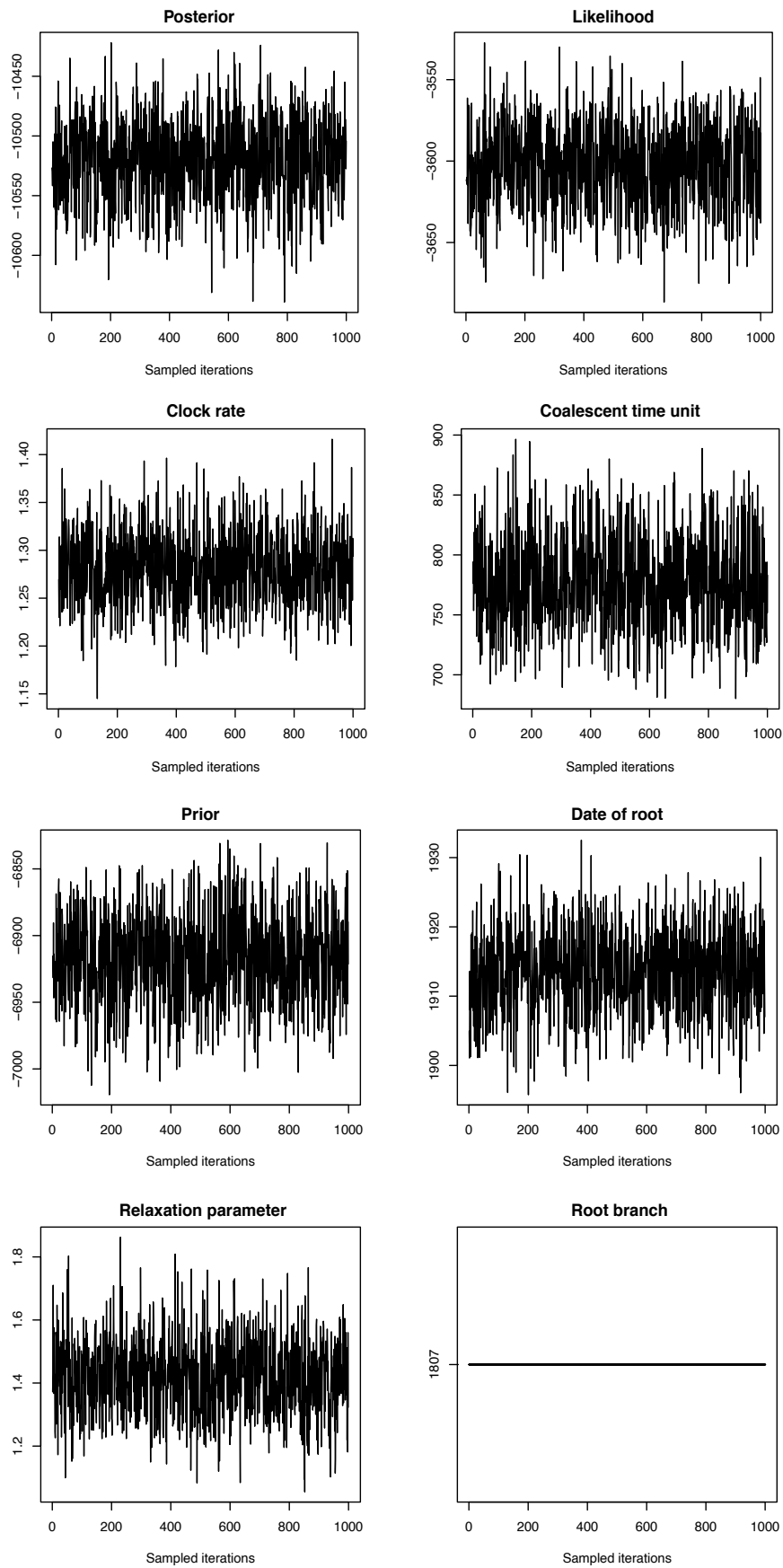
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving

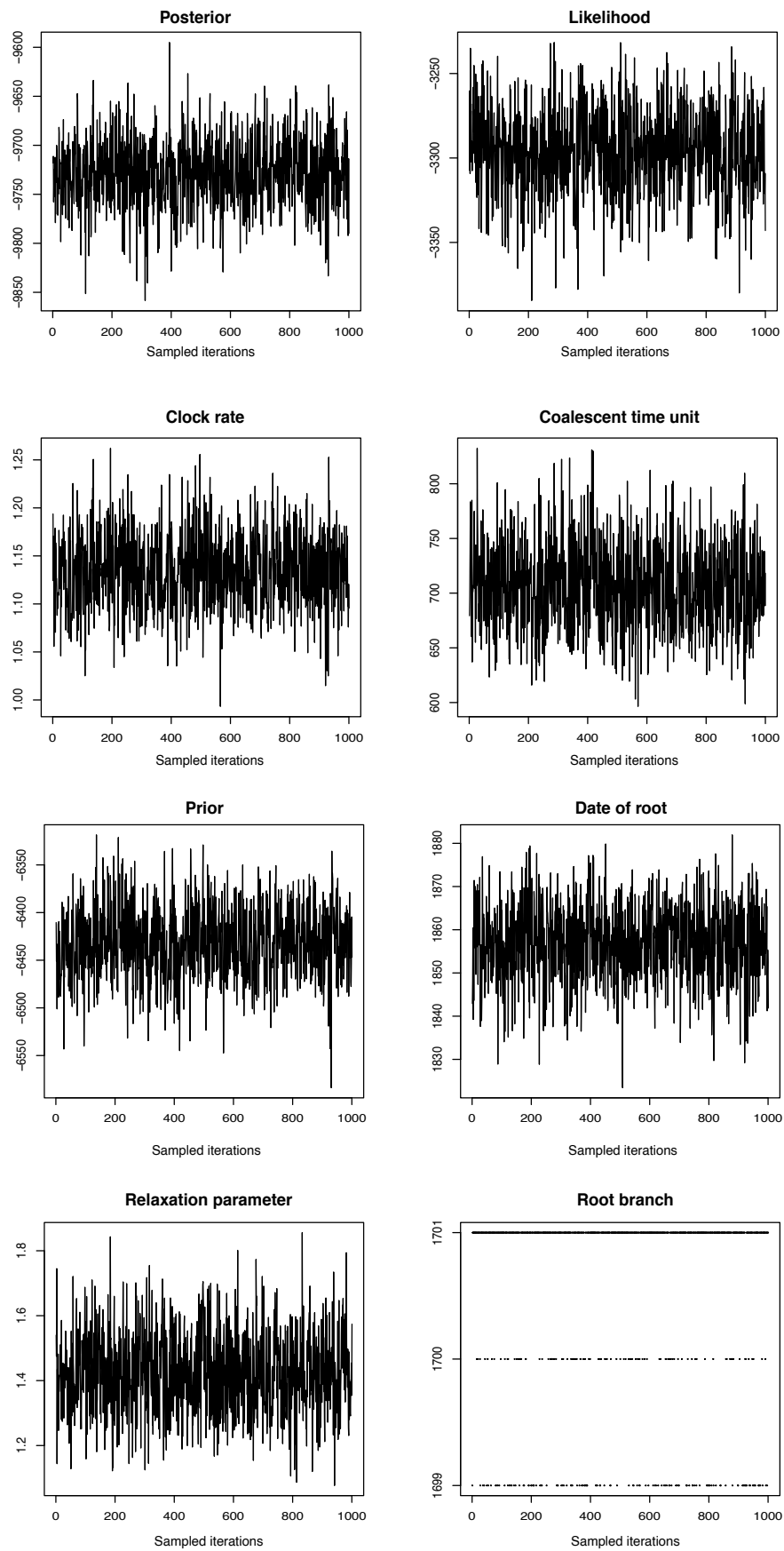
of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2026

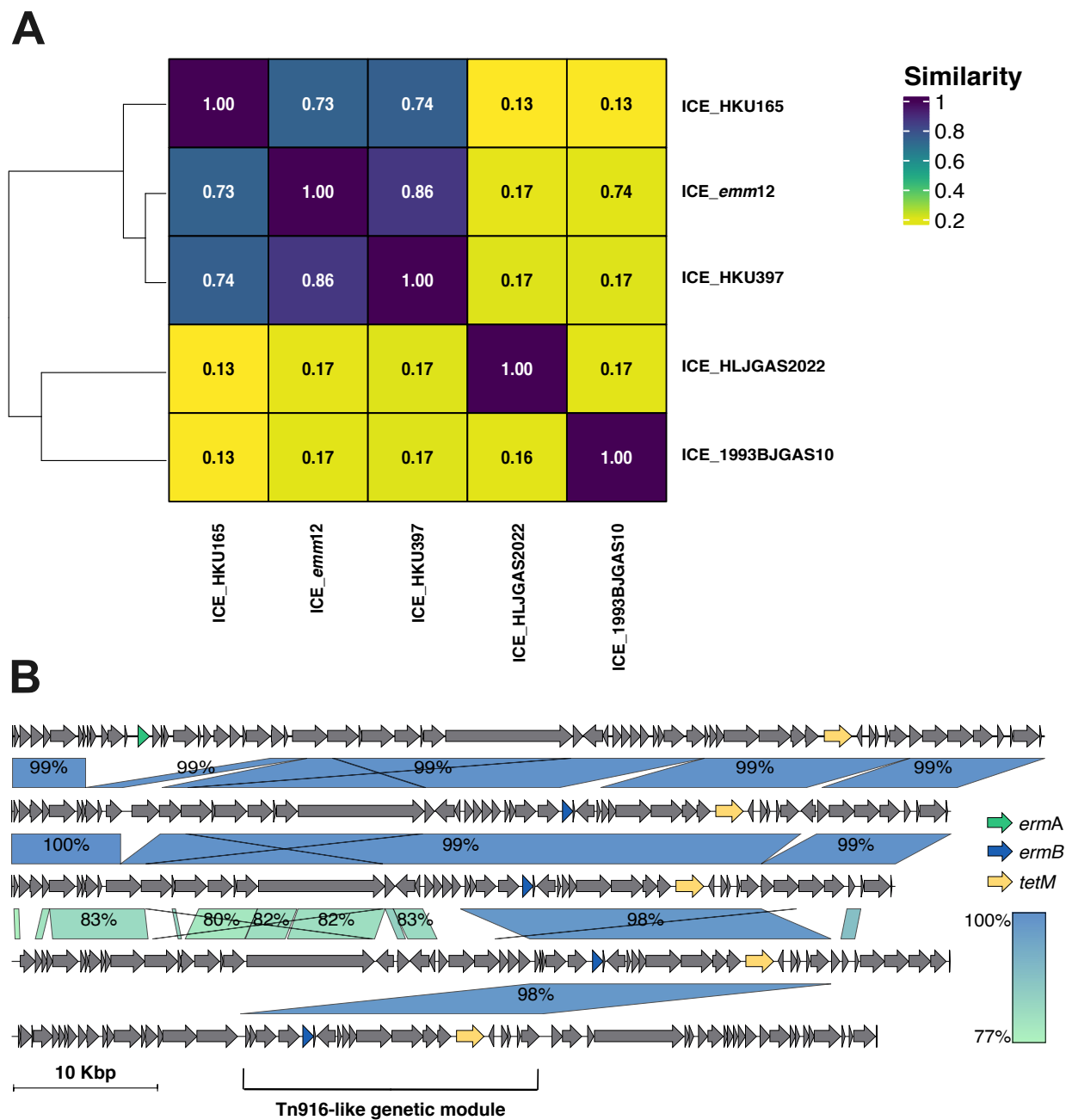
¹National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China. ²Department of Respiration, Shenzhen Children's Hospital, Shenzhen University, Shantou University Medical College, Shenzhen, China. ³Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China. ⁴Beijing Key Laboratory of Surveillance, Early Warning and Pathogen Research on Emerging Infectious Diseases, Beijing Center for Disease Prevention and Control, Beijing, China. ⁵The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia. ⁶Monash Infectious Diseases, Monash Health, Melbourne, Victoria, Australia. ⁷Department of Pediatrics, Beijing Chaoyang Hospital, Capital Medical University, Beijing, China. ⁸Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China. ⁹Suzhou Key Laboratory of Pathogenic Microorganisms for Emerging and Re-emerging Infectious Diseases, Suzhou Center for Disease Control and Prevention, Suzhou, China. ¹⁰Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia. ¹¹Department of Computational Biology, Institut Pasteur, Paris, France. ¹²Shandong Provincial Key Laboratory of Intelligent Monitoring, Early Warning Prevention and Control for Infectious Diseases, Jinan, China. ¹³Hainan Provincial Center for Disease Control and Prevention, Haikou, China. ¹⁴Department of Clinical Laboratory Medicine, The First Affiliated Hospital of Shandong First Medical University & Shandong Provincial Qianfoshan Hospital, Shandong Medicine and Health Key Laboratory of Laboratory Medicine, Jinan, China. ¹⁵Department of Pediatrics, The First Affiliated Hospital of Anhui Medical University, Hefei, China. ¹⁶Microbiology Laboratory, Beijing Pediatric Research Institute, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China. ¹⁷These authors contributed equally: Yuanhai You, Dingle Yu, Chao Yang, Xiaomin Peng, Ouli Xie, Hesheng Chang, Chunzhen Hua.



Extended Data Fig. 1 | Trace files of key Markov Chain Monte Carlo (MCMC) parameters from the *emm12* BactDating analysis. Traces derived from 1e7 iterations with the first half discarded as MCMC burnin. Overall mixing is supportive of MCMC convergence.



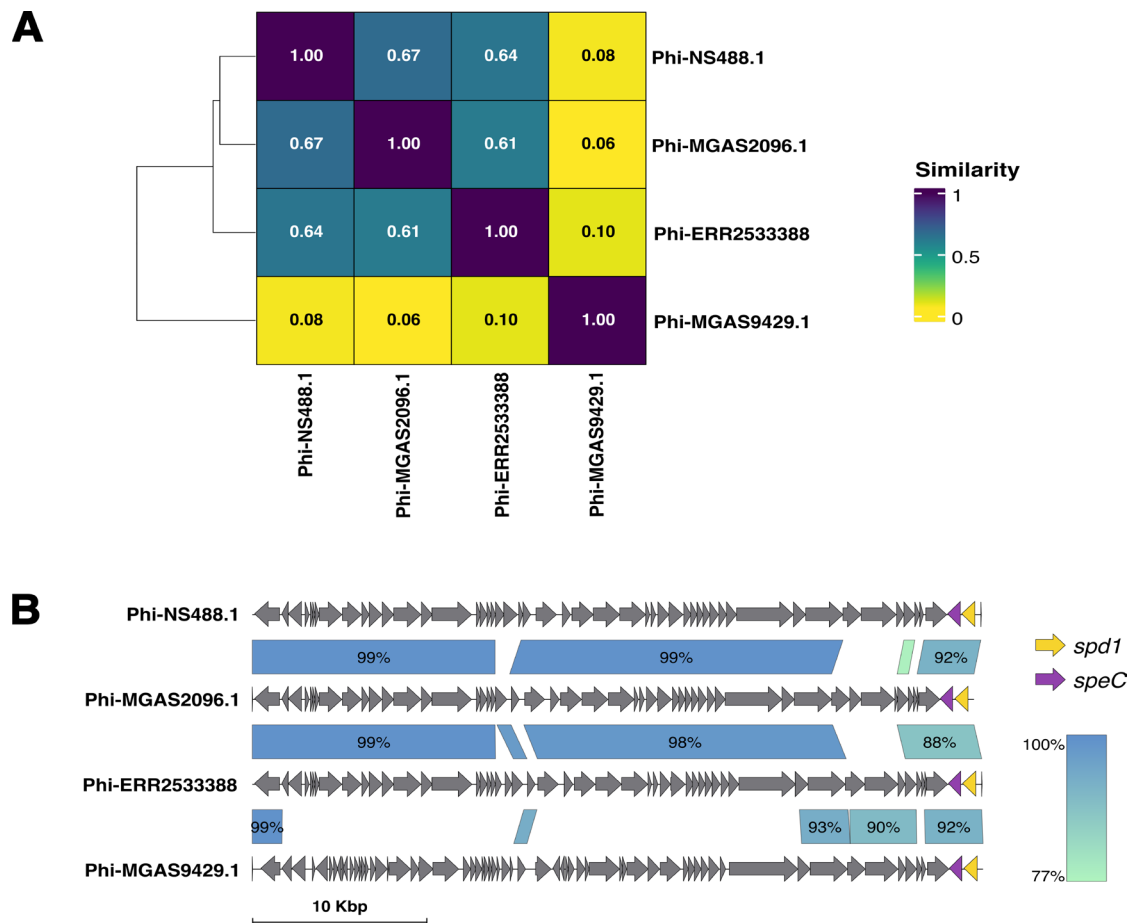
Extended Data Fig. 2 | Trace files of key Markov Chain Monte Carlo (MCMC) parameters from the *emm1* BactDating analysis. Traces derived from 1e7 iterations with the first half discarded as MCMC burnin. Overall mixing is supportive of MCMC convergence.



Extended Data Fig. 3 | Comparison of common multidrug resistant integrative conjugative elements in the GAS *emm1* and *emm12* population.

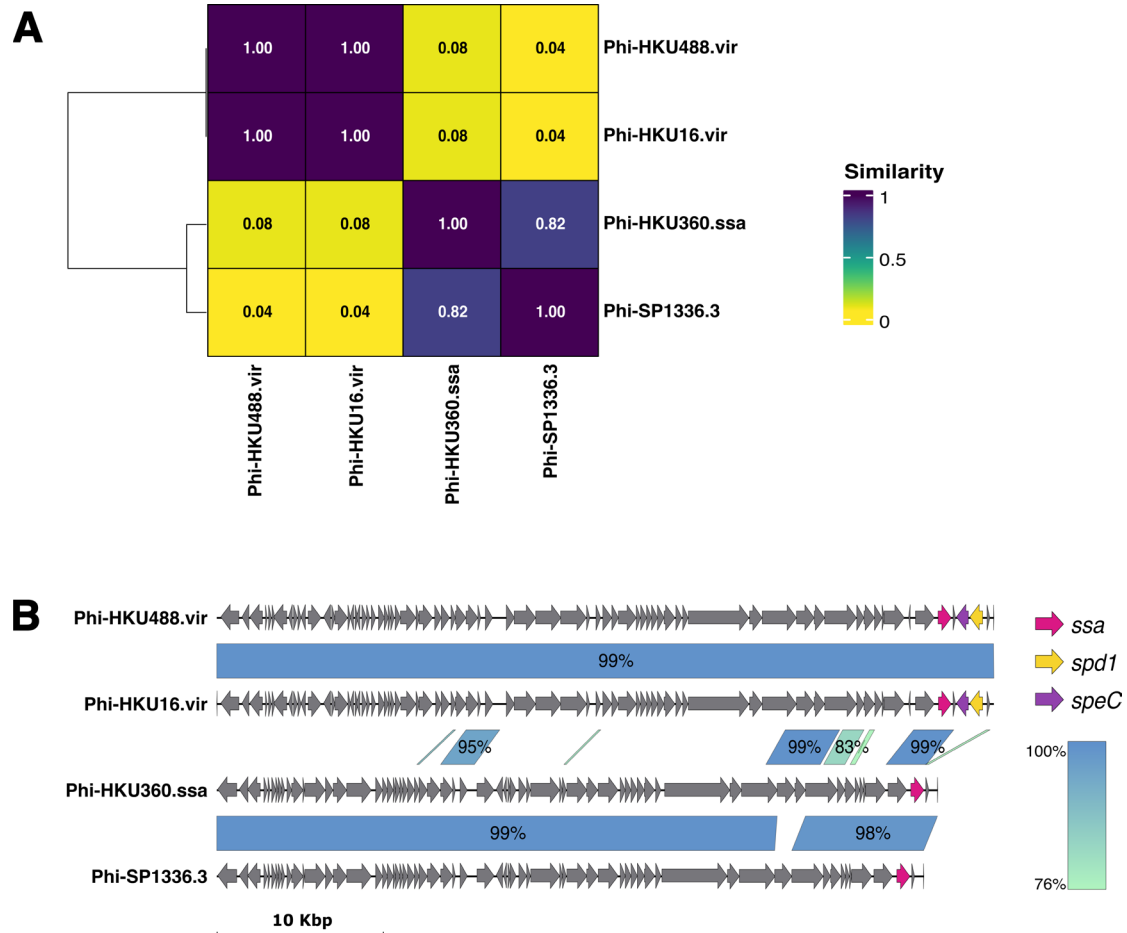
A, Heatmap shows the Jaccard similarity between 5 ICE variants, calculated based on k-mers generated by MinHash sketching. **B**, Pairwise alignment of the 5 ICE variants with blocks indicating regions of shared nucleotide sequence identity as

defined by BlastN. Macrolide (*ermA* and *ermB*) and tetracycline (*tetM*) resistance genes are highlighted in yellow, blue, and red respectively. The relative position of the modular ~20 kb Tn916-like transposable element common to 4 of the 5 ICE variants is displayed at the bottom of the figure as defined by Davies et al.⁷



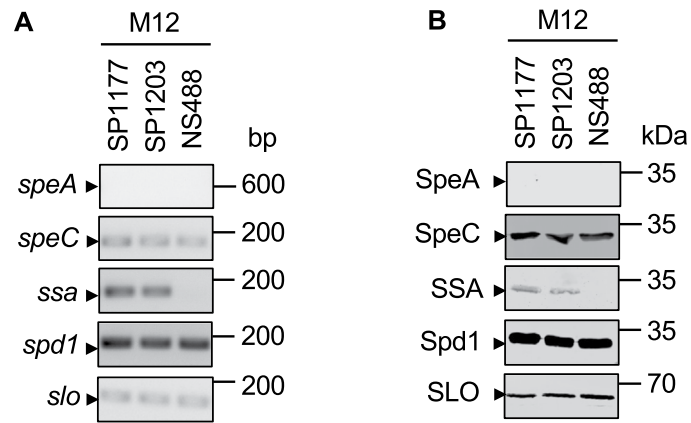
Extended Data Fig. 4 | Comparison of common *speC* and *spd1* virulence carrying prophage in the GAS *emm1* and *emm12* population. A, Heatmap shows the Jaccard similarity between 4 phage variants (present in ≥ 10 isolates from either *emm1* or *emm12* population), calculated based on k-mers generated

by MinHash sketching. **B,** Pairwise alignment of the four phage variants with blocks indicating regions of shared nucleotide sequence identity as defined by BlastN. The deoxyribonuclease *spd1* and the streptococcal pyrogenic exotoxin *speC* are highlighted in yellow and purple respectively.



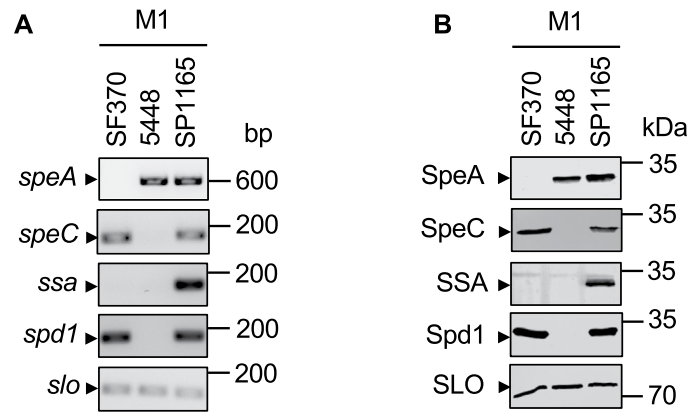
Extended Data Fig. 5 | Comparison of common *ssa* virulence carrying prophage in the GAS *emm1* and *emm12* population. A, Heatmap shows the Jaccard similarity between 4 phage variants (present in ≥ 10 isolates from either *emm1* or *emm12* population), calculated based on k-mers generated by MinHash

sketching. **B,** Pairwise alignment of the 4 phage variants with blocks indicating regions of shared nucleotide sequence identity as defined by BlastN. The streptococcal superantigen *ssa* and *speC*, and the deoxyribonuclease *spd1* are highlighted in pink, purple and yellow respectively.



Extended Data Fig. 6 | Toxin repertoire and virulence of representative *emm12* strains, *emm12*-Clade I (SP1177), *emm12*-Clade II (SP1203) and *emm12*-Clade III (NS488). (a) Carriage of scarlet fever-associated toxins in *emm12* representative strains was confirmed by polymerase chain reaction

(PCR). (b) Western blot analysis of SpeA, SpeC, SSA, Spd1, and SLO in culture supernatants of indicated *emm12* strains grown to late logarithmic phase. Experiments were undertaken n = 2.



Extended Data Fig. 7 | Toxin repertoire and virulence of representative *emm1* strains, *emm1* ancestral (SF370), M1global (5448) and M1China (SP1165). (a) Carriage of scarlet fever-associated toxins in *emm1* representative strains was

confirmed by polymerase chain reaction (PCR). (b) Western blot analysis of SpeA, SpeC, SSA, Spd1, and SLO in culture supernatants of indicated *emm1* strains grown to late logarithmic phase. Experiments were undertaken $n = 2$.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a | Confirmed |
|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis https://www2a.cdc.gov/ncidod/biotech/strepblast.asp)
Dorado v0.4.2 (dna_r10.4.1_e8.2_400bps_sup@v5.0.0)
Kraken2 v2.1.2
Shovill v1.1.0 (<https://github.com/tseemann/shovill>)
SPAdes assembler v3.14.0
Filtlong v0.2.1 (<https://github.com/rrwick/Filtlong>)
Flye v2.9.5
Prokav1.14.6
emmtyper v0.2.0 (<https://github.com/MDU-PHL/emmtyper>)
MLST v2.23 (<https://github.com/tseemann/mlst>)
Snippy v4.6.0 (<https://github.com/tseemann/snippy>)
Gubbins v3.4
IQ-tree v3.0.1
BactDating v1.1.4
BEAST v1.10.5
Panaroo v1.5.2 (<https://github.com/gtonkinhill/panaroo>)
Corekaurav0.0.5
HMMer v3.4

eggNOG-mapper v2.1.7
 Sourmash v4.9.3
 ABRicate v1.0.1 (<https://github.com/tseemann/abricate>)
 Genofig v1.1
 BlastN (National Center for Biotechnology Information (NCBI))
<http://vassarstats.net/prop1.html>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2025) in the National Genomics Data Center, (Nucleic Acids Res 2025), China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA033131) that are publicly accessible at <https://ngdc.cnbc.ac.cn/gsa>. Accession numbers for individual isolates and publicly available Illumina reads are listed in Supplementary Table 1. Long-read ONT data are available at the NCBI under the BioProject PRJEB103775. Further additional data are available from the corresponding author Dr. Yuanhai You upon request.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	Not applicable
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable
Population characteristics	Cases of suspected scarlet fever are notified by clinicians to the Chinese National Notifiable Infectious Disease Surveillance System on the basis of symptoms consistent with scarlet fever, with or without laboratory confirmation of GAS infection. GAS isolates were collected from eight provinces across China from 1993-2024.
Recruitment	Not applicable
Ethics oversight	The National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control approved this study under ethics number ICDC-2023003.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No formal sample size calculations were performed. However, where available, samples for contextualization were chosen to be as broadly representative from regions around the globe as possible.
Data exclusions	Illumina genome sequencing reads with >5% reads assigned to another species were excluded due to suspected contamination. To manage the effects of genome re-arrangements, only segments extracted from insertion sites with the most common arrangement between the core genes were retained. Specifically, uncommon insertion sites with an occurrence <10 across genomes and containing a conflicting core gene within other arrangements were excluded.
Replication	Extended data figure 6 & 7 have been repeated.
Randomization	All available strains from China were included in the analysis. For phylogenetic contextualization, global sequences from UK, Australia and

Randomization	Denmark were randomized based on global temporal and geographic distribution. To minimize over-representation of these datasets, isolates were randomly subsampled by R function sample based on temporal and geographic distribution (n = 10 isolates per year/study).
Blinding	No blinding was performed in this study. Blinding was not required as the results are quantitative and did not require subjective judgment or interpretation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Included in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	Affinity-purified rabbit antibody to SpeA (PAI111, Toxin Technology; 1:1000 dilution), affinity-purified rabbit antibody to SpeC (PCI333, Toxin Technology; 1:1000 dilution), affinity-purified rabbit antibody to SSA (produced by Mimotopes, Clayton, Australia raised against the peptide HCGGSSQPDPTPEQLNKSSQFTG-OH coupled to Keyhole Limpet Hemocyanin; 1:500 dilution). Mouse antibody to Spd1 (1:1000 dilution) were generated as previously described (PMID: 33024089). Anti-rabbit IgG (H+L) (DyLight 800 4X PEG Conjugate, NEB, 5151P) and anti-mouse IgG (H+L) (DyLight 800 4x PEG Conjugate, NEB, 52575) were used as the secondary antibodies (1:10,000 dilution).
Validation	Specificity of each primary antibody was validated by Western Blotting using purified recombinant protein and culture supernatants of respective isogenic mutant strains of <i>Streptococcus pyogenes</i> . Detection with secondary antibodies validated the species-specific source of each primary antibody.

Plants

Seed stocks	<i>Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.</i>
Novel plant genotypes	<i>Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.</i>
Authentication	<i>Describe any authentication procedures for each seed-stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.</i>