nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0417-8

Corrected: Author Correction

Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics

Mark R. Davies[1,2,3,4*], Liam McIntyre[1], Ankur Mutreja[2,5], Jake A. Lacey[6], John A. Lees[7], Rebecca J. Towers[8], Sebastián Duchêne[1,9], Pierre R. Smeesters[10,11,12], Hannah R. Frost[10,11,12], David J. Price[13,14], Matthew T. G. Holden[2,15], Sophia David[2], Philip M. Giffard[8], Kate A. Worthing[1], Anna C. Seale[16], James A. Berkley[17], Simon R. Harris[2], Tania Rivera-Hernandez[3,4], Olga Berking[3,4], Amanda J. Cork[3,4], Rosângela S. L. A. Torres[18,19], Trevor Lithgow[20], Richard A. Strugnell[1], Rene Bergmann[21], Patric Nitsche-Schmitz[21], Gusharan S. Chhatwal[21], Stephen D. Bentley[2], John D. Fraser[22], Nicole J. Moreland[22], Jonathan R. Carapetis[23], Andrew C. Steer[12], Julian Parkhill[2], Allan Saul[5], Deborah A. Williamson[24], Bart J. Currie[8], Steven Y. C. Tong[6,8,25], Gordon Dougan[2,26] and Mark J. Walker[3,4*]

Group A *Streptococcus* (GAS; *Streptococcus pyogenes*) is a bacterial pathogen for which a commercial vaccine for humans is not available. Employing the advantages of high-throughput DNA sequencing technology to vaccine design, we have analyzed 2,083 globally sampled GAS genomes. The global GAS population structure reveals extensive genomic heterogeneity driven by homologous recombination and overlaid with high levels of accessory gene plasticity. We identified the existence of more than 290 clinically associated genomic phylogroups across 22 countries, highlighting challenges in designing vaccines of global utility. To determine vaccine candidate coverage, we investigated all of the previously described GAS candidate antigens for gene carriage and gene sequence heterogeneity. Only 15 of 28 vaccine antigen candidates were found to have both low naturally occurring sequence variation and high (>99%) coverage across this diverse GAS population. This technological platform for vaccine coverage determination is equally applicable to prospective GAS vaccine antigens identified in future studies.

Group A *Streptococcus* (GAS) causes >700 million cases y^{-1} of superficial diseases such as pharyngitis and impetigo, and >600,000 cases y^{-1} of serious invasive infection. Immune sequelae such as acute rheumatic fever and acute post-streptococcal glomerulonephritis each account for >400,000 cases y^{-1} (refs. [1,2]). As a consequence of acute rheumatic fever, >30 million people live with rheumatic heart disease, involving mitral and/or aortic regurgitation[3]. GAS ranks within the top ten infectious disease causes of death worldwide[1]. Despite over 100 years of research, a commercial vaccine has not been developed[2]. Obstacles that have hindered the development of a GAS vaccine include serotype diversity, GAS antigen carriage and variation, and vaccine safety concerns due to the immune sequelae caused by repeated GAS infection[2,4]. A limited number of phase I clinical trials have since been conducted,

[1]Department of Microbiology and Immunology, The Peter Doherty Institute for Infection and Immunity, The University of Melbourne and The Royal Melbourne Hospital, Melbourne, Victoria, Australia. [2]The Wellcome Trust Sanger Institute, Hinxton, UK. [3]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Queensland, Australia. [4]Australian Infectious Diseases Research Centre, The University of Queensland, Brisbane, Queensland, Australia. [5]GSK Vaccines Institute for Global Health, Siena, Italy. [6]Doherty Department, The Peter Doherty Institute for Infection and Immunity, The University of Melbourne and The Royal Melbourne Hospital, Melbourne, Victoria, Australia. [7]Department of Microbiology, New York University School of Medicine, New York, NY, USA. [8]Menzies School of Health Research, Darwin, Northern Territory, Australia. [9]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Melbourne, Victoria, Australia. [10]Molecular Bacteriology Laboratory, Université Libre de Bruxelles, Brussels, Belgium. [11]Department of Pediatrics, Queen Fabiola Childrens University Hospital, Université Libre de Bruxelles, Brussels, Belgium. [12]Murdoch Childrens Research Institute, Melbourne, Victoria, Australia. [13]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia. [14]Victorian Infectious Diseases Reference Laboratory Epidemiology Unit, The Peter Doherty Institute for Infection and Immunity, The University of Melbourne and The Royal Melbourne Hospital, Melbourne, Victoria, Australia. [15]School of Medicine, University of St Andrews, St Andrews, UK. [16]Wellcome Trust Research Centre, Kilifi, Kenya. [17]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [18]Laboratory of Bacteriology, Epidemiology Laboratory and Disease Control Division, Laboratório Central do Estado do Paraná, Curitiba, Brazil. [19]Department of Medicine, Universidade Positivo, Curitiba, Brazil. [20]Infection and Immunity Program, Biomedicine Discovery Institute and Department of Microbiology, Monash University, Melbourne, Victoria, Australia. [21]Helmholtz Centre for Infection Research, Braunschweig, Germany. [22]Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand. [23]Telethon Kids Institute, University of Western Australia and Perth Children's Hospital, Perth, Western Australia, Australia. [24]Microbiological Diagnostic Unit Public Health Laboratory, Department of Microbiology and Immunology, The Peter Doherty Institute for Infection and Immunity, The University of Melbourne and The Royal Melbourne Hospital, Melbourne, Victoria, Australia. [25]Victorian Infectious Disease Service, The Royal Melbourne Hospital, Melbourne, Victoria, Australia. [26]Department of Medicine, University of Cambridge, Cambridge, UK.
*e-mail: mark.davies1@unimelb.edu.au; mark.walker@uq.edu.au

focused primarily on multivalent amino (N)-terminal M protein vaccine candidates[5,6]. Other candidate GAS vaccine antigens that have shown efficacy in preclinical (animal) vaccine studies include the J8 peptide incorporated in the carboxy (C)-terminal repeats of M protein[7], and non-M protein candidate vaccine antigens such as the group A carbohydrate[8,9] and other surface or secreted proteins (Supplementary Table 1)[2,4]. While a number of GAS antigens have been selected to avoid autoimmune concerns[10,11] or specifically engineered to remove potential autoimmune-involved epitopes[7,9], the capacity to investigate issues of serotype diversity, antigen carriage and antigenic variation is impeded by the considerable genetic diversity within the global GAS population[12]. To address this issue, we have developed a compendium of all of the GAS vaccine antigen sequences from 2,083 isolates by employing high-throughput genomic technology.

Results

GAS population genetics. To our knowledge, we have compiled the most geographically and clinically diverse database of GAS genome sequences to date, comprising 2,083 strains, of which 645 isolates are reported for the first time (Supplementary Table 2). Extracting the classical GAS epidemiological and genotypic markers of differentiation from 2,083 genome assemblies, the database constitutes 150 *emm* types (347 *emm* subtypes)[13], 39 known M protein clusters[14] and 484 multilocus sequence types (MLSTs)[15].

To assess the genome-wide relationships within this global database, we identified the core genome of GAS to be 1,306 coding DNA sequences (CDS), based on an 80% nucleotide sequence coverage threshold and presence in >99% of the 2,083 genomes. To examine signatures of recombination within the core 1,306 genes, we analyzed each core gene separately for evidence of mosaicism using the homologous recombination detection tool fastGEAR[16]. Using this algorithm, we found 890 core genes with a recombinatorial evolutionary history (Supplementary Fig. 1 and Supplementary Table 3), leaving 416 non-recombinogenic core genes (Supplementary Table 4) encoded by 266,960 base pairs (bp) of sequence (~15% of a complete GAS genome). This number is likely to be an under-representation of the total levels of GAS core genome recombination based on the limitations in sampling (for example, the potential donor genome not being represented in the collection and/or larger recombination blocks encompassing multiple genes being missed). A pseudo-core sequence alignment was generated using these 416 core GAS genes. After the removal of repeat sequences that could confound read mapping, a total of 30,738 single nucleotide polymorphisms (SNPs) and 23,923 parsimony-informative sites were identified within the 266,960-bp pseudo-reference. Phylogenetic analysis of the 416-gene pseudo-core GAS genome identified a deep branching star-like population structure indicative of an early radiation of GAS into distinct lineages (Fig. 1a). While the overall branching topology of the tree is supported by comparing genome-specific and lineage-specific SNPs (Supplementary Fig. 2), low bootstrap support towards the polytomous root of the tree prevents accurate inferences regarding the evolutionary relationships of lineage-specific radiations (Fig. 1a). Comparative analyses of the core phylogenetic tree topologies before (1,306 genes) and after (416 genes) removal of the predicted recombinogenic CDS did not affect the overall clustering of the isolates at the terminal branches of the tree (Supplementary Fig. 3), indicating that recombination events within the 'core' GAS genome have blurred the ancestral evolutionary relationships between GAS lineages, yet have not introduced sufficient homoplasy to disrupt recent evolutionary signals.

Applying the population network approach of Population Partitioning Using Nucleotide *K*-mers (PopPUNK)[17], we identified 299 distinct genetic clusters of evolutionarily related lineages, herein termed phylogroups (Fig. 1a and Supplementary Figs. 4 and 5a). This clustering approach is derived from core and accessory genetic distances between all 2,083 genomes, using optimization of a clustering network score to find a global distance boundary to define phylogroups (Supplementary Fig. 4a,b), and is designed to be iterative, meaning that new genomes can be added to this database by using the same parameters and nomenclature as presented in this study without needing to refit the model. The median nucleotide divergence between phylogroups was 0.47% (range 0.25–0.56%), whereas genomes within the same phylogroup differed by a median divergence of 0.01% (range 0–0.14%). Of the 299 phylogroups, 206 phylogroups were represented by 2 or more isolates (Supplementary Fig. 4c). Overlaying the geographical origin of the isolates suggests that over half of these 206 phylogroups have a diverse geographical distribution (Fig. 1a). The maintenance of so many distinct genetic lineages of GAS not appearing to be restricted by geographical boundaries is suggestive of rapid international spread followed by diversifying selection probably driven through immune selection and/or strain competition between phylogroups. Furthermore, these lineages do not appear to be restricted by clinical association. For example, 172 of the 206 phylogroups (83%) contain at least one clinically defined invasive GAS isolate (Supplementary Fig. 5b). The imbalanced nature of geographical and clinical sampling in this study prevents formal statistical inferences, and such phylogroup-informed associations would require representative genomic epidemiological surveillance of the underlying population of GAS worldwide, which, to date, does not exist. Examination of the distribution of the classic GAS molecular epidemiological markers relative to the 206 multi-isolate phylogroups revealed that 179 (87%) carried a single *emm* sequence type, 140 (68%) carried a single *emm* subtype and 129 (63%) were of a single MLST (Supplementary Fig. 6). Only 3 (1.5%) of the *emm* sequence types and 55 (27%) of the *emm* subtypes were unique to a single phylogroup of 2 or more isolates, thus suggesting extensive heterogeneity within GAS *emm* types. To further investigate these associations, we plotted the pairwise genetic distance of isolates based on common GAS epidemiological markers (*emm* type, *emm* subtype, M cluster and MLST). Greater than 66% of *emm* types (84/128 multi-isolate representatives) and 32% of *emm* subtypes (65/204 multi-isolate representatives) exceeded the minimal median nucleotide divergence between any two phylogroups (0.25%, which equates to 655 SNPs within 416 core genes), showing that many *emm* types, *emm* subtypes and M clusters do not share a close evolutionary history and, in many cases, represent different genetic lineages (Supplementary Fig. 7). Conversely, <1% of MLST (2/269 multi-isolate representatives) exceeded the minimal median nucleotide divergence between phylogroups, yet MLST was a defining marker in only 27% of phylogroups. Furthermore, six of the seven MLST genes (*murI*, *xpt*, *gtr*, *gki*, *recP* and *mutS*) had evidence of homologous recombination within their evolutionary history, while another MLST gene (*yqiL*) is not part of the core GAS genome (Supplementary Tables 3 and 4). Additionally, 3 *emm*18 genomes were also identified to have a deleted *xpt* gene[18], and have been assigned the null allele xpt0 by MLST database curators. While *emm* type and MLST have served as important markers for clonal associations within high-income settings, our data suggest that *emm* type and MLST may have limited capacity for assigning evolutionary relationships within a globally evolving GAS population.

The identification of hundreds of distinct genetic lineages (299 phylogroups) represents a challenge to unraveling the microevolution of dynamically evolving bacterial populations. Indeed, only 32 of the phylogroups identified in this study contain a complete GAS reference genome ($n=68$). Furthermore, the vast majority of publicly available GAS reference genomes are of strains and *emm* types from North America and Europe, with very few reference types from high-disease-burden geographical regions. Moreover, the *emm* types circulating in these high-burden settings are often rarely encountered within high-income regions. To enable future

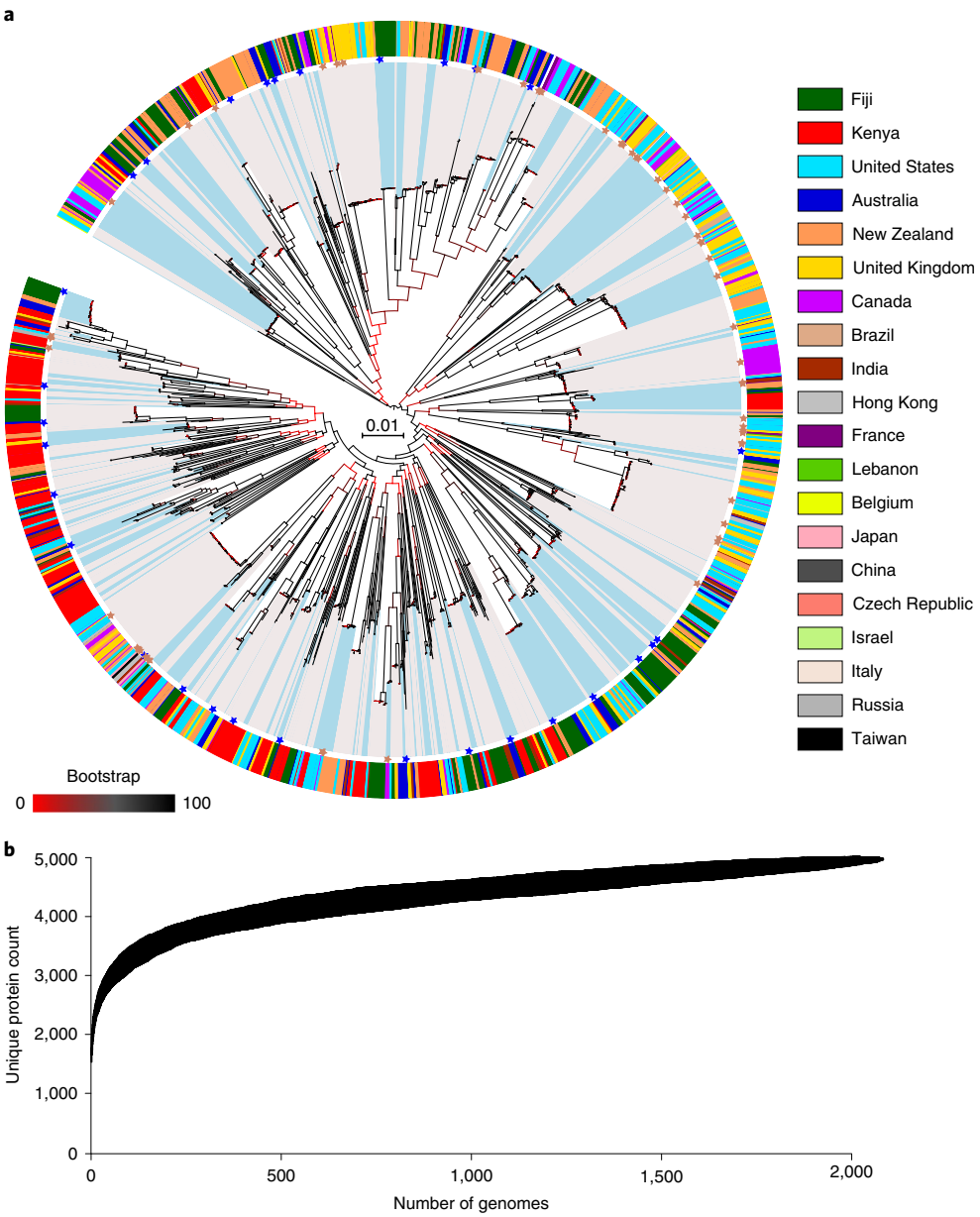

Fig. 1 | Population structure and pangenome of the 2,083 globally distributed GAS strains. a, Maximum-likelihood phylogenetic tree of 30,738 SNPs generated from an alignment of 416 core genes. Branch colors indicate bootstrap support according to the legend. Distinct genetic lineages ($n=299$) are highlighted in alternating colors (blue and gray) from the tips of the tree. Colored asterisks refer to the relative position of complete GAS reference genome sequences (existing references are shown in brown; 30 new reference genomes are shown in dark blue). Color coded around the outside of the phylogenetic tree is the country of isolation for each isolate. **b**, Pangenome accumulation curve of the 2,083 GAS genomes based on protein sequence clustering at 70% homology.

research into global and regional GAS population and evolutionary dynamics, 30 geographically and genetically distinct isolates were completely sequenced using the long-read PacBio platform (Supplementary Table 5). Based on our estimated structure of the global GAS population, these reference genomes represent 27 previously unsampled phylogroups (Fig. 1a). These high-quality geographically, clinically and evolutionarily diverse genomes will act as an important reference tool for new studies into the context of global GAS genome evolution, transmission and disease signatures.

To further assess the relative contribution of recombination on individual phylogroups, we quantifed the genome-wide rate and fragment length of recombination within 36 of the most highly sampled phylogroups (constituting 1,062 genomes). The microevolution of each lineage was assessed by mapping to a phylogroup-specific reference genome and recombination assessed by Gubbins[19]—a tool previously shown to exhibit high concordance with other recombination detection approaches[20]. The average number of SNPs observed within the 36 phylogroups was 5,536 (range 191–24,899), of which an average 20.5% of SNPs (range: 0.1–100%) were found to be vertically inherited within a phylogroup (Supplementary Table 6). Overall, the mean ratio of recombination-derived mutation versus vertically inherited mutation (r/m) was found to be 4.95 (median: 3.12), which, notably, is significantly greater than 1 (one-sample Wilcoxon test, $P=7\times10^{-7}$), suggesting that recombination is the primary driver of SNP-derived variation in GAS (Supplementary Fig. 8). The average number of recombination events per phylogroup

was 58.9 (range: 0–299) (Supplementary Table 6). Plotting the length of recombination blocks/fragments revealed that the majority of the events were small in length (<5,000 bp) with large events occurring infrequently (Supplementary Fig. 9). The average recombination fragment length in each of the 36 phylogroups was 5,437 bp, ranging from 0 bp (phylogroup 23) to 101,894 bp (phylogroup '0'). The removal of recombination events associated with putative mobile genetic elements (MGEs) had a limited effect on the total number of recombination events per phylogroup (Supplementary Fig. 9b), suggesting that heritable heterogeneity is largely MGE independent. These data highlight that evolution across the core genome of GAS lineages is not uniform and is primarily driven by small homologous recombination events.

Analysis of the variable gene content (defined as protein-coding genes present in less than 99% of the 2,083 genomes) across the entire 2,083 genomes identified 3,672 'accessory' genes when homologues were clustered at a conservative 80% amino acid identity using Roary[21] (an average of 1,717 protein-coding genes per genome). Plotting of the unique protein counts per new genome added showed that GAS has an 'open' pangenome (Fig. 1b), indicating that further genes will continue to be identified as new GAS genomes are sequenced. Annotation of the accessory genome derived from prophage analysis of the draft genome assemblies estimated ~50% of the accessory gene pool of GAS to be phage related. Plotting of the accessory content relative to the core genome phylogenetic structure of the global population revealed extensive variation both in total overall and prophage content, within and between GAS core genome lineages (Supplementary Fig. 10), in line with observations from GAS microevolutionary analyses[22–25]. Collectively, this high level of heterogeneity both in the context of core genome sequence and accessory gene content provides a unique database for the examination of disease signatures, as well as exploring conservation and sequence variation within GAS proteins such as vaccine antigens.

Disease signatures within the global GAS database. The lack of correlation between evolutionary lineages and clinical association, such as invasive infection, suggests that disease propensity is not restricted to an evolutionary lineage or clone. The interrogation of genomic databases enables an assessment of whether there are common genetic factors over-represented with a clinical phenotype, within a globally disseminated, genetically diverse bacterial population. Invasive propensity in GAS has been linked to a number of bacterial genetic factors and regulatory mutations[2,26]. To ascertain statistical support for gene content, gene polymorphisms or combinations thereof with clinical GAS invasiveness within this global genomic framework, we used the bacterial genome-wide association study method of pyseer[27]. In this study, we defined GAS isolated from a normally sterile site (blood, cerebrospinal fluid or bronchopulmonary aspirate) or severe cellulitis with positive GAS culture as invasive ($n=1,048$), and those from clinical superficial infections such as the throat, skin or urine as non-invasive ($n=896$). We included country of origin as a regression covariate, to correct for geographical bias, as previously defined[26]. Through this approach, we identified 184 hits provisionally associated with GAS invasiveness. The underlying population structure was identified as a cofounder, even though correction was applied. The confounding effect caused associations at the same *P* value across the entire genome (Supplementary Fig. 11). The top five *k*-mers that exceeded this confounder threshold include a GAS virulence marker *isp* (immunogenic secreted protein)[28], a LacI family transcriptional regulator and a hypothetical open reading frame neighboring the cysteine protease *speB* (Supplementary Table 7). Further studies are required to ascertain a link between genotype and an invasive phenotype. This analysis shows the utility of the global database for generating new disease insights.

GAS vaccine target variation. To examine natural variation of proposed GAS vaccine antigens within this genetically diverse GAS population, antigen carriage (gene presence versus absence) and amino acid sequence variation of 29 proteinaceous GAS antigens, including 4 peptide fragments, was determined (Supplementary Table 1). All identified vaccine antigens analyzed in this study have been shown to convey protection in various murine models (reviewed by Henningham et al.[4]), but little is known about the conservation of these antigens within the global GAS population. Applying a sequence-homology-based screening approach to the 2,083 GAS genome assemblies, 13 antigen genes were identified in >99% of isolates (Fig. 2a) at a 70% BLASTn cut-off. The group A carbohydrate antigen is derived from a 12-gene biosynthetic cluster (*gac*) that has displayed protective properties in an animal model[9]. Some 2,017 GAS genomes (97%) shared all 12 protein-coding genes with high DNA sequence conservation. Some genomes harbored frameshift mutations in several *gac* genes, suggesting that not all 12 genes are critical for GAS survival, commensurate with previous findings on 520 *gac* loci[29].

In addition to being omnipresent within the GAS population, an ideal GAS vaccine candidate would exhibit low levels of naturally occurring sequence variation within a genetically diverse dataset. To examine this question, pairwise BLASTp cut-off values for 25 protein antigens were calculated. A total of 18 antigens exhibited low levels (<2%) of amino acid sequence variation (Supplementary Fig. 12). When plotted relative to overall carriage within the 2,083 genomes, 13 of the 25 antigens were not only carried by >99% of the 2,083 genome sequences but also exhibited low levels of allelic variation (<2% sequence divergence) (Fig. 2b and Supplementary Fig. 12). Furthermore, 11 of these 14 core genome vaccine antigens had signatures of homologous recombination in their evolutionary history (Supplementary Fig. 13). The highest level of sequence heterogeneity in preclinical vaccine antigens was observed within the M protein. Collectively, only 33% of genomes had an N-terminal *emm* subtype (685 out of 2,083) represented within the 30-valent M protein vaccine formulation[30] (Fig. 2a). We also examined the prevalence of other GAS peptide-based vaccine antigens, namely the C-terminal M protein sequences of J8 (ref. [31]) and StreptInCor[32], as well as the S2 peptide from the serine protease SpyCEP[33]. Carriage of these peptide antigens was assessed at an exact 100% match with the query peptide sequence within the 2,083 GAS genomes. The analyses revealed that 37% of the 2,083 isolates carry the J8.0 allele of the M protein; 17% carry the conserved overlapping B- and T-cell epitope of the StreptInCor M protein vaccine candidate and 56% encode the S2 peptide from SpyCEP protein. Further interrogation of known J8 sequence variants within the multicopy M and M-like C-repeat sequences represented in the 2,083 genome assemblies identified J8.12 (79%) and J8.40 (76%) as the most frequently encountered variants (Supplementary Fig. 14).

The characterization of core gene products under different selection pressures may be used to identify potential novel vaccine antigen targets. Using the ratio of non-synonymous to synonymous codon substitutions (the d_N/d_S ratio) of each of the non-recombinogenic 416 genes, we identified that the average d_N/d_S ratio across the core GAS genome was greater than expected under a neutrality ratio of 1 (1.16), constituting 49% of core genes (205 out of 416), suggestive of overall positive selection across the GAS genome (Supplementary Table 4). Of the three 'non-recombinogenic' core vaccine targets analyzed in this study, the streptococcal haemoprotein receptor (Shr) had signatures of positive selection ($d_N/d_S=1.22$) while the hypothetical membrane-associated protein Spy0762 and the putative nucleoside-binding protein Spy0942 both exhibited signatures of purifying selection with d_N/d_S ratios of 0.57 and 0.66, respectively (Supplementary Table 4).

Antigenic heterogeneity within GAS vaccine antigens. The ascertainment of antigenic variation within genome sequence databases

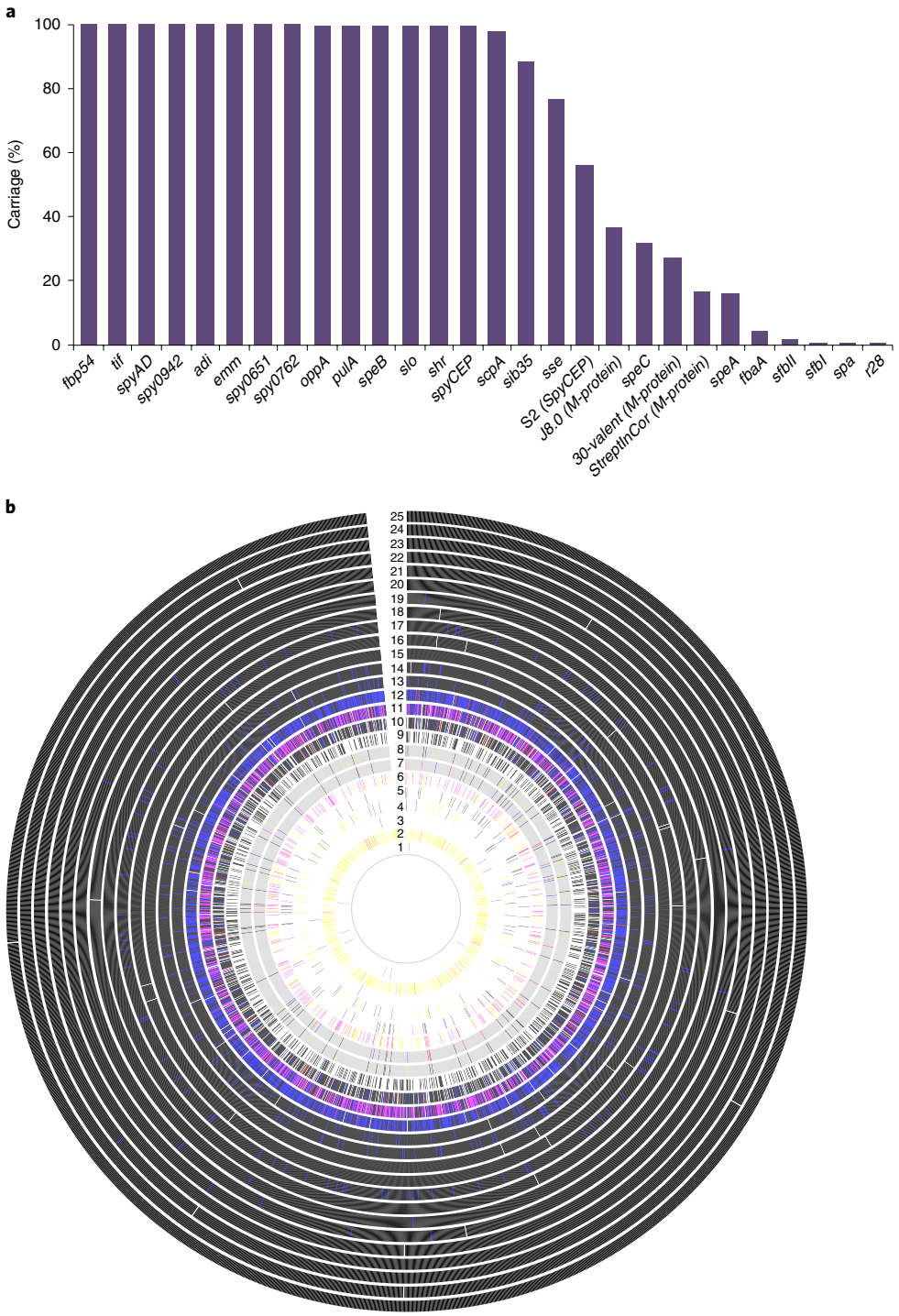

Fig. 2 | Antigenic variation within vaccine targets from the 2,083 GAS genomes. a, Gene carriage (presence versus absence) of vaccine antigens. **b**, Amino acid sequence variation within 25 protein antigens for each of the 2,083 GAS genomes. Each ring represents a single antigen, with protein similarity color coded according to pairwise BLASTp similarity: black (>98%); blue (95–98%); red (90–95%); pink (80–90%); yellow (70–80%); gray (<70%); and white (protein absence). Rings correspond to: (1) R28; (2) Sfb1; (3) Spa; (4) SfbII;(5) FbaA; (6) SpeA; (7) M1 (whole protein); (8) M1 (180-bp N terminal); (9) SpeC; (10) Sse; (11) Sib35; (12) ScpA; (13) SpyCEP; (14) PulA; (15) SLO; (16) Shr; (17) OppA; (18) SpeB; (19) Fbp54; (20) SpyAD; (21) Spy0651; (22) Spy0762;(23) Spy0942; (24) ADI; and (25) TF.

allows such data to be overlaid onto protein structures, yielding important insight regarding potential sites of structural plasticity or immunodominance, which in turn can be used to inform vaccine design through the identification of invariant surface regions and/or structurally constrained domains or subdomains. Two crystal structures are publicly available for GAS proteins that fulfil the criteria of global vaccine antigen coverage as defined in this study (>98% carriage and <2% amino acid sequence variation): streptolysin O[34] and C5a peptidase[35]. We identified polymorphism locations and polymorphism frequencies within the 2,083 GAS genomes for the streptolysin O (Fig. 3a and Supplementary Table 8) and C5a peptidase (Fig. 3b and Supplementary Table 9) proteins. Using these data,

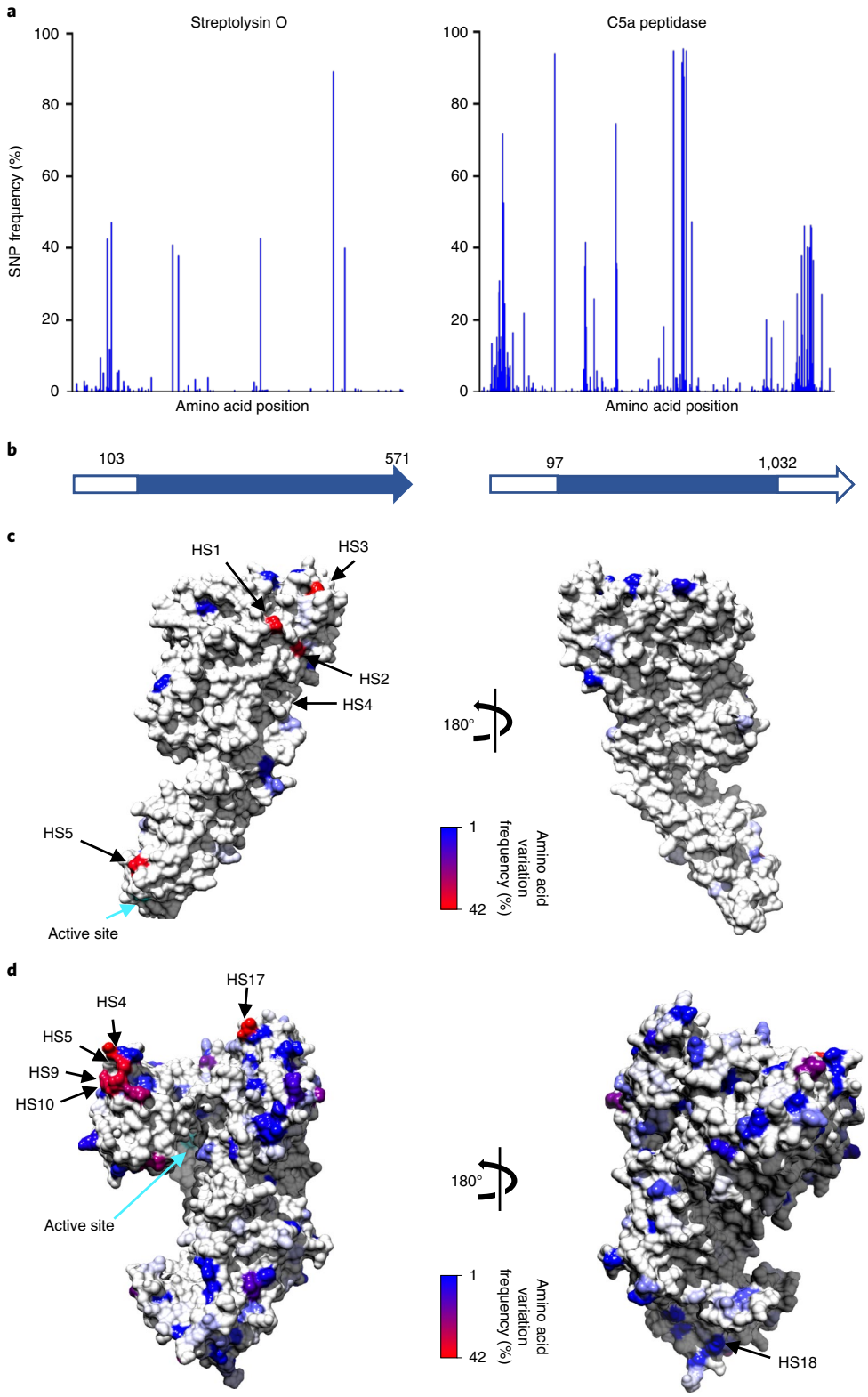

Fig. 3 | Global amino acid variation mapped onto the protein crystal structure of the mature GAS streptolysin O and C5a peptidase. a, Frequency of amino acid variations within the 2,083 genomes for streptolysin O[34] (left) and C5a peptidase[35] (right). **b**, Schematic of the streptolysin O and C5a peptidase open reading frames, representing the locations of amino acids within the mature enzymes (blue blocks). **c**,**d**, Models of the consensus sequences of the streptolysin O (**c**) and C5a peptidase (**d**) mature enzymes. Plotted against each structure is the amino acid variation frequency within the 2,083 GAS genomes, as represented in the color gradient from 1% (blue) to 42% variable (red); invariant sites are colored light gray. The positions of the top five most variable surface hot spots (HS) are annotated (as defined in Supplementary Tables 10 and 11). Active sites for each enzyme are also indicated (cyan arrow).

Table 1 | Comparative ratio of nucleotide changes resulting from recombination relative to point mutation (*r/m*) in selected bacterial pathogens

Species	*r/m* ratio (genome wide)	*r/m* ratio (MLST derived)[a]	References
S. pyogenes	4.95	17.2	This study and Enright et al.[15]
S. pneumoniae	6.36	23.1	Chaguza et al.[53] and Hanage et al.[54]
Staphylococcus aureus	0.6	0.1	Driebe et al.[55] and Enright et al.[56]
L. pneumophila	47.8	0.9	David et al.[46] and Coscolla and Gonzalez-Candelas[57]
K. pneumoniae	4.75	0.3	Diancourt et al.[58] and Wyres et al.[59]

[a]MLST allele-derived *r/m* ratios, as defined by Vos and Didelot[44].

we derived the consensus amino acid sequence for each protein. We then modeled the consensus sequence and population-derived polymorphisms onto the corresponding crystal structures of the mature streptolysin O protein (amino acids 103–501; Fig. 3b,c)[34] and C5a peptidase (amino acids 97–1,032; Fig. 3b,d)[35]. Using data extracted from the 2,083 genomes, further examination of amino acid heterogeneity present within the mature streptolysin O protein revealed 5 sequence diversity hot spots (Fig. 3c). All hot spot polymorphisms were bimorphic in nature, indicating restrictions in streptolysin O plasticity (Supplementary Table 10). In comparison, we identified 20 sequence diversity hot spots within the mature C5a peptidase protein, of which half were bimorphic (Fig. 3a and Supplementary Table 11), indicating that more plasticity can be accommodated within the C5a peptidase than streptolysin O. To ascertain the functional consequence of the most common protein variations, we examined the mutational sensitivity and structural integrity of these amino acid variants by using Phyre2 (ref. [36]) and the SuSPect platform[37]. All substitutions in both streptolysin O and C5a peptidase were at locations where it was predicted that a change in the amino acid would probably not impact protein structure or activity (Supplementary Tables 10 and 11). To further examine the selective pressures within these antigens, we assessed the selective constraints at each codon position. We found that 10.5% (60/571) of amino acid residues had higher diversity at first and second codon positions than at third codon positions for streptolysin O (versus 16.5% (170/1,032) for C5a peptidase), indicating that these sites are undergoing positive selection (Supplementary Tables 8 and 9). Of the diversity hot spots, 40% (2/5) of the streptolysin O sites and 60% (12/20) of the C5a peptidase sites showed signatures of positive selection. These data may reflect immune selection and/or the amount of plasticity that can be encompassed without compromising protein function.

Discussion

There is a strong case for the development of a safe and efficacious GAS vaccine[1,2]. One of several hurdles to be addressed in the development of a GAS vaccine suitable for worldwide use is the extensive genetic diversity of the global GAS population. To address issues of vaccine antigen gene carriage within the global GAS population and the extensive variation of antigen amino acid sequences between isolates, we have developed a platform for the interrogation of candidate antigens at unprecedented resolution. We have demonstrated that GAS is a genetically diverse species containing a large dispensable gene pool. Within the core or 'conserved' genome, we have identified extensive evidence of recombination that will initiate future research into the biology and underlying drivers of such dynamic evolution. This diversity also has consequences for vaccine-induced evolutionary sweeps of bacterial populations and the subsequent emergence of vaccine escape clones, as has been observed in serotype-specific *Streptococcus pneumoniae*[38] and *Bordetella pertussis*[39] vaccination programs. Our findings identify that selection pressures are variable across the core GAS genome and proposed vaccine candidates, and probably reflective of distinct and ongoing evolutionary adaptation. Collectively, within an evolving global bacterial pathogen such as GAS, we have identified that a number of proposed preclinical GAS vaccine antigens fulfil the criteria for a global vaccine. It is tempting to speculate that multiantigenic formulations would provide an ideal approach against a rapidly evolving pathogen, as well as increasing global coverage. Indeed, the incorporation of additional antigens within existing serotype-specific approaches in GAS enhances theoretical vaccine coverage[40] (Supplementary Table 12).

We reveal that the global population structure of GAS is one of extensive genetic diversity, which is probably reflective of the rapid international spread of genetically diverse lineages driven by diversifying selection from the immune system and/or competition between lineages. This may lead to negative frequency-dependant selection, as has been proposed for other human bacterial pathogens such as *S. pneumoniae* and *Escherichia coli*[41,42]. Recombination has previously been identified to be high in GAS[43,44] and, at a genome-wide population level, our findings suggest a major role for homologous recombination of small DNA fragments in driving the evolutionary dynamics of GAS, indicating that evolution of GAS lineages is more likely to arise by recombination rather than by mutation[43]. All GAS lineages do not evolve at the same rate and this is likely to have key, yet undefined, biological significance. Similar impacts and rates of homologous recombination have been observed in other bacterial pathogens such as *S. pneumoniae*[45] and *Legionella pneumophila*[46]. A comparison of the relative rates of recombination versus mutation, based on whole-genome and gene-restricted MLST approaches, places *Streptococcus pyogenes* with other highly recombinogenic species such as *Klebsiella pneumoniae* and *S. pneumoniae* (Table 1).

The generation of high-quality, well-curated reference genomes acts as a landmark for understanding the evolutionary context of a species, especially given the high levels of genetic diversity encountered in bacterial populations such as GAS and the contrasting epidemiology of infection observed between high-income countries and lower socioeconomic regions of the world, which accounts for the overwhelming burden of GAS disease. The availability of new GAS reference genomes will enable targeted evolutionary and pathobiological studies of this genetically diverse pathogen. The 30 new GAS reference genomes reveal that despite an open pangenome where accessory gene content varies significantly across the population and recombination appears frequent, the overall size of the GAS genome remains at a steady state. Only recently have plasmids been characterized within the GAS genome[47,48]. We have identified a further 5 small plasmids in GAS ranging in size from 2,645–6,485 bp, harboring bacteriocin-like genetic markers that are suggested to play a role in interbacterial inhibition[49]. In the context of vaccination, the availability of a globally representative reference database will provide a platform for examining the effect of future vaccination programs[38,39].

The modeling of population-based antigenic variation against protein crystal structures enables the identification of residues that may be under functional or structural constraints, or alternatively, selection pressure. This population-derived sequence approach could be assessed alongside immunological studies to define protective epitopes. Such information can be incorporated into further refinement of vaccine antigens such as peptide-based approaches

that factor in naturally occurring population heterogeneity, enabling the targeting of immunogenic epitopes within antigens that are less amenable to variation.

This platform for population-genomics-informed vaccine design is equally applicable to all known GAS antigens and those that remain to be defined. Thus, informed selection of putative vaccine antigens for human trial evaluation will now be possible, allowing the identification of highly conserved antigens or combinations of antigens that ensure complete vaccine coverage across GAS *emm* types from differing geographic regions. For example, GAS vaccine antigens such as streptolysin O, the IL8 protease SpyCEP, arginine deaminase, trigger factor and C5a peptidase, which were found here to be highly conserved across geographic regions, protect against multiple GAS *emm* types in animal models[10,50,51,52]. An approach similar to that used in this study would also be applicable to other pathogens that exhibit high levels of global strain diversity.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0417-8.

Received: 16 May 2018; Accepted: 10 April 2019;
Published online: 27 May 2019

References

1. Carapetis, J. R., Steer, A. C., Mulholland, E. K. & Weber, M. The global burden of group A streptococcal diseases. *Lancet Infect. Dis.* **5**, 685–694 (2005).
2. Walker, M. J. et al. Disease manifestations and pathogenic mechanisms of group A *Streptococcus*. *Clin. Microbiol. Rev.* **27**, 264–301 (2014).
3. Watkins, D. A. et al. Global, regional, and national burden of rheumatic heart disease, 1990–2015. *N. Engl. J. Med.* **377**, 713–722 (2017).
4. Henningham, A., Gillen, C. M. & Walker, M. J. Group A streptococcal vaccine candidates: potential for the development of a human vaccine. *Curr. Top. Microbiol. Immunol.* **368**, 207–242 (2013).
5. Kotloff, K. L. et al. Safety and immunogenicity of a recombinant multivalent group A streptococcal vaccine in healthy adults: phase 1 trial. *J. Am. Med. Assoc.* **292**, 709–715 (2004).
6. McNeil, S. A. et al. Safety and immunogenicity of 26-valent group A *Streptococcus* vaccine in healthy adult volunteers. *Clin. Infect. Dis.* **41**, 1114–1122 (2005).
7. Brandt, E. R. et al. New multi-determinant strategy for a group A streptococcal vaccine designed for the Australian Aboriginal population. *Nat. Med.* **6**, 455–459 (2000).
8. Sabharwal, H. et al. Group A *Streptococcus* (GAS) carbohydrate as an immunogen for protection against GAS infection. *J. Infect. Dis.* **193**, 129–135 (2006).
9. Van Sorge, N. M. et al. The classical lancefield antigen of group A *Streptococcus* is a virulence determinant with implications for vaccine design. *Cell Host Microbe* **15**, 729–740 (2014).
10. Henningham, A. et al. Conserved anchorless surface proteins as group A streptococcal vaccine candidates. *J. Mol. Med. (Berl.)* **90**, 1197–1207 (2012).
11. Valentin-Weigand, P., Talay, S. R., Kaufhold, A., Timmis, K. N. & Chhatwal, G. S. The fibronectin binding domain of the Sfb protein adhesin of *Streptococcus pyogenes* occurs in many group A streptococci and does not cross-react with heart myosin. *Micro. Pathog.* **17**, 111–120 (1994).
12. Steer, A. C., Law, I., Matatolu, L., Beall, B. W. & Carapetis, J. R. Global *emm* type distribution of group A streptococci: systematic review and implications for vaccine development. *Lancet Infect. Dis.* **9**, 611–616 (2009).
13. 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).
14. Sanderson-Smith, M. et al. A systematic and functional classification of *Streptococcus pyogenes* that serves as a new tool for molecular typing and vaccine development. *J. Infect. Dis.* **210**, 1325–1338 (2014).
15. Enright, M. C., Spratt, B. G., Kalia, A., Cross, J. H. & Bessen, D. E. Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between *emm* type and clone. *Infect. Immun.* **69**, 2416–2427 (2001).
16. Mostowy, R. et al. Efficient inference of recent and ancestral recombination within bacterial populations. *Mol. Biol. Evol.* **34**, 1167–1182 (2017).
17. Lees, J. A. et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29**, 304–316 (2019).
18. Chochua, S. et al. Population and whole genome sequence based characterization of invasive group A streptococci recovered in the United States during 2015. *MBio* **8**, e01422-17 (2017).
19. 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).
20. Marttinen, P. et al. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* **40**, e6 (2012).
21. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
22. Beres, S. B. et al. Genome-wide molecular dissection of serotype M3 group A *Streptococcus* strains causing two epidemics of invasive infections. *Proc. Natl Acad. Sci. USA* **101**, 11833–11838 (2004).
23. Nasser, W. et al. Evolutionary pathway to increased virulence and epidemic group A *Streptococcus* disease derived from 3,615 genome sequences. *Proc. Natl Acad. Sci. USA* **111**, E1768–E1776 (2014).
24. Turner, C. E. et al. Emergence of a new highly successful acapsular group A *Streptococcus* clade of genotype *emm89* in the United Kingdom. *MBio* **6**, e00622 (2015).
25. You, Y. et al. Scarlet fever epidemic in China caused by *Streptococcus pyogenes* serotype M12: epidemiologic and molecular analysis. *EBioMedicine* **28**, 128–135 (2018).
26. Lees, J. A. et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.* **7**, 12797 (2016).
27. Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* **34**, 4310–4312 (2018).
28. McIver, K. S., Subbarao, S., Kellner, E. M., Heath, A. S. & Scott, J. R. Identification of *isp*, a locus encoding an immunogenic secreted protein conserved among group A streptococci. *Infect. Immun.* **64**, 2548–2555 (1996).
29. Henningham, A. et al. Virulence role of the GlcNAc side chain of the Lancefield cell wall carbohydrate antigen in non-M1-serotype group A *Streptococcus*. *MBio* **9**, e02294-17 (2018).
30. Dale, J. B., Penfound, T. A., Chiang, E. Y. & Walton, W. J. New 30-valent M protein-based vaccine evokes cross-opsonic antibodies against non-vaccine serotypes of group A streptococci. *Vaccine* **29**, 8175–8178 (2011).
31. Batzloff, M. R. et al. Protection against group A *Streptococcus* by immunization with J8-diphtheria toxoid: contribution of J8- and diphtheria toxoid-specific antibodies to protection. *J. Infect. Dis.* **187**, 1598–1608 (2003).
32. Guilherme, L. et al. Towards a vaccine against rheumatic fever. *Clin. Dev. Immunol.* **13**, 125–132 (2006).
33. Pandey, M. et al. Combinatorial synthetic peptide vaccine strategy protects against hypervirulent CovR/S mutant streptococci. *J. Immunol.* **196**, 3364–3374 (2016).
34. Feil, S. C., Ascher, D. B., Kuiper, M. J., Tweten, R. K. & Parker, M. W. Structural studies of *Streptococcus pyogenes* streptolysin O provide insights into the early steps of membrane penetration. *J. Mol. Biol.* **426**, 785–792 (2014).
35. Kagawa, T. F. et al. Model for substrate interactions in C5a peptidase from *Streptococcus pyogenes*: a 1.9 A crystal structure of the active form of ScpA. *J. Mol. Biol.* **386**, 754–772 (2009).
36. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
37. Yates, C. M., Filippis, I., Kelley, L. A. & Sternberg, M. J. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J. Mol. Biol.* **426**, 2692–2701 (2014).
38. Croucher, N. J. et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* **45**, 656–663 (2013).
39. Bart, M. J. et al. Global population structure and evolution of *Bordetella pertussis* and their relationship with vaccination. *MBio* **5**, e01074 (2014).
40. Courtney, H. S. et al. Trivalent M-related protein as a component of next generation group A streptococcal vaccines. *Clin. Exp. Vaccin. Res.* **6**, 45–49 (2017).
41. Corander, J. et al. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat. Ecol. Evol.* **1**, 1950–1960 (2017).
42. McNally, A. et al. Signatures of negative frequency dependent selection in colonisation factors and the evolution of a multi-drug resistant lineage of *Escherichia coli*. Preprint at *bioRxiv* https://doi.org/10.1101/400374 (2018).
43. Bao, Y. J., Shapiro, B. J., Lee, S. W., Ploplis, V. A. & Castellino, F. J. Phenotypic differentiation of *Streptococcus pyogenes* populations is induced by recombination-driven gene-specific sweeps. *Sci. Rep.* **6**, 36644 (2016).
44. Vos, M. & Didelot, X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* **3**, 199–208 (2009).
45. Chewapreecha, C. et al. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46**, 305–309 (2014).
46. David, S. et al. Dynamics and impact of homologous recombination on the evolution of *Legionella pneumophila*. *PLoS Genet.* **13**, e1006855 (2017).

47. Bergmann, R., Nerlich, A., Chhatwal, G. S. & Nitsche-Schmitz, D. P. Distribution of small native plasmids in *Streptococcus pyogenes* in India. *Int. J. Med. Microbiol.* **304**, 370–378 (2014).
48. Woodbury, R. L. et al. Plasmid-borne *erm*(T) from invasive, macrolide-resistant *Streptococcus pyogenes* strains. *Antimicrob. Agents Chemother.* **52**, 1140–1143 (2008).
49. Wescombe, P. A., Heng, N. C., Burton, J. P., Chilcott, C. N. & Tagg, J. R. Streptococcal bacteriocins and the case for *Streptococcus salivarius* as model oral probiotics. *Future Microbiol.* **4**, 819–835 (2009).
50. Bensi, G. et al. Multi high-throughput approach for highly selective identification of vaccine candidates: the group A *Streptococcus* case. *Mol. Cell Proteom.* **11**, 015693 (2012).
51. Ji, Y., Carlson, B., Kondagunta, A. & Cleary, P. P. Intranasal immunization with C5a peptidase prevents nasopharyngeal colonization of mice by the group A *Streptococcus*. *Infect. Immun.* **65**, 2080–2087 (1997).
52. Rivera-Hernandez, T. et al. An experimental group A vaccine that reduces pharyngitis and tonsillitis in a nonhuman primate model. *MBio* **10**, e00693-19 (2019).
53. Chaguza, C. et al. Recombination in *Streptococcus pneumoniae* lineages increase with carriage duration and size of the polysaccharide capsule. *MBio* **7**, e01053-16 (2016).
54. Hanage, W. P. et al. Using multilocus sequence data to define the pneumococcus. *J. Bacteriol.* **187**, 6223–6230 (2005).
55. Driebe, E. M. et al. Using whole genome analysis to examine recombination across diverse sequence types of *Staphylococcus aureus*. *PLoS ONE* **10**, e0130955 (2015).
56. Enright, M. C., Day, N. P., Davies, C. E., Peacock, S. J. & Spratt, B. G. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J. Clin. Microbiol.* **38**, 1008–1015 (2000).
57. Coscolla, M. & Gonzalez-Candelas, F. Population structure and recombination in environmental isolates of *Legionella pneumophila*. *Environ. Microbiol.* **9**, 643–656 (2007).
58. Diancourt, L., Passet, V., Verhoef, J., Grimont, P. A. & Brisse, S. Multilocus sequence typing of *Klebsiella pneumoniae* nosocomial isolates. *J. Clin. Microbiol.* **43**, 4178–4182 (2005).
59. Wyres, K. L. et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. *PLoS Genet.* **15**, e1008114 (2019).

Acknowledgements

This work was supported by National Health and Medical Research Council project and program grants for: Protein Glycan Interactions in Infectious Diseases and Cellular Microbiology; the Coalition to Accelerate New Vaccines Against *Streptococcus* (CANVAS; an Australian and New Zealand joint initiative); and The Wellcome Trust, UK. For part of this study, M.R.D. was supported by a National Health and Medical Research Council postdoctoral training fellowship (635250) and A.M. was a GENDRIVAX fellow funded by the European Union's Seventh Framework Programme FP7/2007–2013/ under REA grant agreement number 251522. We acknowledge assistance from the sequencing and pathogen informatics core teams at The Wellcome Trust Sanger Institute. We acknowledge and thank the database curators of the *S. pyogenes* MLST and *emm* databases (especially D. Bessen). We dedicate this work to the memory of our friend and colleague Gusharan Singh Chhatwal.

Author contributions

M.R.D., G.D. and M.J.W. conceived the project. M.R.D., A.M., J.A.Lacey, J.A.Lees, S.Duchene, P.R.S., M.T.G.H., S.Y.C.T., P.M.G., A.C.S., J.A.B., G.S.C., S.D.B., R.A.S., T.L., J.D.F., N.J.M., J.R.C., A.C.S., J.P., A.S., D.A.W., B.J.C. and M.J.W. designed the experiments. M.R.D., L.M., J.A.Lacey, J.A.Lees, S.David, A.M., R.J.T., K.A.W., S.R.H., T.R.-H., H.R.F., R.S.L.A.T., O.B., A.J.C., R.B., P.N.-S., N.J.M. and D.A.W. performed the experimental protocols. M.R.D., L.M., J.A.Lacey, J.A.Lees, S.Duchene, D.J.P., A.M., P.R.S., N.J.M., G.D. and M.J.W. analyzed the experimental results. M.R.D. and M.J.W. wrote the manuscript and all authors reviewed the manuscript.

Competing interests

A.S. is an employee of the GSK group of companies with a commercial interest in GAS vaccine development. These companies had no influence over study design.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0417-8.

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

Correspondence and requests for materials should be addressed to M.R.D. or M.J.W.

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

Methods

Bacterial isolates. The global collection of 2,083 *S. pyogenes* isolates examined in this study included short-read genome sequence data from population-based studies that we have generated within Kenya[60] and Fiji[26], and other disease-specific, population-based studies of invasive GAS from Canada[61], the United States[18] and the United Kingdom[62,63], that were available as of 1 July 2018. We selected a small subset of isolates from published microevolution (outbreak) studies to avoid biasing the collection on single genetically related lineages. A total of 68 GAS reference genomes and publicly available draft genomes from Lebanon[64] were also included. To increase genomic representation from regions endemic for GAS infection and other undersampled geographical regions, we collected a further 271 isolates from Australia, 279 isolates from New Zealand, 50 isolates from Brazil, 45 isolates from India and 7 isolates from Belgium. The rationale underpinning isolate selection was a difference in epidemiological markers (*emm* type), the anatomical site of isolation (skin, throat and blood) and clinical presentation, all of which are key factors in GAS vaccine design. Metadata pertaining to the database of isolates are provided in Supplementary Table 2.

Genome sequencing and assembly. Genomic DNA was extracted, and paired-end multiplex libraries were created and sequenced using the Illumina HiSeq 2500 platform at the read length between 75 and 125 bp (The Wellcome Trust Sanger Institute, United Kingdom). Draft genome sequences were generated using an iterative Velvet-based assembly pipeline with secondary read mapping validation[65] or using SKESA version 2.3.0 (ref. [66]) with default parameters. Gene predictions and annotations were generated using PROKKA[67] and streptococcal RefSeq-specific databases[65]. Annotations pertaining to the *mga* locus (including *emm* and *emm*-like genes) were manually curated using in-house databases due to ambiguity when using pipeline procedures. The assembly pipeline generated assemblies of an average length of 1,791,171 bp (range: 1,641,039–1,986,343 bp) and an N50 of 252,789 bp (range: 2,276–1,953,601 bp). On average, 1,711 CDS were identified per draft genome (range: 1,495–1,976 CDS). All draft genome assemblies are publicly available through GenBank. Accession numbers are listed in Supplementary Table 2.

Sequence mapping. To examine the genetic relationships of the 2,083 GAS genome sequences, we employed a single reference-based mapping approach using subsampled Illumina FASTQs at an estimated coverage of 75×. Published reference and draft genome datasets accessed from public databases were each shredded into an estimated 75× coverage of paired-end 100-bp reads using SAMtools wgsim. Sequence reads were mapped to the M1 GAS reference genome MGAS5005 (GenBank accession number CP000017)[68] with BWA MEM (version 0.7.16) and read depths were calculated with SAMtools (version 1.6) with a Phred quality score of ≥20. SNPs with a Phred quality score of ≥30 were identified in each isolate by using SAMtools pileup with a minimum coverage of 10×. Core genes were defined as a minimum 80% of the MGAS5005 reference gene with a minimum 10× coverage. Using this approach, we identified 1,306 MGAS5005 genes with 99% carriage in 2,083 genomes. A core SNP genome alignment of 171,273 SNPs was generated by concatenating the SNPs located within the 1,306 core genes, giving a total of 1,201,767 bp. SNPs residing within repeat regions (minimum length: 20 nucleotides) and mobile genetic elements are considered evolutionary confounders and were identified as previously described[69] or by using PHASTER[70]. SNPs within these regions were excised from the core alignment, reducing the length from 1,201,767 bp to 1,197,326 bp and the SNP count from 171,273 to 170,653. Therefore, a total of 170,653 SNPs were aligned for phylogenetic analysis of the 1,306-gene 'core' genome (Supplementary Fig. 3).

Recombination detection. To examine evidence of recombination within the core GAS genome, FastGEAR[16] was run on 1,306 individual gene alignments, comprising all 2,083 GAS strains included in the study. This method infers population structure for each alignment, allowing for the detection of lineages that have ancestral and recent recombinations between them. Default parameters were used with a minimum threshold of 4 bp applied for the recombination length. A total of 890 genes had signatures of recombination and were excluded from evolutionary analyses. The remaining 416 genes were concatenated, corresponding to 268,003 bp of sequence. SNPs residing within repeat regions were removed as described above, resulting in 266,960 bp of sequence used as a best estimate for the global GAS population structure.

For intraphylogroup recombination analyses, the 36 most highly represented PopPUNK phylogroups were chosen to investigate the influence of recombination (1,062 isolates). For each phylogroup, core genome alignments were performed using Snippy version 4.3.5 (https://github.com/tseemann/snippy) against a reference strain within each phylogroup (Supplementary Table 6), and maximum-likelihood trees were inferred using IQ-TREE version 1.6.5 (ref. [71]), These trees were used as inputs for the recombination detection tool Gubbins version 2.3.4 (ref. [19]). Gubbins was run with a maximum of 20 iterations, a minimum of 5 SNPs to identify a recombination block, a window size of 100–10,000 bp, and any taxa with more than 25% gaps filtered from the analysis. Recombinogenic blocks that overlapped with predicted MGEs in the reference genome were discarded. Phage regions were determined using PHASTER[70], and integrative conjugative elements were determined by manual inspection of reference genomes based on the similarity of BLAST hits from known integrative conjugative elements. Recombination versus *r*/*m* ratios for each lineage were calculated as the average *r*/*m* including all isolates within the phylogroup. For the species, *r*/*m* was determined by averaging across all 36 phylogroups (Table 1).

Phylogenetic analysis. Maximum-likelihood trees were generated for the 416- and 1,306-gene core genome alignments, using IQ-TREE version 1.6.5 (ref. [71]). The generalized time-reversible nucleotide substitution with gamma correction for site-specific rate variation was performed with 100 bootstrap random resamplings of the alignment data to support maximum-likelihood bipartitions. For figure generation, phylogenetic trees and associated metadata were collated using the web portal Interactive Tree of Life[72].

Population genomics and cluster designation. To define evolutionarily related clusters (phylogroups) in the population, we used PopPUNK, which has previously been shown to give high-quality clusters in a subset of the *S. pyogenes* isolates included in this study[17]. We used *k*-mers between 15 and 29 nucleotides in length, in steps of two, to calculate core and accessory distances between all pairs of isolates (Supplementary Fig. 4a). We clustered these distances first with the default two-component Bayesian Gaussian mixture model, then used the 'refine fit' mode to move the boundary of this fit such that the network was highly transitive and sparse, obtaining a network score of 0.980 (Supplementary Fig. 4b,c). To increase the utility of the GAS population clusters defined here, we created a database so that others can assign sample clusters using the same model and nomenclature as we present here. To do this, we used PopPUNK to extract one sample per clique in the network, giving a reduced-size query database containing 359 sequences. This database can be accessed at https://doi.org/10.6084/m9.figshare.6931439.v1 and contains an example command for database query and future expansion. The PopPUNK cluster designations ('phylogroups') for the 2,083 genomes have been added to Supplementary Table 2 and the Microreact[73] interactive web application (https://microreact.org/project/5DEFpeck4).

Nucleotide divergence was derived by calculating the pairwise hamming distance from the 416-gene core genome alignment (266,960 bp). For pairwise hamming distance plots based on epidemiological markers (Supplementary Fig. 7), a reference genome was assigned for each marker based on the most representative distance within each type (minimum combined hamming distance) from the 416-gene core genome alignment.

Pangenome analysis. The pangenome was defined using Roary version 3.11.2 (ref. [21]) without splitting paralogues and with clustering at 80%. The accessory genome was defined as the pan less the core, totaling 3,672 genes. The identification of prophage CDS within each of the 2,083 genomes was performed using PHASTER[70]. Clustering with CD-HIT-EST[74] at ≤90% nucleotide homology resulted in 1,438 gene clusters. Some 584 core genes and 1,567 accessory genes hit these phage regions with BLASTn version 2.3.0+ with a 90% nucleotide cut-off over 90% of the gene length. These data were then processed to generate a binary gene content matrix in which the presence of a gene was defined as >90% coverage in a corresponding phage gene cluster.

Vaccine antigen screening pipeline. To examine the naturally occurring antigenic variation of proposed GAS vaccine targets within this genetically diverse GAS population, the carriages of 29 vaccine antigens (Supplementary Table 1) and the group A carbohydrate biosynthesis loci were determined. The vaccine antigens screened have been shown to convey a significant level of protection in murine models[4], but less is known about the conservation of these antigens within a global context. The presence of vaccine antigen genes was determined by BLASTn analysis of 2,083 genome assemblies based on a 70% nucleotide cut-off over 70% of the gene length. Nucleotide sequences of whole and N-terminal regions of the M protein were extracted using publicly available databases to account for known higher levels of allelic variation. These data were then converted into a binary gene content matrix in which gene presence was defined as >70% homology across a minimum of 70% of the query gene length. Allelic variation was examined by plotting tBLASTn (or BLASTn for group A carbohydrate genes) scores relevant to the query reference sequence. To facilitate future studies assessing vaccine antigen carriage and sequence variation within GAS genome sequences, we generated a bioinformatics pipeline for assessing antigenic variation from genome assemblies. This script, as used in this study, is available at https://github.com/shimbalama/screen_assembly. It requires a query sequence (such as a vaccine antigen) and will run BLASTn, tBLASTn or BLASTp at a user-defined cut-off, generating numerous outputs and plots as represented in this study (see Fig. 3a, Supplementary Fig. 13 and Supplementary Tables 8 and 9). Furthermore, this screening approach is applicable to any pathogen where genome assemblies are supplied.

Streptolysin O and C5a peptidase surface variation. Protein sequences of streptolysin O and C5a peptidase were chosen for further analyses since well-characterized crystal structures exist for each of these GAS antigens. Protein alignments corresponding to the published crystalized structures of streptolysin O (amino acid residues 103–571; Protein Data Bank accession number PDB 4HSC[34])

and C5a peptidase (amino acid residues 97–1,032; Protein Data Bank accession number PDB 3EIF[35]) were generated. Using these data, we derived the consensus amino acid sequence for each protein as defined by the most common amino acid identified within the global GAS genome database, and modeled the consensus against the mature crystal structures. Amino acid polymorphic sites were converted into a binary matrix and presented as a percentage of 2,083 genomes in Fig. 3. Visualization of polymorphic sites on the crystal structure was determined using Chimera (version 1.11.2)[75]. Mutational sensitivity and structural integrity analyses were performed using Phyre2 (ref. [36]), which incorporates the SuSPect platform[37].

Signatures of molecular adaptation. We investigated molecular signatures of selective constraints in all non-recombinogenic core genes ($n=416$) by fitting a codon model to each of the individual genes and estimating the d_N/d_S ratio (also known as ω). Recombinogenic core genes ($n=890$), as identified by fastGEAR, were excluded from the analyses as such evolutionary processes invalidate phylogenetic codon model fitting. For each gene alignment, ambiguous codon sites were first excluded, before fitting the M0 codon model in CODEML, (part of the PAML version 4.0 package[76]). This model estimates a global d_N/d_S, which allows for straightforward comparison between genes. For the streptolysin O and C5a peptidase protein-coding genes, we conducted more detailed analyses by assessing selective constraints across codon sites. To do this, we counted the number synonymous and non-synonymous substitutions in each codon position, to obtain a similar quantity to the d_N/d_S value above[77]. Although this method does not explicitly use a codon model, it is scalable for the large number of samples used here. Despite the objective of this study being centred around global diversity, our database contains sample bias in the context of clinical and geographical sampling, and the selection analyses should be interpreted carefully, as they may not represent current global selective trends.

Generation of 30 new GAS reference genomes. The vast majority of publicly available completely sequenced reference genomes are of *emm* types from North America and Europe, while very few are of *emm* types from high-disease-burden geographical regions. To facilitate the expansion of studies within the highest-disease-burden regions, 30 isolates were completely sequenced using long-read sequencing technology. Long-read sequences were obtained using the PacBio RS II platform from a single-molecule real-time cell, as described previously[78]. Briefly, genome sequences were assembled using the SMRTpipe version 2.1.0 using the Hierarchical Genome Assembly Process (HGAP.2) and Quiver for post-assembly consensus validation. Secondary validation of the assemblies was performed using the Canu assembler[79]. To correct long-read sequence errors, primarily around homopolymeric regions, Illumina short-read sequences from each of the 30 genomes were mapped using BWA MEM version 0.7.16. Single contigs were achieved for all genomes and associated plasmids where present, with an average coverage depth of 80×. Genomes were annotated using the same pipeline as for the Illumina draft genomes[65], with putative prophage regions defined using the PHASTER server[70]. The average size of these new reference genomes was 1,810,671 bp (ranging from 1,701,466–1,950,606), with 5 strains containing circular plasmids ranging from 2,645–6,485 bp in size (Supplementary Table 5).

Genome-wide association of GAS invasiveness. To identify genomic signatures within the global GAS population over-represented with severe GAS infection, we ran pyseer[27] on 1,944 samples (1,048 defined as invasive), using the linear mixed model. A total of 87 million *k*-mers between 9 and 100 bases in length were counted using fsm-lite. We tested only common *k*-mers (that is, those with a minor allele frequency > 1%, of which 18 million were counted in our dataset). We created a kinship matrix from our recombination-free core phylogenetic tree of 2,083 genomes (416 genes; Fig. 1a). The country of isolation was used as a covariate in the pyseer model to account for geographical signal, as defined previously[26]. All *k*-mers were mapped to the MGAS5005 GAS reference genome using bwa, and visualized with R. We used Bonferroni correction to adjust the significance threshold past the number of unique patterns tested, which gave 9.4×10^{-7} for a 0.05 family-wise error rate. In total, 184 *k*-mers were significantly associated with severe infection.

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

Data availability

Illumina sequence reads and draft genome assemblies were deposited to the European Nucleotide Archive under the accession numbers specified in Supplementary Table 2. GenBank accession numbers for the 30 new GAS reference genomes are provided in Supplementary Table 5. To facilitate community accessibility and interrogation of the data presented in this study, the phylogenetic tree (Fig. 1a), PopPUNK phylogroup designations and associated metadata components have been uploaded to the interactive web interface Microreact[66] (https://microreact.org/project/5DEFpeck4). The PopPUNK database for assigning new genomes is available at https://doi.org/10.6084/m9.figshare.6931439.v1.

Code availability

The script for assessing antigenic variation from genome assemblies, as used in this study, is available at https://github.com/shimbalama/screen_assembly.

References

60. Seale, A. C. et al. Invasive group A *Streptococcus* infection among children, rural Kenya. *Emerg. Infect. Dis.* **22**, 224–232 (2016).
61. Athey, T. B. et al. Deriving group A *Streptococcus* typing information from short-read whole-genome sequencing data. *J. Clin. Microbiol.* **52**, 1871–1876 (2014).
62. Chalker, V. et al. Genome analysis following a national increase in scarlet fever in England 2014. *BMC Genom.* **18**, 224 (2017).
63. Kapatai, G., Coelho, J., Platt, S. & Chalker, V. J. Whole genome sequencing of group A *Streptococcus*: development and evaluation of an automated pipeline for emmgene typing. *PeerJ* **5**, e3226 (2017).
64. Ibrahim, J. et al. Genome analysis of *Streptococcus pyogenes* associated with pharyngitis and skin infections. *PLoS ONE* **11**, e0168177 (2016).
65. Page, A. J. et al. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Micro. Genom.* **2**, e000083 (2016).
66. Souvorov, A., Agarwala, R. & Lipman, D. J. SKESA: strategic *k*-mer extension for scrupulous assemblies. *Genome Biol.* **19**, 153 (2018).
67. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
68. Sumby, P. et al. Evolutionary origin and emergence of a highly successful clone of serotype M1 group A *Streptococcus* involved multiple horizontal gene transfer events. *J. Infect. Dis.* **192**, 771–782 (2005).
69. He, M. et al. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat. Genet.* **45**, 109–113 (2013).
70. Arndt, D. et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).
71. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
72. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
73. Argimon, S. et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Micro. Genom.* **2**, e000093 (2016).
74. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
75. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
76. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
77. Weyrich, L. S. et al. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* **544**, 357–361 (2017).
78. 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).
79. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

nature research

Corresponding author(s): NG-A48976-RZ
Mark R. Davies
Mark J. Walker

Last updated by author(s): 2nd April 2019

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees 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	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	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)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	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

To examine the genetic relationship of the 2,083 GAS genome sequences, we employed a single reference based mapping approach using sub-sampled Illumina fastqs at an estimated coverage of 75x. Published reference and draft genome datasets accessed from public databases were each shredded into an estimated 75x coverage of paired-end 100 bp reads using SAMtools wgsim. Sequence reads were mapped to the M1 GAS reference genome MGAS5005 (GenBank accession number CP000017) with BWA MEM (version 0.7.16) and read depth calculated with SAMtools (version 1.6) with a Phred quality score ≥20. Single nucleotide polymorphism (SNPs) with a Phred quality score ≥30 were identified in each isolate using SAMtools pileup with a minimum coverage of 10x. Core genes were defined as a minimum 80% of the MGAS5005 reference gene with a minimum 10x coverage.

1,306 MGAS5005 genes with 99% carriage in 2,083 genomes. A core SNP genome alignment of 171,273 SNPs was generated by concatenating the SNPs located within the 1,306 core genes, giving a total of 1,201,767 bp. SNPs residing within repeat regions (minimum length of 20 nucleotides) and mobile genetic elements are considered evolutionary confounders and were identified using PHASTER. SNPs within these regions were excised from the core alignment, reducing the length from 1,201,767 bp to 1,197,326 bp and the SNP count from 171,273 to 170,653. Therefore, a total of 170,653 SNPs were aligned for phylogenetic analysis of the 1,306 'core' genome.

FastGEAR was run on 1,306 individual gene alignments, comprising all 2,083 GAS strains included in the study. Default parameters were used with a minimum threshold of 4 bp applied for recombination length.

For intra-phylogroup recombination analyses, 36 most highly represented PopPunk phylogroups were chosen to investigate the influence of recombination (1,062 isolates). For each phylogroup, core genome alignments were performed using Snippy v4.3.5 (https://github.com/tseemann/snippy), against a reference strain within each phylogroup, maximum likelihood trees were inferred using IQtree v1.6.5, which were used as inputs for the recombination detection tool Gubbins v.2.3.4. Gubbins was run with maximum number of iterations of 20 with the minimum number of 5 SNPs to identify a recombination block, with a window size of 100 to 10,000 bp, with any taxa with more than 25% gaps filtered from the analysis. Recombinogenic blocks that overlapped with predicted mobile genetic elements (MGEs) in the reference genome were discarded. Phage regions were determined using PHASTER and integrative conjugative elements (ICE) were determined by manual inspection of reference genomes based on similarity of blast hits from known ICE. Recombination

versus vertically inherited mutation (r/m) ratios for each lineage were calculated as the average r/m including all isolates within the phylogroup. For the species values of r/m was determined by the average across all 36 phylogroups.

Maximum-likelihood trees were generated for the 416 and 1,306 core genome alignments using IQ-tree v1.6.5. The generalized time-reversible nucleotide substitution with gamma correction for site-specific rate variation was performed with 100 bootstrap random resampling's of the alignment data to support for maximum-likelihood bipartitions.

To define evolutionary related clusters (phylogroups) in the population we used PopPUNK (Population Partitioning Using Nucleotide K-mers), which has previously been shown to give high quality clusters in a subset of S. pyogenes isolates included in this study. We used k-mers between 15 and 29 nucleotides long in steps of two to calculate core and accessory distances between all pairs of isolates. We clustered these distances first with the default two-component Bayesian Gaussian Mixture Model, then used the 'refine fit' mode to move the boundary of this fit such that the network was highly transitive and sparse, obtaining a network score (ns) of 0.980.

Nucleotide divergence was derived by calculating the pairwise hamming distance from the 416 core genome alignment (266,960 bp). For pairwise hamming distance plots based on epdemiological markers, a reference genome was assigned for each marker based on the most representative distance within each type (minimum combined hamming distance) from the 416 core genome alignment.

The pangenome was defined using Roary v3.11.2 without splitting paralogs and with clustering at 80%. Accessory genome was defined as the pan less the core, totalling 3,672 genes. Identification of prophage CDS within each of the 2,083 genomes was performed using PHASTER. Clustering with CD-HIT-EST at ≤90% nucleotide homology resulted in 1,438 gene clusters. 584 core genes and 1,567 accessory genes hit these phage regions with blastn v2.3.0+ with a 90% nucleotide cut-off over 90% of the gene length. These data were then processed to generate a binary gene content matrix in which the presence of a gene is defined as >90% coverage to a corresponding phage gene cluster.

The presence of vaccine antigen genes was determined by BlastN analysis of 2,083 genome assemblies based on a 70% nucleotide cut-off over 70% of the gene length. Nucleotide sequences of whole and N-terminal regions of the M-protein were extracted using publicly available databases to account for known higher levels of allelic variation. This data was then converted into a binary gene content matrix in which gene presence was defined as >70% homology across a minimum 70% of the query gene length. Allelic variation was examined by plotting tBlastN (or BlastN for group A carbohydrate genes) scores relevant to the query reference sequence.

Protein sequences of streptolysin O and C5a peptidase were chosen for further analyses as well characterised crystal structures exist for each of these GAS antigens. A protein alignment corresponding to the published crystalised structures of streptolysin O (amino acid residues 103 – 571, Protein Data Bank [PDB] accession number 4HSC) and C5a peptidase (amino acid residues 97 - 1032, PDB accession number 3EIF) was generated. Using this data, we derived the consensus amino acid sequence for each protein as defined by the most common amino acid identified within the global GAS genome database and modelled the consensus against the mature crystal structures. Mutational sensitivity and structural integrity analyses was performed using Phyre2 that incorporates the SuSPect platform.

We investigated molecular signatures of selective constraints in all non-recombinogenic core genes (n = 416) by fitting a codon model to each of the individual genes and estimating the ratio of synonymous to nonsynonymous substitutions, dN/dS (also known as ω). Recombinogenic core genes (n = 890), as identified by fastGEAR, were excluded from analyses as such evolutionary processes invalidate phylogenetic codon model fitting. For each gene alignment, ambiguous codon sites were first excluded, before fitting the M0 codon model in CODEML, part of the PAML v4.0 package. This model estimates a global dN/dS which allows for straight-forward comparison between genes. For the Streptolysin O and C5a peptidase protein coding genes we conducted more detailed analyses, by assessing selective constraints across codon sites. To do this we counted the number synonymous and nonsynonymous substitutions in each codon position, to obtain a similar quantity to the dN/dS value above75. Although this method does not explicitly use a codon model, it is scalable for the large number of samples used here. Despite the objective of this study being centered around global diversity, our database does contain sample bias in the context of clinical and geographical sampling, and the selection analyses should be interpreted carefully, as they may not represent current global selective trends.

To facilitate the expansion of studies within the highest disease burden regions, 30 isolates were completely sequenced using long-read sequencing technology. Long-read sequences were obtained using the Pacific Biosciences RS II platform from a single molecule real-time (SMRT) cell. Briefly, genome sequences were assembled using the SMRTpipe version v2.1.0 using the Hierarchical Genome Assembly Process (HGAP.2) and Quiver for post-assembly consensus validation. Secondary validation of the assemblies was performed using the Canu assembler. To correct long-read sequence errors, primarily around homopolymeric regions, Illumina short read sequences from each of the 30 genomes were mapped using BWA MEM v0.7.16. Single contigs were achieved for all genomes and associated plasmids where present, with an average coverage depth of 80x. Genomes were annotated using the same pipeline as for the Illumina draft genomes with putative prophage regions defined using the PHASTER server.

To identify genomic signatures within the global GAS population overrepresented with severe GAS infection ('invasive') we ran pyseer on 1,944 samples (1,048 defined as invasive) using the linear mixed model. A total of 87M k-mers between 9 and 100 bases long were counted using fsm-lite. We only tested common k-mers, those with a minor allele frequency >1% (of which 18M were counted in our dataset). We created a kinship matrix from our recombination-free core phylogenetic tree of 2,083 genomes (416 genes). The country of isolation was used as a covariate in pyseer's model to account for geographical signal as defined previously. All k-mers were mapped to the MGAS5005 GAS reference genome using bwa and visualised with R. We used a Bonferroni correction to adjust the significance threshold passed the number of unique patterns tested, which gave 9.4x10-7 for a 0.05 family-wise error rate. 184 k-mers were significantly associated with severe infection.

Mutational sensitivity and structural integrity analyses was performed using Phyre2 that incorporates the SuSPect platform.

Gene predictions and annotations were generated using PROKKA and streptococcal RefSeq specific databases.

Allelic variation was examined by plotting BlastP (or BlastN for group A carbohydrate genes) scores relevant to the query reference sequence.

Data analysis

To facilitate future studies assessing vaccine antigen carriage and sequence variation within GAS genome sequences, we have generated

Data analysis	a bioinformatics pipeline for assessing antigenic variation from genome assemblies. This script, as used in this study, is available at https://github.com/shimbalama/screen_assembly and requires a query sequence (such as avaccine antigen) and will run BlastN at a user defined cut-off generating numerous outputs and plots as represented in this study.

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Illumina sequence reads and draft genome assemblies were deposited into the European Nucleotide Archive under the accession numbers specified in Supplementary Table 2. Genbank accession numbers for the 30 new GAS reference genomes are provided in Supplementary Table 5. To facilitate community accessibility and interrogation of the data presented in this study, the phylogenetic (Fig 1a), PopPUNK phylogroup designations, and associated metadata components have been uploaded to the interactive web interface Microreact (https://microreact.org/project/5DEFpeck4). The PopPUNK database for assigning new genomes is available at https://doi.org/10.6084/m9.figshare.6931439.v1.

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

Life sciences study design

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

Sample size	The global collection of 2,083 Streptococcus pyogenes isolates examined in this study was sequenced as follows, and included short read genome sequence data from population-based studies that we have generated within Kenya and Fiji, and other disease specific population-based studies of invasive GAS from Canada, USA and the United Kingdom that was available as of 1st July 2018. We selected a small subset of isolates from published microevolution (outbreak) studies to avoid biasing the collection on single genetically related lineages. sixty-eight GAS reference genomes and publicly available draft genomes from Lebanon were also included. To increase genomic representation from regions endemic for GAS infection and other under-sampled geographical regions, we collected a further 271 isolates from Australia, 279 isolates from New Zealand, 50 isolates from Brazil, 45 isolates from India and 7 isolates from Belgium. The rationale underpinning isolate selection was difference in epidemiological markers (emm type), anatomical site of isolation (skin, throat, blood) and clinical presentation, all key factors in GAS vaccine design. Metadata pertaining to the database of isolates are provided in Supplementary Table 2.
Data exclusions	N/A
Replication	N/A
Randomization	N/A
Blinding	N/A

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	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging