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

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

[1]Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia

[2]The Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom.

[3]School of Chemistry and Molecular Biosciences and Australian Infectious Diseases Research Centre, The University of Queensland, Brisbane, Queensland, Australia.

[4]GSK Vaccines Institute for Global Health, Siena, Italy

[5]Doherty Department, University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia

[6]Department of Microbiology, New York University School of Medicine, New York, USA.

[7]Menzies School of Health Research, Darwin, NT, Australia.

[8]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Victoria, Australia

[9]Molecular Bacteriology Laboratory, Université libre de Bruxelles, Brussels, Belgium; Department of Pediatrics, Academic Children Hospital Queen Fabiola, Université libre de Bruxelles, Brussels, Belgium

[10]Murdoch Childrens Research Institute, Melbourne, Victoria, Australia.

[11]School of Medicine, University of St Andrews, St Andrews, UK

[12]Wellcome Trust Research Centre, Kilifi, Kenya

[13]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK

[14]Laboratory of Bacteriology, Epidemiology Laboratory and Disease Control Division, Laboratório Central do Estado do Paraná, Curitiba, PR, Brazil and Department of Medicine, Universidade Positivo, Curitiba, PR, Brazil

[15]Infection and Immunity Program, Biomedicine Discovery Institute and Department of Microbiology, Monash University, Clayton, Australia

[16]Helmholtz Centre for Infection Research, Braunschweig, Germany

[17]Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand

[18]Telethon Kids Institute, University of Western Australia and Perth Children's Hospital, Perth, Western Australia, Australia

[19]Microbiological Diagnostic Unit Public Health Laboratory, Department of Microbiology and Immunology, Doherty Institute, The University of Melbourne, Melbourne, Victoria, Australia

[20]Victorian Infectious Diseases Service, Royal Melbourne Hospital, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia

[21]Department of Medicine, University of Cambridge, Cambridge, UK

[*]**For correspondence:** Professor Mark J. Walker, School of Chemistry and Molecular Biosciences, The University of Queensland, Cooper Road, St. Lucia, QLD, 4072, Australia.

Tel: 0061-7-3346 1623; Fax: 0061-7-3365 4273; E-mail: mark.walker@uq.edu.au; Doctor Mark Davies, Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, 3000, Australia. Tel: 0061-3-9035 6519; E-mail: mark.davies1@unimelb.edu.au

Key words: *Streptococcus pyogenes*, group A *Streptococcus*, vaccine, genomics, epidemiology, bacterial GWAS, molecular adaptation

Manuscript word count: (xxxx words)

Abstract word count: (197 words)

Group A *Streptococcus* (GAS; *Streptococcus pyogenes*) is a bacterial pathogen for which a vaccine is not available[1,2]. Employing the advantages of high-throughput DNA sequencing technology to vaccine design, we have analysed 2,083 GAS genomes from isolates causing significant morbidity and mortality in both developing and high-income countries. The global GAS population structure reveals extensive genomic heterogeneity 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. We report the extent of natural genetic diversity across 150 GAS molecular *emm* types[3], 484 multi-locus sequence types[4] and 39 M-protein clusters[5]. To determine vaccine candidate coverage, we investigated all previously described GAS antigens[2,6] 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. Mapping global antigenic heterogeneity onto antigen protein structure provides a new approach for the identification of conserved epitopes on the surface of vaccine antigens.

This technological platform for vaccine coverage determination is equally applicable to prospective GAS antigens identified in future studies.

GAS causes >700 million cases per year of superficial diseases such as pharyngitis and impetigo, and >600,000 cases per year of serious invasive infection. Immune sequelae such as acute rheumatic fever (ARF) and acute post-streptococcal glomerulonephritis each account for >400,000 cases per year[1,2]. As a consequence of ARF, >30 million people live with rheumatic heart disease, involving mitral and/or aortic regurgitation[7]. GAS ranks within the top 10 infectious disease causes of human mortality worldwide[1]. Despite over 100 years of research, a commercial vaccine has not been developed[2]. Obstacles that have hindered 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[2,6]. In 1978 the US Food and Drug Administration imposed a moratorium on human GAS vaccine trials due to concerns surrounding the potential of vaccine antigens to trigger autoimmunity. The US National Institute of Allergy and Infectious Diseases convened an expert workshop in 2004, which led to the lifting of the ban, but noted the possible involvement of M protein and group A carbohydrate antigens in autoimmunity[8]. A limited number of phase 1 clinical trials have since been conducted, focused primarily on multivalent N-terminal M protein vaccine candidates[9,10]. Other candidate GAS vaccine antigens that have demonstrated efficacy in animal models include the J8 peptide incorporated in the C-terminal repeats of M protein[11], and non-M protein candidate vaccine antigens. The group A carbohydrate[12,13] and multiple other surface or secreted proteins have been examined in preclinical vaccine studies (Supplementary Table 1)[2,6]. While a number of GAS antigens have been selected to avoid autoimmune concerns[14,15] or specifically engineered to remove potential autoimmune-involved epitopes[11,13], the capacity to investigate issues of serotype diversity, antigen carriage and

antigenic variation is impeded by the tremendous genetic diversity within the global GAS population[16]. To address this issue, we have developed a compendium of all GAS vaccine antigen sequences from 2,083 isolates employing high-throughput genomic technology.

RESULTS

GAS population genetics

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). Our sampling strategy targeted geographical regions where GAS infection is endemic and encompassed isolates from both asymptomatic carriage and various clinical disease states. We included population-based studies from published databases and a limited number of representative isolates from *emm*-type specific microevolution studies, to prevent substantial epidemiological bias in data interpretation. Extracting the classical GAS epidemiological and genotypic markers of differentiation from 2,083 genome assemblies, the database constitutes 150 *emm* types (347 *emm* sub-types), 39 known M-protein clusters and 484 multi-locus sequence types **(**MLSTs).

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 analysed each core gene separately for evidence of mosaicism using the homologous recombination detection tool fastGEAR[17]. Using this algorithm, we estimated 890 core genes as having a recombinatorial evolutionary history (Supplementary Fig. 1, Supplementary Table 3), leaving 416 non-recombinogenic core genes (Supplementary Table 4) encoded by 266,960 bp of sequence (~15% of a complete GAS genome). This 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 of a donor genome not being represented in the collection) and/or the limitation that larger blocks of

recombination encompassing multiple genes may be missed. A pseudo-core sequence alignment was generated using these 416 core GAS genes. After removal of repeat sequences that can 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 the lineage-specific radiations (Fig 1a). Comparative analyses of the core phylogenetic tree topologies prior (1,306 genes) and post (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 PopPUNK[18], we identified 299 distinct genetic clusters of evolutionarily related lineages, herein termed phylogroups (Figure 1a, Supplementary Fig. 4, 5a). This clustering approach is derived from core and accessory genetic distances between all 2,083 genomes using optimisation 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 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, Supplementary Fig 5a). Overlaying the geographical origin of the isolates suggests that over half 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 likely 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 a 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* sub-type and 129 (63%) were of a single multi-locus sequence type (Supplementary Fig. 6). Only 3 (1.5%) of the *emm* sequence types and 55 (27%) of the *emm* sub-types were unique to a single phylogroup of 2 or more isolates, inferring 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* sub-type, and MLST). Greater than 66% of *emm* types (84/128 multi-isolate representatives) and 32% of the *emm* sub-types (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* sub-types 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, 6 of the 7 MLST genes (*murI*, *xpt*, *gtr*, *gki*, *recP*, and *mutS*) were identified to have 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[19], and have been assigned the null allele xpt0 by MLST database curators. Collectively, these 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 unravelling the microevolution of dynamically evolving pathogenic 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 research into global and regional GAS population and evolutionary dynamics, 30 isolates representing geographically and genetically distinct samples were completely sequenced using the long-read PacBio platform. The average size of these new reference genomes was 1,810,671 bp (ranging from 1,701,466 bp to 1,950,606) with 5 strains containing circular plasmids ranging from 2,645 bp to 6,485 bp in size (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 evolutionary diverse genomes will act as an important reference tool for vaccine

developers, microbiologists, and molecular biologists 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 quanitifed 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[20], a tool previously shown to exhibit high concordance with other recombination detection approaches[21]. The average number of SNPs observed within the 36 phylogroups was 5,536 SNPs (range 191 to 24,899 SNPs) of which an average 20.5% of SNPs (range of 0.1 to 100%) were found to be vertically inherited within a phylogroup (Supplementary Table 6). Overall the ratio of recombination derived mutation versus vertically inherited mutation (r/m) was found to be 4.95 (median of 3.12), and noteably, is significantly greater than 1 (one-sample Wilcoxon test p-value of 7×10^{-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 found to be 58.94 (range 0 to 299) (Supplementary Table 6). Plotting the length of recombination blocks/fragments revealed that the majority of the events were small in length (< 5000bp) 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"). Removal of recombination events associated with putative mobile genetic elements had a limited effect on the total number of recombination events per phylogroup (Supplementary Fig. 9), suggesting that hertitable heterogeneity is largely mobile genetic element (MGE) independent. These data highlights 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[22] (average of 1,717 protein coding genes per genome). Plotting of unique protein counts per new genome added shows 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[23-26]. 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 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 on 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 with a

number of bacterial genetic factors and regulatory mutations[2,27]. To ascertain statistical support of gene content, gene polymorphisms or combinations thereof with clinical GAS invasiveness within this global genomic framework, we used the bacterial GWAS method of pyseer[73]. In this study, we defined invasiveness as those GAS isolated from a normally sterile site (blood, cerebrospinal fluid, bronchopulmonary aspirate) or severe cellulitis with positive GAS culture as invasive (n = 1,048); and those from clinical superfical infections such as 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[27]. Through this approach, we identified 184 hits provisionally associated with GAS invasiveness. Even though it was corrected for, at this significance level population structure confounding effects were apparent (which cause associations at the same p-value across the entire genome) (Supplementary Fig. 11). The top five k-mers which exceeded this threshold include a GAS virulence marker *isp* (immunogenic secreted protein)[28]; a LacI family transcriptional regulator; and a hypothetical open reading frame neighbouring the cysteine protease *speB* (Supplementary Table 7). Further studies are required to ascertain a link between genotype and an invasive phenotype. This analysis demonstrates 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/absence) and amino acid sequence variation of 29 proteinaceous GAS antigens, including 4 peptide fragments, was determined (Supplementary Table 1). The list of identified vaccine antigens analysed in this study have all been shown to convey protection in various murine models (reviewed by Henningham et al. 2012[6]) 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[13]. 2,017 GAS genomes (97%) shared all 12 protein coding genes with high DNA sequence conservation. Some genomes harboured 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. Eighteen antigens exhibited low levels (<2%) of amino-acid sequence variation (Supplementary Fig. 12). When plotted relative to overall carriage within 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, Supplementary Fig. 12). Furthermore, 11 of these 14 core genome vaccine antigens were identified to have signatures of homologous recombination in their evolutionary history (Supplementary Fig. 13). The highest level of sequence heterogeneity in pre-clinical vaccine antigens was observed within the M-protein. Collectively 33% of genomes had an N-terminal *emm* sub-type (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[31] and StreptInCor[32]; and the S2 peptide from the serine protease SpyCEP[33]. Given conformational and binding constraints afforded by peptide vaccine antigens relative to the complete protein antigens investigated above, carriage of these peptide antigens were assessed at an exact 100% match with the query peptide sequence within the 2,083 GAS genomes. 37% of the 2,083 isolates harboured 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% of isolates encode the S2 peptide from SpyCEP protein. Further interrogation of known J8 sequence variants within the multi-copy M- and M-like C-repeat sequences represented in the 2,083 genome assemblies identified carriage of J8.12 (79%) and J8.40 (76%) to be the most frequently encountered variants (Supplementary Fig. 14).

The identification of high homoplasy across core GAS antigens, including proposed vaccine antigens, emphasises that the evolution of GAS gene products is likely to be an ongoing process driven by recombination, genetic drift and diversifying selection. The characterisation of core gene products under different selection pressures may be used to identify putative vaccine antigen targets. Using the ratio of non-synonomous to synonomous codon subsititutions (d_N/d_S ratio) of each of the non-recombinogenic 416 genes, we identify that the average d_N/d_S ratio across the core GAS genome is greater than expected under a neutrality ratio of 1 (1.16), consitituing 49% of core genes (205 out of 416), suggestive of an overall positive selection across the GAS genome (Supplementary Table 4). Of the 3 'non-recombinogenic' core vaccine targets analysed in this study, the streptococcal hemoprotein receptor (Shr) had signatures of positive selection (d_N/d_S 1.22) while the hypothetical membrane associated protein Spy0762 and the 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

Structural analysis of antigens through protein crystallography yields insights regarding the identification of key functional amino acid residues and juxtaposition of surface peptide sequences. The ascertainment of antigenic variation within genome sequence databases allows such data to be overlaid onto protein structures, yielding important insight regarding potential

sites of structural plasticity or immunodominance, that in turn can be used to inform vaccine design through identification of invariant surface regions and/or structurally constrained domains or subdomains. Two crystal structures are publically 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]. Identification of polymorphism location and polymorphism frequency within the 2,083 GAS genomes for the Streptolysin O (Fig. 3a, Supplementary Table 8) and C5a peptidase (Fig. 3b, Supplementary Table 9) proteins were determined. Using this data, we derived the consensus amino acid sequence for each protein. We then modelled 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 97-1032; 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 hotspots (Fig. 3c). All hotspot polymorphisms were bimorphic in nature indicating restrictions in Streptolysin O plasticity (Supplementary Table 10). In comparison, we identified 20 sequence diversity hotspots within the mature C5a peptidase protein of which half were bimorphic (Fig. 3a, Supplementary Table 11), indicating 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 mutational sensitivity and structural integrity of these amino acids variants using Phyre2[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 to any amino acid would not impact protein structure or activity (Supplementary Tables 10 and 11). To further examine 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 and 16.5%

(170/1032) for C5a peptidase, indicating that these sites are undergoing positive selection (Supplementary Tables 8 and 9). Of these sites with signatures of positive selection, 40% (2/5) were diversity hotspots for Streptolysin O and 60% (12/20) for C5a peptidase. 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 subsequent emergence of vaccine escape clones, as has been observed in targeted *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, likely 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 pre-clinical GAS vaccine antigens fulfil the criteria for a global vaccine. It is tempting to speculate that multi-antigenic formulations would provide an ideal approach against a rapidly evolving pathogen as well as increasing global coverage. Indeed,

the incorporation of additional antigens to 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, likely to be reflective of 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 *E. 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 impact 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 *S. pyogenes* with other highly recombinogenic species such as *K. 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 less-developed economic regions of the world where the overwhelming burden of GAS disease resides. 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 bp to 6,485 bp, harbouring bacteriocin-like genetic markers that are suggested to play a role in inter-bacterial 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].

Modelling 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 discovered. Thus, informed selection of putative vaccine antigens for human trial evaluation will now be possible, allowing 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 SLO, SpyCEP, ADI, TF and C5a peptidase, found here to be highly conserved across geographic regions, protect against multiple GAS *emm* types in animal

models[14,50,51]. An approach similar to that used in this study would also be applicable to other pathogens that exhibit high levels of global strain diversity.

ACKNOWLEDGMENTS

This work was supported by the National Health and Medical Research Council (NHMRC) project and program grants for protein glycan interactions in infectious diseases and cellular microbiology; an Australian and New Zealand joint initiative, the Coalition to Accelerate New Vaccines Against *Streptococcus* (CANVAS); and The Wellcome Trust, UK. For part of this study, MRD was supported by a NHMRC postdoctoral training fellowship (635250) and AM was a GENDRIVAX fellow funded by European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n°251522. We acknowledge the assistance of 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 Prof. Debra Bessen). We dedicate this work to the memory of our friend and colleague Prof. Gusharan Singh Chhatwal.

AUTHOR CONTRIBUTIONS

MRD, GD and MJW conceived the project. MRD, AM, JAL, JALees, SD, PRS, DJP, MTGH, SYT, PMG, ACS, JAB, GSC, SDB, RAS, TL, JDF, NJM, JRC, ACS, JP, AS, DAW, BJC and MJW designed experiments. MRD, LM, JAL, JALees, SD, AM, RJT, KAW, SRH, TRH, HRF, OB, AJC, RSLAT, RB, PNS, NJM and DAW performed experimental protocols. MRD, LM, JAL, JALees, SD, AM, PRS, NJM, GD and MJW analyzed experimental results. MRD and MJW wrote the manuscript and all authors reviewed the manuscript.

COMPETING INTERESTS STATEMENT

AS is an employee of GlaxoSmithKline (GSK) that has a commercial interest in GAS vaccine development. The company had no influence over study design. The remaining authors report no competing commercial interests.

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**, 685-94 (2005).
2. Walker, M.J. *et al.* Disease manifestations and pathogenic mechanisms of group A *Streptococcus*. *Clin Microbiol Rev* **27**, 264-301 (2014).
3. Beall, B., Facklam, R. & Thompson, T. Sequencing emm-specific PCR products for routine and accurate typing of group A streptococci. *J Clin Microbiol* **34**, 953-8 (1996).
4. 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**, 2416-27 (2001).
5. 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**, 1325-38 (2014).
6. 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**, 207-42 (2013).
7. Watkins, D.A. *et al.* Global, Regional, and National Burden of Rheumatic Heart Disease, 1990-2015. *N Engl J Med* **377**, 713-722 (2017).
8. Bisno, A.L. *et al.* Prospects for a group A streptococcal vaccine: rationale, feasibility, and obstacles--report of a National Institute of Allergy and Infectious Diseases workshop. *Clin Infect Dis* **41**, 1150-6 (2005).
9. Kotloff, K.L. *et al.* Safety and immunogenicity of a recombinant multivalent group A streptococcal vaccine in healthy adults: phase 1 trial. *JAMA* **292**, 709-15 (2004).
10. McNeil, S.A. *et al.* Safety and immunogenicity of 26-valent group A *Streptococcus* vaccine in healthy adult volunteers. *Clin Infect Dis* **41**, 1114-22 (2005).
11. Brandt, E.R. *et al.* New multi-determinant strategy for a group A streptococcal vaccine designed for the Australian Aboriginal population. *Nat Med* **6**, 455-9 (2000).
12. Sabharwal, H. *et al.* Group A *Streptococcus* (GAS) carbohydrate as an immunogen for protection against GAS infection. *J Infect Dis* **193**, 129-35 (2006).
13. 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).
14. Henningham, A. *et al.* Conserved anchorless surface proteins as group A streptococcal vaccine candidates. *J Mol Med (Berl)* **90**, 1197-207 (2012).
15. 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. *Microb Pathog* **17**, 111-20 (1994).
16. 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**, 611-6 (2009).
17. Mostowy, R. *et al.* Efficient inference of recent and ancestral recombination within bacterial populations. *Mol Biol Evol* **34**, 1167-1182 (2017).
18. Lees, J.A. *et al.* Fast and flexible bacterial genomic epidemiology with PopPUNK. *bioRxiv*, 360917 (2018).
19. 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**(2017).

20. 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).
21. Marttinen, P. *et al.* Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res* **40**, e6 (2012).
22. Page, A.J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3 (2015).
23. 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 U S A* **101**, 11833-8 (2004).
24. 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 U S A* **111**, E1768-76 (2014).
25. Turner, C.E. *et al.* Emergence of a new highly successful acapsular group A *Streptococcus* clade of genotype *emm*89 in the United Kingdom. *MBio* **6**, e00622 (2015).
26. You, Y. *et al.* Scarlet fever epidemic in China caused by *Streptococcus pyogenes* serotype M12: Epidemiologic and molecular analysis. *EBioMedicine* (2018).
27. Lees, J.A. *et al.* Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun* **7**, 12797 (2016).
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**, 2548-55 (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**(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**, 8175-8 (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**, 1598-608 (2003).
32. Guilherme, L. *et al.* Towards a vaccine against rheumatic fever. *Clin Dev Immunol* **13**, 125-32 (2006).
33. Pandey, M. *et al.* Combinatorial synthetic peptide vaccine strategy protects against hypervirulent CovR/S mutant streptococci. *J Immunol* **196**, 3364-74 (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**, 785-92 (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**, 754-72 (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**, 845-58 (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**, 2692-701 (2014).
38. Croucher, N.J. *et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* **45**, 656-63 (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 Vaccine 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*. *bioRxiv*, 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**, 370-8 (2014).
48. Woodbury, R.L. *et al.* Plasmid-Borne *erm*(T) from invasive, macrolide-resistant *Streptococcus pyogenes* strains. *Antimicrob Agents Chemother* **52**, 1140-3 (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**, 819-35 (2009).
50. Bensi, G. *et al.* Multi high-throughput approach for highly selective identification of vaccine candidates: the group A *Streptococcus* case. *Mol Cell Proteomics* **11**, M111 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**, 2080-7 (1997).
52. Chaguza, C. *et al.* Recombination in *Streptococcus pneumoniae* lineages increase with carriage duration and size of the polysaccharide capsule. *MBio* **7**(2016).
53. Hanage, W.P. *et al.* Using multilocus sequence data to define the pneumococcus. *J Bacteriol* **187**, 6223-30 (2005).
54. Driebe, E.M. *et al.* Using whole genome analysis to examine recombination across diverse sequence types of *Staphylococcus aureus*. *PLoS One* **10**, e0130955 (2015).
55. 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**, 1008-15 (2000).
56. Coscolla, M. & Gonzalez-Candelas, F. Population structure and recombination in environmental isolates of *Legionella pneumophila*. *Environ Microbiol* **9**, 643-56 (2007).
57. Diancourt, L., Passet, V., Verhoef, J., Grimont, P.A. & Brisse, S. Multilocus sequence typing of *Klebsiella pneumoniae* nosocomial isolates. *J Clin Microbiol* **43**, 4178-82 (2005).
58. Wyres, K.L. *et al.* Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. *bioRxiv*, 414235 (2018).

FIGURE LEGENDS

Figure 1. Population structure and pangenome of 2,083 globally distributed GAS strains. (**a**) Maximum-likelihood phylogenetic tree of 30,738 SNPs generated from an alignment of 416 core genes. Branch colours indicate bootstrap support according to the legend. Distinct genetic lineages (n = 299) are highlighted in alternating colours (blue and grey) from the tips of the tree. Coloured 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). Colour coded around the outside of the phylogenetic tree is the country of isolation for each isolate. (**b**) Pangenome accumulation curve of 2,083 GAS genomes based on clustering of protein sequence at 70% homology.

Figure 2. Antigenic variation within vaccine targets from 2,083 GAS genomes. (**a**) Gene carriage (presence/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 colour coded according to pairwise BlastP similarity: Black (>98%); Blue (between 95 – 98%); Red (between 90 - 95%); Pink (80 - 90%); Yellow (70 - 80%); Grey (< 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 (180bp 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.

Figure 3. Global amino acid variation mapped onto the protein crystal structure of the mature GAS Streptolysin O[34] and C5a peptidase[35]. (**a**) Frequency of amino acid variations within 2,083 genomes. (**b**) Schematic of the Streptolysin O and C5a peptidase open reading frame

representing the location of amino acids within the mature enzymes (blue block). Model of the consensus sequence of the Streptolysin O (**c**) and C5a peptidase (**d**) mature enzymes. Plotted against the structure is the amino acid variation frequency within the 2,083 GAS genomes as represented in the colour gradient from 1% variable (blue) to 42% variable (red); invariant sites are coloured in light grey. Position of the top 5 most variable surface hotspots ("HS") are annotated (as defined in Supplementary Tables 10 and 11). Active sites for each enzyme are 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)***	**References**
Streptococcus pyogenes	4.95	17.2	This Study and Enright *et al*[4]
Streptococcus pneumoniae	6.36	23.1	Chaguza *et al*[52,53]
Staphylococcus aureus	0.6	0.1	Driebe et al,[54,55]
Legionella pneumophilla	47.8	0.9	David et al[46,56]
Klebsiella pneumoniae	4.75	0.3	Diancourt et al[57,58]

Footnotes

*Multi-locus Sequence Type (MLST) allele-derived r/m ratios as defined by Vos and Didelot[44].

ONLINE METHODS

Bacterial isolates

The global collection of 2,083 *Streptococcus pyogenes* isolates examined in this study included short read genome sequence data from population-based studies that we have generated within Kenya[59] and Fiji[27], and other disease specific population-based studies of invasive GAS from Canada[60], USA[19] and the United Kingdom[61,62] 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 publically available draft genomes from Lebanon[63] 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.

Genome sequencing and assembly

Genomic DNA was extracted and paired-end multiplex libraries were created and sequenced using the Illumina Hi-seq 2500 platform at the read-length between 75 to 125 bp (Wellcome Trust Sanger Institute, UK). Draft genome sequences were generated using an iterative Velvet-based assembly pipeline with secondary read mapping validation[64] or using SKESA v2.3.0[65] with default parameters. Gene predictions and annotations were generated using PROKKA[66] and streptococcal RefSeq specific databases[64]. 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 bp – 1,986,343) and an N50 of 252,789 bp (range 2,276 – 1,953,601 bp). On average, 1,711 coding sequences were identified per draft genome (range 1,495 – 1,976 coding sequences [CDS]). All draft genome assemblies are publically available through GenBank. Accession numbers are listed in Supplementary Table 2.

Sequence mapping

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)[67] 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. 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 of 20 nucleotides) and mobile genetic elements are considered evolutionary confounders and were identified as previously described[68] or identified using PHASTER[69]. 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 (Supplementary Fig. 3).

Recombination detection

To examine evidence of recombination within the core GAS genome, FastGEAR[17] 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 detection of lineages that have ancestral and recent recombinations between them. Default parameters were used with a minimum threshold of 4 bp applied for 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 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 (Supplementary Table 6), maximum likelihood trees were inferred using IQtree v1.6.5[70], which were used as inputs for the recombination detection tool Gubbins v.2.3.4[20]. 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[69] 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 inferited 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 (Table 1).

Phylogenetic analysis

Maximum-likelihood trees were generated for the 416 and 1,306 core genome alignments using IQ-tree v1.6.5[70]. 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. For figure generation, phylogenetic trees and associated metadata were collated using the web portal, Interactive Tree of Life[71].

Population genomics and cluster designation

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[18]. 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 (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 (n_s) 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 designation ("phylogroup") for each of 2,083 genomes have been added to Supplementary Table 2 and to the Microreact[79] interative web application (rJbD5w2nZ).

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 (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 core genome alignment.

Pangenome analysis

The pangenome was defined using Roary v3.11.2[22] 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[69]. Clustering with CD-HIT-EST[72] 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.

Vaccine antigen screening pipeline

To examine naturally occurring antigenic variation of proposed GAS vaccine targets within this genetically diverse GAS population, carriage of 29 vaccine antigens (Supplementary Table

1) and the group A carbohydrate biosynthesis loci was determined. The list of vaccine antigens screened have been shown to convey a significant level of protection in murine models[6], 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. N-terminal *emm* peptide and whole *emm* 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. To facilitate future studies assessing vaccine antigen carriage and sequence variation within GAS genome sequences, we have 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 and 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. 12 and Supplementary Tables 8 and 9). Futhermore, 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 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[34]) and C5a peptidase (amino acid residues 97 - 1032, PDB accession number 3EIF[35]) 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. Amino acid polymorphic sites were converted into a binary matix and presented as a percentage of 2,083 genomes in Fig. 3. Visualisation of polymorphic sites on the crystal structure was determined using Chimera (version 1.11.2)[73]. Mutational sensitivity and structural integrity analyses was performed using Phyre2[36] that 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 ratio of synonymous to nonsynonymous substitutions, d_N/d_S (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[74]. This model estimates a global d_N/d_S 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 d_N/d_S value above[75]. 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.

825

826 **Generation of 30 new GAS reference genomes**

827 The vast majority of publically available completely sequenced reference genomes are of *emm*-
828 types from North America and Europe and very few are of *emm*-types from high-disease
829 burden geographical regions. To facilitate the expansion of studies within the highest disease
830 burden regions, 30 isolates were completely sequenced using long-read sequencing
831 technology. Long-read sequences were obtained using the Pacific Biosciences RS II platform
832 from a single molecule real-time (SMRT) cell as described previously[76]. Briefly, genome
833 sequences were assembled using the SMRTpipe version v2.1.0 using the Hierarchical Genome
834 Assembly Process (HGAP.2) and Quiver for post-assembly consensus validation. Secondary
835 validation of the assemblies was performed using the Canu assembler[77]. To correct long-read
836 sequence errors, primarily around homopolymeric regions, Illumina short read sequences from
837 each of the 30 genomes were mapped using BWA MEM v0.7.16. Single contigs were achieved
838 for all genomes and associated plasmids where present, with an average coverage depth of 80x.
839 Genomes were annotated using the same pipeline as for the Illumina draft genomes[64] with
840 putative prophage regions defined using the PHASTER server[69].

841

842 **Genome-Wide Association of GAS Invasiveness**

843 To identify genomic signatures within the global GAS population overrepresented with severe
844 GAS infection ('invasive') we ran pyseer[78] on 1,944 samples (1,048 defined as invasive) using
845 the linear mixed model. A total of 87M k-mers between 9 and 100 bases long were counted
846 using fsm-lite. We only tested common k-mers, those with a minor allele frequency >1% (of
847 which 18M were counted in our dataset). We created a kinship matrix from our recombination-
848 free core phylogenetic tree of 2,083 genomes (416 genes, Figure 1a). The country of isolation
849 was used as a covariate in pyseer's model to account for geographical signal as defined

previously[27]. 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.

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[79] (identification number rJbD5w2nZ). The PopPUNK database for assigning new genomes is available at https://doi.org/10.6084/m9.figshare.6931439.v1.

SUPPLEMENTARY REFERENCES

59. Seale, A.C. *et al.* Invasive Group A Streptococcus Infection among Children, Rural Kenya. *Emerg Infect Dis* **22**, 224-32 (2016).
60. Athey, T.B. *et al.* Deriving group A Streptococcus typing information from short-read whole-genome sequencing data. *J Clin Microbiol* **52**, 1871-6 (2014).
61. Chalker, V. *et al.* Genome analysis following a national increase in Scarlet Fever in England 2014. *BMC Genomics* **18**, 224 (2017).
62. 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).
63. Ibrahim, J. *et al.* Genome Analysis of Streptococcus pyogenes Associated with Pharyngitis and Skin Infections. *PLoS One* **11**, e0168177 (2016).
64. Page, A.J. *et al.* Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb Genom* **2**, e000083 (2016).
65. Souvorov, A., Agarwala, R. & Lipman, D.J. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol* **19**, 153 (2018).
66. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-9 (2014).

67. 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**, 771-82 (2005).
68. He, M. *et al.* Emergence and global spread of epidemic healthcare-associated Clostridium difficile. *Nat Genet* **45**, 109-13 (2013).
69. Arndt, D. *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* **44**, W16-21 (2016).
70. 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**, 268-74 (2015).
71. 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**, W242-5 (2016).
72. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680-2 (2010).
73. Pettersen, E.F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-12 (2004).
74. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-91 (2007).
75. Weyrich, L.S. *et al.* Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* **544**, 357-361 (2017).
76. 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**, 84-7 (2015).
77. 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).
78. 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).
79. Argimon, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* **2**, e000093 (2016).
80. Fox, E.N., Waldman, R.H., Wittner, M.K., Mauceri, A.A. & Dorfman, A. Protective study with a group A streptococcal M protein vaccine. Infectivity challenge of human volunteers. *J Clin Invest* **52**, 1885-92 (1973).
81. Ma, C.Q. *et al.* Similar ability of FbaA with M protein to elicit protective immunity against group A streptococcus challenge in mice. *Cell Mol Immunol* **6**, 73-7 (2009).
82. Kawabata, S. *et al.* Systemic and mucosal immunizations with fibronectin-binding protein FBP54 induce protective immune responses against Streptococcus pyogenes challenge in mice. *Infect Immun* **69**, 924-30 (2001).
83. Guzman, C.A., Talay, S.R., Molinari, G., Medina, E. & Chhatwal, G.S. Protective immune response against Streptococcus pyogenes in mice after intranasal vaccination with the fibronectin-binding protein SfbI. *J Infect Dis* **179**, 901-6 (1999).
84. Courtney, H.S., Hasty, D.L. & Dale, J.B. Serum opacity factor (SOF) of Streptococcus pyogenes evokes antibodies that opsonize homologous and heterologous SOF-positive serotypes of group A streptococci. *Infect Immun* **71**, 5097-103 (2003).
85. Ulrich, R.G. Vaccine based on a ubiquitous cysteinyl protease and streptococcal pyrogenic exotoxin A protects against Streptococcus pyogenes sepsis and toxic shock. *J Immune Based Ther Vaccines* **6**, 8 (2008).

86. McCormick, J.K. *et al.* Development of streptococcal pyrogenic exotoxin C vaccine toxoids that are protective in the rabbit model of toxic shock syndrome. *J Immunol* **165**, 2306-12 (2000).
87. Kapur, V. *et al.* Vaccination with streptococcal extracellular cysteine protease (interleukin-1 beta convertase) protects mice against challenge with heterologous group A streptococci. *Microb Pathog* **16**, 443-50 (1994).
88. Zingaretti, C. *et al.* Streptococcus pyogenes SpyCEP: a chemokine-inactivating protease with unique structural and biochemical features. *FASEB J* **24**, 2839-48 (2010).
89. Liu, M., Zhu, H., Zhang, J. & Lei, B. Active and passive immunizations with the streptococcal esterase Sse protect mice against subcutaneous infection with group A streptococci. *Infect Immun* **75**, 3651-7 (2007).
90. Stalhammar-Carlemalm, M., Areschoug, T., Larsson, C. & Lindahl, G. The R28 protein of Streptococcus pyogenes is related to several group B streptococcal surface proteins, confers protective immunity and promotes binding to human epithelial cells. *Mol Microbiol* **33**, 208-19 (1999).
91. Huang, Y.S., Fisher, M., Nasrawi, Z. & Eichenbaum, Z. Defense from the Group A Streptococcus by active and passive vaccination with the streptococcal hemoprotein receptor. *J Infect Dis* **203**, 1595-601 (2011).
92. Okamoto, S., Tamura, Y., Terao, Y., Hamada, S. & Kawabata, S. Systemic immunization with streptococcal immunoglobulin-binding protein Sib 35 induces protective immunity against group: a Streptococcus challenge in mice. *Vaccine* **23**, 4852-9 (2005).
93. Dale, J.B., Chiang, E.Y., Liu, S., Courtney, H.S. & Hasty, D.L. New protective antigen of group A streptococci. *J Clin Invest* **103**, 1261-8 (1999).
94. Reglinski, M., Lynskey, N.N., Choi, Y.J., Edwards, R.J. & Sriskandan, S. Development of a multicomponent vaccine for Streptococcus pyogenes based on the antigenic targets of IVIG. *J Infect* **72**, 450-9 (2016).
95. Rivera-Hernandez, T. *et al.* Differing Efficacies of Lead Group A Streptococcal Vaccine Candidates and Full-Length M Protein in Cutaneous and Invasive Disease Models. *MBio* **7**(2016).
96. Bowen, A.C. *et al.* Whole genome sequencing reveals extensive community-level transmission of group A Streptococcus in remote communities. *Epidemiol Infect* **144**, 1991-8 (2016).
97. Bao, Y.J. *et al.* Genomic Characterization of a Pattern D Streptococcus pyogenes emm53 Isolate Reveals a Genetic Rationale for Invasive Skin Tropicity. *J Bacteriol* **198**, 1712-24 (2016).
98. Athey, T.B. *et al.* High Incidence of Invasive Group A Streptococcus Disease Caused by Strains of Uncommon emm Types in Thunder Bay, Ontario, Canada. *J Clin Microbiol* **54**, 83-92 (2016).
99. Fiebig, A. *et al.* Comparative genomics of Streptococcus pyogenes M1 isolates differing in virulence and propensity to cause systemic infection in mice. *Int J Med Microbiol* **305**, 532-43 (2015).
100. Teatero, S. *et al.* Canada-Wide Epidemic of emm74 Group A Streptococcus Invasive Disease. *Open Forum Infect Dis* **5**, ofy085 (2018).
101. Fittipaldi, N. *et al.* Full-genome dissection of an epidemic of severe invasive disease caused by a hypervirulent, recently emerged clone of group A Streptococcus. *Am J Pathol* **180**, 1522-34 (2012).

102. Soriano, N. *et al.* Complete Genome Sequence of Streptococcus pyogenes M/emm44 Strain STAB901, Isolated in a Clonal Outbreak in French Brittany. *Genome Announc* **2**(2014).
103. Soriano, N. *et al.* Closed Genome Sequence of Noninvasive Streptococcus pyogenes M/emm3 Strain STAB902. *Genome Announc* **2**(2014).
104. Soriano, N. *et al.* Full-Length Genome Sequence of Type M/emm83 Group A Streptococcus pyogenes Strain STAB1101, Isolated from Clustered Cases in Brittany. *Genome Announc* **3**(2015).
105. Meygret, A. *et al.* Genome Sequence of the Uncommon Streptococcus pyogenes M/emm66 Strain STAB13021, Isolated from Clonal Clustered Cases in French Brittany. *Genome Announc* **4**(2016).
106. Longo, M. *et al.* Complete Genome Sequence of Streptococcus pyogenes emm28 Clinical Isolate M28PF1, Responsible for a Puerperal Fever. *Genome Announc* **3**(2015).
107. de Andrade Barboza, S. *et al.* Complete Genome Sequence of Noninvasive Streptococcus pyogenes M/emm28 Strain STAB10015, Isolated from a Child with Perianal Dermatitis in French Brittany. *Genome Announc* **3**(2015).
108. Rochefort, A. *et al.* Full Sequencing and Genomic Analysis of Three emm75 Group A Streptococcus Strains Recovered in the Course of an Epidemiological Shift in French Brittany. *Genome Announc* **5**(2017).
109. Ben Zakour, N.L. *et al.* Transfer of scarlet fever-associated elements into the group A Streptococcus M1T1 clone. *Sci Rep* **5**, 15877 (2015).
110. Tse, H. *et al.* Molecular characterization of the 2011 Hong Kong scarlet fever outbreak. *J Infect Dis* **206**, 341-51 (2012).
111. Sagar, V. *et al.* Variability in the distribution of genes encoding virulence factors and putative extracellular proteins of Streptococcus pyogenes in India, a region with high streptococcal disease burden, and implication for development of a regional multisubunit vaccine. *Clin Vaccine Immunol* **19**, 1818-25 (2012).
112. Haggar, A. *et al.* Clinical and microbiologic characteristics of invasive Streptococcus pyogenes infections in north and south India. *J Clin Microbiol* **50**, 1626-31 (2012).
113. Hertzog, B.B. *et al.* A Sub-population of Group A Streptococcus Elicits a Population-wide Production of Bacteriocins to Establish Dominance in the Host. *Cell Host Microbe* **23**, 312-323 e6 (2018).
114. Beres, S.B. *et al.* Transcriptome Remodeling Contributes to Epidemic Disease Caused by the Human Pathogen Streptococcus pyogenes. *MBio* **7**(2016).
115. Nakagawa, I. *et al.* Genome sequence of an M3 strain of Streptococcus pyogenes reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res* **13**, 1042-55 (2003).
116. Miyoshi-Akiyama, T., Watanabe, S. & Kirikae, T. Complete genome sequence of Streptococcus pyogenes M1 476, isolated from a patient with streptococcal toxic shock syndrome. *J Bacteriol* **194**, 5466 (2012).
117. Yoshida, H. *et al.* Comparative Genomics of the Mucoid and Nonmucoid Strains of Streptococcus pyogenes, Isolated from the Same Patient with Streptococcal Meningitis. *Genome Announc* **3**(2015).
118. Watanabe, S. *et al.* Complete Genome Sequence of Streptococcus pyogenes Strain JMUB1235 Isolated from an Acute Phlegmonous Gastritis Patient. *Genome Announc* **4**(2016).
119. Minogue, T.D. *et al.* Complete Genome Assembly of Streptococcus pyogenes ATCC 19615, a Group A beta-Hemolytic Reference Strain. *Genome Announc* **2**(2014).

120. Port, G.C., Paluscio, E. & Caparon, M.G. Complete Genome Sequence of emm Type 14 Streptococcus pyogenes Strain HSC5. *Genome Announc* **1**(2013).
121. Port, G.C., Paluscio, E. & Caparon, M.G. Complete Genome Sequences of emm6 Streptococcus pyogenes JRS4 and Parental Strain D471. *Genome Announc* **3**(2015).
122. Bao, Y. *et al.* Unique genomic arrangements in an invasive serotype M23 strain of Streptococcus pyogenes identify genes that induce hypervirulence. *J Bacteriol* **196**, 4089-102 (2014).
123. McShan, W.M. *et al.* Genome sequence of a nephritogenic and highly transformable M49 strain of Streptococcus pyogenes. *J Bacteriol* **190**, 7773-85 (2008).
124. Suvorova, M.A. *et al.* Complete Genome Sequences of emm111 Type Streptococcus pyogenes Strain GUR, with Antitumor Activity, and Its Derivative Strain GURSA1 with an Inactivated emm Gene. *Genome Announc* **5**(2017).
125. Zheng, P.X. *et al.* Complete Genome Sequence of emm1 Streptococcus pyogenes A20, a Strain with an Intact Two-Component System, CovRS, Isolated from a Patient with Necrotizing Fasciitis. *Genome Announc* **1**(2013).
126. Beres, S.B. & Musser, J.M. Contribution of exogenous genetic elements to the group A Streptococcus metagenome. *PLoS One* **2**, e800 (2007).
127. Bessen, D.E. *et al.* Whole-genome association study on tissue tropism phenotypes in group A Streptococcus. *J Bacteriol* **193**, 6651-63 (2011).
128. Beres, S.B. *et al.* Genome sequence of a serotype M3 strain of group A Streptococcus: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc Natl Acad Sci U S A* **99**, 10078-83 (2002).
129. Sichtig, H. *et al.* FDA-ARGOS: A Public Quality-Controlled Genome Database Resource for Infectious Disease Sequencing Diagnostics and Regulatory Science Research. *bioRxiv*, 482059 (2018).
130. Ferretti, J.J. *et al.* Complete genome sequence of an M1 strain of Streptococcus pyogenes. *Proc Natl Acad Sci U S A* **98**, 4658-63 (2001).
131. Jacob, K.M., Spilker, T., LiPuma, J.J., Dawid, S.R. & Watson, M.E., Jr. Complete Genome Sequence of emm28 Type Streptococcus pyogenes MEW123, a Streptomycin-Resistant Derivative of a Clinical Throat Isolate Suitable for Investigation of Pathogenesis. *Genome Announc* **4**(2016).
132. Jacob, K.M., Spilker, T., LiPuma, J.J., Dawid, S.R. & Watson, M.E., Jr. Complete Genome Sequence of emm4 Streptococcus pyogenes MEW427, a Throat Isolate from a Child Meeting Clinical Criteria for Pediatric Autoimmune Neuropsychiatric Disorders Associated with Streptococcus (PANDAS). *Genome Announc* **4**(2016).
133. Smoot, J.C. *et al.* Genome sequence and comparative microarray analysis of serotype M18 group A Streptococcus strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci U S A* **99**, 4668-73 (2002).
134. Long, S.W., Kachroo, P., Musser, J.M. & Olsen, R.J. Whole-Genome Sequencing of a Human Clinical Isolate of emm28 Streptococcus pyogenes Causing Necrotizing Fasciitis Acquired Contemporaneously with Hurricane Harvey. *Genome Announc* **5**(2017).
135. Green, N.M. *et al.* Genome sequence of a serotype M28 strain of group a streptococcus: potential new insights into puerperal sepsis and bacterial disease specificity. *J Infect Dis* **192**, 760-70 (2005).
136. Holden, M.T. *et al.* Complete genome of acute rheumatic fever-associated serotype M5 Streptococcus pyogenes strain manfredo. *J Bacteriol* **189**, 1473-7 (2007).
137. Banks, D.J. *et al.* Progress toward characterization of the group A Streptococcus metagenome: complete genome sequence of a macrolide-resistant serotype M6 strain. *J Infect Dis* **190**, 727-38 (2004).

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Recombination analysis of 890 recombinogenic core GAS genes. Maximum-likelihood phylogenetic tree of core global GAS genome (416 'non-recombinogenic' genes; 30,738 SNPs) is shown on the left. Middle panel shows fastGEAR[17] outputs where each gene was analysed independently with recombinations coloured by donor lineage for each gene. Yellow is used to represent the most frequent lineage for each gene to optimise visualisation. Homologous recombination blocks are represented as gene fragments harbouring multiple 'colors'. The plot on top shows the number of homologous recombination events detected per gene.

Supplementary Figure 2. Support for the core genome phylogeny. Correlation between private SNPs (i.e. SNPs unique to each genome) and the length of the branch leading to that genome in the maximum-likelihood phylogenetic tree (416 genes), displayed as a proportion of the total length to the root. This correlation indicates that the raw SNP data supports the deep branching observed in the phylogeny and this is not an artefact of enforcing a tree-like structure onto the data.

Supplementary Figure 3. Tanglegram comparison of maximum-likelihood phylogenetic tree topologies based on the 1,306 gene core genome (left: 170,653 SNPs) and the 416 gene core genome (right: 30,738 SNPs) after the removal of 'recombinogenic' genes. Colours refer to 299 different PopPUNK phylogroupings. While the terminal clustering of the trees has not changed after removal of 'recombinogenic' genes, recombination has distorted the ancestral evolutionary signal within the core GAS genome.

Supplementary Figure 4. PopPUNK model fitting and refined network clustering of 2,083 GAS genomes. (a) Scatter plot showing core (π) and accessory (a) distance between all pairs of isolates with density contours displayed. (b) Same scatter plot as (a) after network score fit refinement for within cluster boundary. Blue points are distances comparing genomes within the same phylogroup, turquoise points are distances comparing genomes in different phylogroups. The red dashed line separates these assignments, grey dashed lines are the fit using core and accessory distances only. These data suggest low within-strain recombination (dense cluster of points near the origin of the graph), but high between-strain recombination (single broad cluster of between-strain points). (c) Network assignment of refined 299 PopPUNK clusters where samples are nodes (coloured by assigned cluster) and edges are pairwise links judged to be within the same cluster using the refined fit shown in (b).

Supplementary Figure 5. Population structure of 2,083 GAS genomes based on (a) 299 phylogroups and (b) their clinical association. Maximum-likelihood phylogenetic tree of core global GAS genome (416 genes) as displayed in Fig. 1. Branch colours indicate bootstrap support according to the legend. Distinct genetic phylogroups (n = 299) are assigned a unique colour to aid in visual designation of clusters in panel (a) while for each phylogroup in panel (b) two alternate colours (blue and grey) are assigned (as in Fig. 1).

Supplementary Figure 6. Population structure of 2,083 GAS genomes and their association with primary GAS epidemiological markers. Maximum-likelihood phylogenetic tree of core global GAS genome (416 genes) as displayed in Fig. 1. Branch colours indicate bootstrap support according to the legend. Distinct genetic lineages (n = 299) are highlight in alternating colours (blue and grey) from the tips of the tree. Represented from inner to outer rings of the

epidemiological data are *emm*-type (n = 150), *emm* sub-type (n = 347), M-cluster (n = 39), and MLST (n = 484).

Supplementary Figure 7. Pairwise genetic (hamming) distance of the non-recombinogenic GAS core genome (416 genes) based on isolates being of the same *emm* type (**a**); *emm* sub-type (b); MLST (c); M-protein cluster (d) and core genome PopPUNK phylogroups (d). Only groups which contain multiple isolates are represented on the X-axis. For each group, one reference was chosen based on having the mimimum median SNP distance from all other samples in the group. Each dot indicates the genetic distance (number of nucleotide SNPs, Y-axis) of samples from this reference, blue (same phylogroup [evolutionary lineage] as the reference), red (different phylogroup to the reference).

Supplementary Figure 8. Box and whisker plot showing the ratio of recombination derived mutation versus vertically inherited mutation (r/m) for the 36 most sampled phylogroups. Overall the average r/m across these phylogroups was 4.95 (median of 3.12), and notably, is significantly greater than 1 (p-value = 7×10^{-7}), using a one-sample Wilcoxon test, with a test-value of 1.

Supplementary Figure 9. Size distribution of intra-phylogroup recombination lengths. Lengths of putative recombination blocks in a subset of the most highly sampled PopPUNK phylogroups (n = 36 [total of 1,062 genomes]). Recombination blocks were defined using the sliding window approach of Gubbins[20] based on intra-phylogroup mapping. Cumulative frequency of all recombination blocks within the 36 phylogroups including putative mobile genetic elements (MGE) (**a**) or excluding putative MGEs (**b**) within a 5000 base pair (bp) range. (**c**) Distribution of homologous recombination blocks (excluding MGE) within each of

the 36 phylogroups (uniquely coloured) plotted from a 500 base pair sliding window (up to 20,000 bp). Collectively, evolution of GAS phylogroups is linked to high rates of small (<5000 bp) homologous recombination blocks.

Supplementary Figure 10. Variation in the size and prophage content of the GAS accessory genome. Counts of accessory genes per genome are overlaid against the maximum-likelihood phylogenetic tree of core global GAS genome (416 genes) as displayed in Fig. 1. Branch colours indicate bootstrap support according to the legend. Distinct genetic lineages (n = 299) are highlight in alternating colours (blue and grey) from the tips of the tree. Red bars relate to the total count of phage-related genes based on PHASTER analysis of the draft genome assemblies and blue bars relate to 'other' genes. Accessory gene scale refers to the number of genes (in 100 gene increments).

Supplementary Figure 11. Manhattan plot of SNPs associated with invasive GAS infection. The significance (y-axis) of each SNPs association with severe infection against its relative position within the MGAS5005 genome (x-axis). The red line denotes a significance cutoff of $p < 9x10^{-7}$. Top five loci reaching significance (Supplementary Table 7) are annotated. Associations were investigated by pyseer[78] using k-mers with a minimum minor allele frequency of 1%.

Supplementary Figure 12. GAS vaccine antigen carriage (gene product) and sequence variation within the 2,083 genome database. Left vertical axis refers to the bar graph showing frequency of antigen carriage of 26 GAS vaccine antigens and the group A carbohydrate (GAC) operon (X-axis) within 2,083 GAS genomes and the right vertical axis refers to the box and whisker plot showing quartile range, median (red line) and minimum/maximum values of each

antigen as inferred by BlastP (as per Fig. 2b). Generally, genes that are 'core' have less sequence heterogeneity than genes that are variably carried (accessory genes).

Supplementary Figure 13. Recombination analysis of 11 conserved GAS vaccine antigens within the context of 2,083 GAS genomes. Maximum-likelihood phylogenetic tree of core GAS genome based on 416 non-recombinogenic genes of the 2,083 genomes is shown on the left. Middle panel shows fastGEAR[17] outputs per gene, with colours representing gene lineages (as per Supplementary Fig. 1). The plot on top shows the number of homologous recombination events detected per gene.

Supplementary Figure 14. Frequency of J8 alleles in the GAS 2,083 genome database.

Supplementary Table 1. Vaccine antigen candidates examined in this study^ and published status of vaccine development (shaded boxes).

ANTIGEN	Pre-clinical	Phase I	Phase II	Proof of concept	Phase III	Reference
M protein: N-terminal peptide (30-valent)	■	■	■			Dale et al 2011[30]
M protein: C-terminal peptide (J8)	■	■				Batzloff et al 2003[31]
M1 protein: (whole protein)	■	■				Fox et al 1973[80]
M protein: C-terminal peptide (StreptInCor)	■					Guilherme et al 2006[32]
Trigger factor (TF)#	■					Henningham et al 2012[14]
Group A carbohydrate (GAC)	■					Sabharwal et al 2006[12]
C5a peptidase (ScpA)+#	■					Ji et al 1997[51]
Fibronectin-binding protein A (FbaA)	■					Ma et al 2009[81]
Fibronectin-binding protein 54 (Fbp54)	■					Kawabata et al 2001[82]
Streptococcal fibronectin binding protein I (SfbI)	■					Guzman et al 1999[83]
Serum opacity factor (SfbII/SOF)	■					Courtney et al 2003[84]
Streptococcal pyrogenic exotoxin A (SpeA)	■					Ulrich et al 2008[85]
Streptococcal pyrogenic exotoxin C (SpeC)	■					McCormick et al 2000[86]
Cysteine protease (SpeB)	■					Kapur et al 1994[87]
Serine protease (SpyCEP)*#	■					Zingaretti C et al 2010[88]
Serine protease (SpyCEP): S2 peptide	■					Pandey et al 2016[33]
Adhesion and division protein (SpyAD)+*	■					Bensi et al 2012[50]
Streptolysin O (SLO)*#	■					Bensi et al 2012[50]
Serine esterase (Sse)	■					Liu et al 2007[89]
Arginine deiminase (ADI)#	■					Henningham et al 2012[14]

Rib-like cell wall protein (R28)			Stalhammar-Carlemalm et al 1999[90]
Streptococcal hemoprotein receptor (Shr)			Huang et al 2011[91]
Streptococcal immunoglobulin-binding protein 35 (Sib35)			Okamoto et al 2005[92]
Streptococcal protective antigen (Spa)			Dale et al 1999[93]
Oligopeptide-binding protein (OppA)+			Reglinski et al 2016[94]
Putative pullulanase (PulA)+			Reglinski et al 2016[94]
Nucleoside-binding protein (Spy0942)+			Reglinski et al 2016[94]
Hypothetical membrane associated protein (Spy0762) +			Reglinski et al 2016[94]
Cell surface protein (Spy0651) +			Reglinski et al 2016[94]

Footnotes:
^All query sequences were based on the M1 GAS strain MGAS5005 as a query reference sequence (if present). Otherwise, query sequences from original published reference were used.
*Components of the Novartis (GSK) combination vaccine (Bensi et al 2012)[50]
+Components of the Spy7 combination vaccine (Reglinski et al., 2016)[94]
#Components of the Combo#5 vaccine (Rivera-Hernandez et al., 2016)[95]

Supplementary Table 2. GAS strains used in this study.

See separate Excel file

Supplementary Table 3. List of 890 core GAS genes identified as having recombinogenic signatures as defined by fastGEAR[17].

See separate Excel file

Supplementary Table 4. List of 416 "non-recombinogenic" core GAS genes with MGAS5005 reference genome annotations.

See separate Excel file

Supplementary Table 5: Strain and genome characteristics of 30 new globally sampled GAS reference genomes.

Strain ID	Country of isolation	Site	*emm*-subtype	Other *emm*-subtype	M-cluster	MLST	genome size (bp)	CDS (no.)	plasmid size (bp)	Prophage (no.)	GenBank Accession
GAS13475	New Zealand	Throat	197.0	-	AC2	998	1797172	1800		3	Pending
NS178	Australia	Skin	54.1	166.2	D1	302	1742565	1708		1	Pending
20123V1I1	Fiji	Blood	100.0	167.0	D2	119	1839531	1864		3	Pending
31010V3S1	Fiji	Skin	123.0	205.0	D3	325	1768816	1708		10	Pending
NS5694	Australia	Skin	230.0	-	D4	205	1826832	1813		4	Pending
31165V2S1	Fiji	Skin	93.4	174.1, 156.0	D4	814	1701466	1642		10	Pending
NS5958	Australia	Skin	56.0	205.0	D4	115	1825427	1833		3	Pending
31041V2S1	Fiji	Skin	70.0	174.1	D4	10	1826467	1818		3	Pending
K23890	Kenya	Soft Tissue	97.1	-	D5	283	1812090	1774		1	Pending
Bra006	Brazil	Throat	68.2	-	E2	989	1747924	1691		1	Pending
30109V1T1	Fiji	Throat	92.0	-	E2	1026	1758778	1718	3453	1	Pending
14GA0958	New Zealand	Blood	90.5	-	E2	184	1764969	1703		8	Pending
NS365	Australia	Blood	58.0	236.1	E3	176	1888806	1902		4	Pending
31132V1S1	Fiji	Skin	25.0	159.0	E3	1032	1835714	1832		2	Pending
A1268	India	Blood	1.0	-	E3	28	1834762	1841		3	Pending
NS7124	Australia	Throat	124.0	-	E4	199	1790668	1759		1	Pending
NS5128	Australia	Throat	77.0	149.2	E4	588	1806314	1782		1	Pending
A995	India	Skin	22.8	-	E4	360	1950616	1960	3626	4	Pending
GAS02198	New Zealand	Throat	78.3	-	E1	1000	1806521	1767		1	Pending
31143V3S1	Fiji	Skin	89.14	236.2	E4	380	1806344	1797	3043	2	Pending
Bra010	Brazil	Throat	64.3	205.1	E5	1008	1779766	1769		2	Pending
NS20	Australia	Skin	75.1	170.0	E6	607	1887700	1870		2	Pending

K3534	Kenya	Blood	65.0	-	E6	716	1789855	1734		1	Pending
GAS11291	New Zealand	Throat	11.0	202.1	E6	547	1809631	1793		2	Pending
31034V1S1	Fiji	Skin	105.0	-	M105	954	1800116	1757	2645	1	Pending
NS4972	Australia	Skin	55.0	-	M55	100	1899479	1908		4	Pending
NS7259	Australia	Throat	NA	138.0	NA	612	1788166	1788		3	Pending
K17300	Kenya	Soft Tissue	stg866.1	166.1	NT	450	1816007	1786		1	Pending
14GA0287	New Zealand	Throat	74.0	156.0	M74	120	1861037	1873		5	Pending
31034V4S1	Fiji	Skin	57.0	166.1	M57	1025	1756622	1718	6485	1	Pending

Supplementary Table 6: Frequency, size (length) and relative rates of recombination within 36 PopPUNK phylogroups.

See separate Excel file

Supplementary Table 7: Top 5 k-mers associated with invasiveness as determined by pyseer[73].

Gene/Locus Tag	Product	k-mer Coordinates	Log10(p-value)
M5005_Spy1733	Hypothetical protein	1696410..1696509	7.987
Intergenic		1810451..1810550	7.410
M5005_Spy1061	LacI family transcriptional regulator	1032656..1032755	7.261
M5005_Spy0554 (*ezrA*)	Cell division regulation	545543..545642	7.257
M5005_Spy1723 (*isp*)	immunogenic secreted protein	1687402..1687501	7.146

Supplementary Table 8: Position of amino acid variants within the Streptolysin O protein (SLO) and the consensus sequence of the SLO mature protein (as plotted in Fig 3a and 3c).

See separate Excel file

Supplementary Table 9: Position of amino acid variants within the C5a peptidase (ScpA) protein and the consensus sequence of the ScpA mature protein (as plotted in Fig 3a and 3d).

See separate Excel file

Supplementary Table 10: Mutation sensitivity analysis of amino acid variants within the mature Streptolysin O protein.

	Amino acid position within Streptolysin O protein				
hotspot[&]	HS1	HS2	HS3	HS4	HS5
aa position	172	182	324	450	470
major aa[$]	R(1242)	N(1305)	E(1204)	T(1843)	Q(1260)
minor aa[$]	M(841)	D(778)	D(879)	S(240)	R(823)
mutational sensitivity (minor aa)[#]	M=6	D=2	D=1	S=2	R=2

Footnotes:
aa (amino acid) .
[&] Diversity hotspot as determined by a minor amino acid frequency of >10% in the mature enzymatic protein (amino acids 103-501). Relative location is plotted in Figure 3c.
[$] Number in brackets refers to the total number of 2,083 GAS genomes carrying the respective amino acid.
[#] Determined using the SuSPect[37] platform (ranked between 1-9; 1 representing "very low" and 9 as "very high" mutational sensitivity). Sensitivity being a measure of likely functional consequence of the observed amino acid mutation.

Supplementary Table 11: Mutation sensitivity analysis of amino acid variants within the mature C5a peptidase protein.

	Amino acid position within C5a peptidase protein											
hotspot	HS1	HS2	HS3	HS4	HS5	HS6	HS7	HS8	HS9	HS10	HS11	HS12
aa position	110	146	247	346	348	350	376	448	450	451	605	637
major aa[$]	Q(1745)	T(1634)	R(1921)	A(1367)	Q(1229)	D(1711)	M(1553)	D(1534)	P(1351)	Q(1381)	K(1709)	H(1948)
minor aa[$]	H(338)	A(360) S(89)	I(150) K(12)	D(622) E(93) V(1)	H(420) K(434)	A(346) G(19) N(7)	T(530)	E(549)	S(617) L(98) R(11) F(6)	K(692) P(10)	T(374)	L(131) Y(4)
mutational sensitivity (minor aa) [#]	H=2	A=2 S=2	I=1 K=3	D=3 D=2 V=2	H=2 K=2	A=2 G=3 N=3	T=3	E=1	S=5 L=5 R=6 F=6	K=2 P=3	T=1	L=1 Y=2

	Amino acid position continued							
hotspot[&]	HS13	HS14	HS15	HS16	HS17	HS18	HS19	HS20
aa position	665	669	671	679	697	942	959	999
major aa[$]	V(1878)	A(1963)	R(1804)	Q(1952)	T(1111)	T(1671)	V(1775)	A(1679)
minor aa[$]	I(200) A(5)	V(120)	Q(279)	P(131)	K(972)	A(412)	I(308)	G(404)
mutational sensitivity (minor aa) [#]	I=1 A=3	V=2	Q=1	P=1	K=2	A=2	I=1	G=1

Footnotes:
aa (amino acid)
[&] Diversity hotspot as determined by a minor amino acid frequency of >10% in the mature enzymatic protein (amino acids 97-1032). Relative location is plotted in Figure 3d
[$] Number in brackets refers to the total number of 2,083 genomes carrying the respective amino acid

[#] Determined using the SuSPect[37] platform (ranked between 1-9; 1 representing “very low” and 9 as “very high” mutational sensitivity). Sensitivity being a measure of likely functional consequence of the observed amino acid mutation.

Supplementary Table 12: Theoretical global coverage of combination vaccines based on the genome database presented in this study.

		Theoretical Vaccine Coverage						
Vaccine	**Vaccine antigens**	Europe	Oceania	North America	South America	Asia	East Africa	**TOTAL**
		(n = 242)	*(n = 906)*	*(n = 474)*	*(n = 51)*	*(n = 79)*	*(n = 328)*	***(n = 2080)***
3-component (GSK)[50]	SLO$^+$	98%	>99%	>99%	>99%	97%	98%	**>99%**
	SpyCEP$^+$	99%	>99%	99%	>99%	99%	>99%	**>99%**
	SpyAD$^+$	>99%	>99%	>99%	>99%	>99%	>99%	**>99%**
	any antigen	**>99%**	**>99%**	**>99%**	**>99%**	**>99%**	**>99%**	**>99%**
Spy7[94]	Spy0651$^+$	>99%	>99%	>99%	>99%	>99%	>99%	**>99%**
	Spy0762$^+$	>99%	>99%	>99%	>99%	>99%	>99%	**>99%**
	Spy0942$^+$	>99%	>99%	>99%	>99%	>99%	>99%	**>99%**
	PulA$^+$	>99%	>99%	>99%	>99%	99%	99%	**>99%**
	OppA$^+$	>99%	>99%	99%	>99%	99%	99%	**>99%**
	SpyAD$^+$	>99%	>99%	99%	>99%	99%	99%	**>99%**
	ScpA$^+$	90%	>99%	96%	>99%	99%	99%	**98%**
	any antigen	**>99%**	**>99%**	**>99%**	**>99%**	**>99%**	**>99%**	**>99%**
Combo #5[95]	TF$^+$	>99%	>99%	>99%	98%	>99%	>99%	**>99%**
	ScpA$^+$	90%	>99%	96%	>99%	99%	99%	**98%**
	SpyCEP$^+$	99%	>99%	99%	>99%	99%	>99%	**>99%**
	ADI$^+$	>99%	>99%	99%	>99%	99%	99%	**>99%**
	SLO$^+$	98%	>99%	>99%	>99%	97%	98%	**>99%**
	any antigen	**>99%**	**>99%**	**>99%**	**>99%**	**>99%**	**>99%**	**>99%**
StreptInCor [32]	B cell epitope$^@$	39%	25%	15%	12%	34%	14%	**22%**
	T cell epitope$^@$	8%	17%	3%	2%	4%	9%	**11%**
	common epitope$^@$	35%	16%	14%	10%	30%	11%	**18%**
	any epitope	**39%**	**26%**	**15%**	**12%**	**34%**	**17%**	**23%**
S2 - J8.0[33]	S2$^@$	>99%	>99%	>99%	>99%	>99%	>99%	**>99%**
	J8$^\$$	45%	41%	31%	31%	57%	25%	**37%**
	any epitope	**>99%**	**>99%**	**>99%**	**>99%**	**>99%**	**>99%**	**>99%**
30-valent [30]	30 *emm* families$^\&$	**71%**	**33%**	**75%**	**53%**	**73%**	**28%**	**48%**
30-valent with Mrp[40]	30 *emm* families$^\&$	71%	33%	75%	53%	73%	28%	**48%**
	MrpI$^@$	8%	9%	12%	4%	5%	4%	**8%**
	MrpII$^@$	6%	13%	6%	18%	6%	3%	**9%**
	MrpIII$^@$	5%	4%	7%	4%	9%	6%	**5%**
	any antigen	**77%**	**51%**	**83%**	**59%**	**83%**	**33%**	**60%**

Footnotes:
[+] Defined by BlastN as 70% homology over 70% length of the nucleotide sequence.
[@] Peptide sequence carriage is defined by BlastP at 95% homology over 95% of query length.
[$] Defined as clustering at 90% of the J8 allelic database (encompasses J8.0, J8.57 or J8.59) by CD-HIT EST.
[&] Defined at the *emm* family level (irrespective of *emm* sub-type).