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

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

[1]Department of Microbiology and Immunology, The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, 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, QLD, Australia.

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

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

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

[7]Murdoch Childrens Research Institute, Melbourne, VIC, Australia.

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

[9]Victorian Infectious Diseases Service, The Royal Melbourne Hospital, and the

University of Melbourne, The Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia

[10]Wellcome Trust Research Centre, Kilifi, Kenya

[11]Centre for Tropical Medicine & Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK

[12]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

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

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

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

[16]Microbiological Diagnostic Unit Public Health Laboratory, Department of Microbiology & Immunology, Doherty Institute, The University of Melbourne, Melbourne, Australia

[17]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 at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne, 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

Manuscript word count: (2,912 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 1,579 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 200 clinically associated genomic phylogroups across 18 geographical regions, highlighting challenges in designing vaccines of global utility. We report the extent of natural genetic diversity across 141 GAS molecular *emm* types[3], 399 multi-locus sequence types[4] and 37 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 (>98%) 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 1,579 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 1,579 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 1,579 genome assemblies, the database constitutes 141 *emm* types (259 *emm* sub-types), 37 M-protein clusters and 399 multi-locus sequence types (MLSTs).

To assess the genome-wide relationships within this global database, we identified the core genome of GAS to be 1,325 coding DNA sequences (CDS), based on an 80% nucleotide sequence coverage threshold and presence in >99% of the 1,579 genomes (Supplementary Table 3). To examine signatures of recombination within the core 1,325 genes, we analysed each core gene separately for evidence of mosaicism using the homologous recombination detection tool fastGEAR[17]. Using this algorithm, we estimated 841 core genes as having a recombinatorial evolutionary history (Supplementary Fig. 1), leaving 484 non-recombinogenic core genes (Supplementary Table 3) encoded by 309,723 bp of sequence (~17% of a complete GAS genome). This is likely to be an underrepresentation 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 484 core GAS genes. After removal of repeat sequences that can confound read mapping, a total of 33,917 single nucleotide polymorphisms (SNPs) were identified within a 308,108 bp pseudo-reference. Phylogenetic analysis of the 484 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 (1,325 genes) and post (484 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 a phylogenetic clustering approach (RAMI[18]) to the refined core 484 gene alignment, we identified 250 distinct genetic clusters of evolutionarily related lineages, herein termed phylogroups (Supplementary Fig. 4a). The median nucleotide divergence between phylogroups was 0.52% (range 0.29 – 0.62%), whereas genomes within the same phylogroup differed by a median divergence of 0.02% (range 0 – 0.13%). Of the 247 phylogroups, 178 phylogroups were represented by 2 or more isolates. Overlaying the geographical origin of the isolates suggests that over half these 178 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 extensive genetic drift within the human adapted host or independent adaptive selection. Furthermore, these lineages do not appear to be restricted by clinical association (Supplementary Fig. 4b). For example, 137 of the 178 phylogroups (77%) contain a clinically defined invasive GAS isolate, defined in this study as an isolate obtained from a normally sterile site. Examination of the distribution of the classic GAS molecular epidemiological markers relative to the 178 multi-isolate phylogroups, revealed that 144 (81%) carried a single *emm* sequence type, 121 (68%) carried a single *emm* sub-type and 70 (39%) were of a single multi-locus sequence type (Supplementary Fig. 5). Only 23 (13%) of the *emm* sequence types and 56 (31%) of the *emm* sub-types were unique to a single phylogroup of 2 or more strains, 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 (80/121 multi-isolate representatives) and 35% of the *emm* sub-types (57/176 multi-isolate representatives) exceeded the minimal median nucleotide divergence between phylogroups (0.29% which equates to ~900 SNPs within 484 core genes), showing that many *emm* types and *emm* sub-types do not share a close evolutionary history and in many cases represent different genetic lineages (Supplementary Fig. 6). Similarly, 17% of MLST (38/221 multi-isolate representatives) also exceeded the minimal median nucleotide divergence between phylogroups. Furthermore, 4 of the 7 MLST genes (*gki*, *gtr*, *mutS*, and *recP*) 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 Table 3 and 4). Collectively, these data suggest that *emm*-type and MLST may have limited capacity for assigning evolutionary relationships within a globally evolving population.

The identification of hundreds of distinct genetic lineages (250 phylogroups) represents a challenge to unravelling the microevolution of dynamically evolving pathogenic populations. Indeed, only 23 of the phylogroups identified in this study contain a complete GAS reference genome (n = 47). 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 to 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 9). Based on our estimated structure of the global GAS population, these reference genomes represent 29 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.

Analysis of the variable gene content (defined as genes present in less than 99% of the 1,579 genomes) identified 4,838 'accessory' genes when homologues were clustered at a conservative 70% amino acid identity (average of 308 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. 7), in-line with observations from GAS microevolutionary analyses[19-22]. 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 conservation or sequence variation within GAS proteins such as vaccine antigens.

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.*[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 1,579 GAS genome assemblies, 15 antigen genes were identified in >99% of isolates (Fig. 2a) at a 70% BlastN cut-off. The species defining marker and vaccine candidate group A carbohydrate is comprised of a 12 gene biosynthesis cluster[13]. 1,554 GAS genomes (98%) shared all 12 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[23].

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. 8). When plotted relative to overall carriage within 1,579 genomes, 14 of the 25 antigens were not only carried by >99% of the 1,579 genome sequences but also exhibited low levels of allelic variation (<2% sequence divergence) (Fig. 2b, Supplementary Fig. 8). Furthermore, 11 of these 14 core genome vaccine antigens were identified to have signatures of homologous recombination in their evolutionary history (Supplementary Fig. 9), emphasising that the evolution of 'core' GAS antigens is likely to be an ongoing process.

The highest level of sequence heterogeneity was observed with the M-protein. Collectively 28% of genomes had an N-terminal *emm* type represented within the 30-valent M-protein vaccine formulation[24] (Fig. 2a). We also examined the prevalence of other GAS peptide-based vaccine antigens, namely the C-terminal M-protein sequences of J8[25] and StreptInCor[26]; and the S2 peptide from the serine protease SpyCEP[27]. 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 1,579 GAS genomes. 37% of the 1,579 isolates harboured the J8.0 allele of the M-protein; 22% carry the conserved overlapping B and T cell epitope of the StreptInCor M-protein vaccine candidate; and 54% 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 1,579 genome assemblies identified carriage of J8.12 (90%) and J8.40 (80%) to be the most frequently encountered variants (Supplementary Fig. 10).

Antigenic heterogeneity within GAS vaccine antigens

Structural analysis of antigens through protein crystallography yields key 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[28] and C5a peptidase[29]. Identification of polymorphism location and polymorphism frequency within the 1,579 GAS genomes for the Streptolysin O (Fig. 3a, Supplementary Table 5) and C5a peptidase (Fig. 3a, Supplementary Table 6) 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)[28] and C5a peptidase (amino acids 97-1032; Fig. 3b, d)[29]. Further examination of amino acid heterogeneity present in at least 10% of the 1,579 genomes within the mature Streptolysin O protein, revealed 5 sequence diversity hotspots (Fig. 3c, Supplementary Table 7). All polymorphisms were bimorphic in nature indicating restrictions in Streptolysin O plasticity (Supplementary Table 7). In comparison, we identified 20 sequence diversity hotspots within the mature C5a peptidase protein of which half were bimorphic (Fig. 3a, Supplementary Table 8), 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[30] and the SuSPect platform[31]. 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 7 and 8). Such variation 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 drivers and biology 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*[32] and *Bordetella pertussis*[33] vaccination programs.

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 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 identified within the GAS genome[34]. 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[35]. 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[32,33].

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 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. 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); 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, 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 dedicate this work to the memory of our friend and colleague Prof. Gusharan Singh Chhatwal.

AUTHOR CONTRIBUTIONS

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

COMPETING INTERESTS STATEMENT

The authors report no competing 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**, 685-694 (2005).
2. Walker, M.J. et al. Disease manifestations and pathogenic mechanisms of Group A *Streptococcus*. *Clin Microbiol Rev* **27**, 264-301 (2014).
3. Beall, B., Facklam, R. & Thompson, T. Sequencing *emm*-specific PCR products for routine and accurate typing of group A streptococci. *J Clin Microbiol* **34**, 953-958 (1996).
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**, 2416-2427 (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**, 1325-1338 (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**, 207-242 (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**, 1150-1156 (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**, 709-715 (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**, 1114-1122 (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**, 455-459 (2000).

12. Sabharwal, H. et al. Group A *Streptococcus* (GAS) carbohydrate as an immunogen for protection against GAS infection. *J Infect Dis* **193**, 129-135 (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**, 1197-1207 (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**, 111-120 (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**, 611-616 (2009).

17. Mostowy, R. et al. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Mol Biol Evol* **34**, 1167-1182 (2017).

18. Pommier, T., Canback, B., Lundberg, P., Hagstrom, A. & Tunlid, A. RAMI: a tool for identification and characterization of phylogenetic clusters in microbial communities. *Bioinformatics* **25**, 736-742 (2009).

19. 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**, 11833-11838 (2004).

20. 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**, E1768-1776 (2014).

21. 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).

22. You, Y. et al. Scarlet Fever Epidemic in China Caused by *Streptococcus pyogenes* Serotype M12: Epidemiologic and Molecular Analysis. *EBioMedicine* (2018).

23. 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).

24. Dale, J.B., Penfound, T.A., Chiang, E.Y. & Walton, W.J. New 30-valent M protein-based vaccine evokes cross-opsonic antibodies against non-vaccine serotypes of group A streptococci. *Vaccine* **29**, 8175-8178 (2011).

25. Batzloff, M.R. et al. Protection against group A *Streptococcus* by immunization with J8-diphtheria toxoid: contribution of J8- and diphtheria toxoid-specific antibodies to protection. *J Infect Dis* **187**, 1598-1608 (2003).

26. Guilherme, L. et al. Towards a vaccine against rheumatic fever. *Clin Dev Immunol* **13**, 125-132 (2006).

27. Pandey, M. et al. Combinatorial Synthetic Peptide Vaccine Strategy Protects against Hypervirulent CovR/S Mutant Streptococci. *J Immunol* **196**, 3364-3374 (2016).

28. Feil, S.C., Ascher, D.B., Kuiper, M.J., Tweten, R.K. & Parker, M.W. Structural studies of *Streptococcus pyogenes* streptolysin O provide insights into the early steps of membrane penetration. *J Mol Biol* **426**, 785-792 (2014).

29. Kagawa, T.F. et al. Model for substrate interactions in C5a peptidase from *Streptococcus pyogenes*: A 1.9 A crystal structure of the active form of ScpA. *J Mol Biol* **386**, 754-772 (2009).

30. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. & Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858 (2015).

31. Yates, C.M., Filippis, I., Kelley, L.A. & Sternberg, M.J. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J Mol Biol* **426**, 2692-2701 (2014).

32. Croucher, N.J. et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* **45**, 656-663 (2013).

33. Bart, M.J. et al. Global population structure and evolution of *Bordetella pertussis* and their relationship with vaccination. *MBio* **5**, e01074 (2014).

34. Bergmann, R., Nerlich, A., Chhatwal, G.S. & Nitsche-Schmitz, D.P. Distribution of small native plasmids in *Streptococcus pyogenes* in India. *Int J Med Microbiol* **304**, 370-378 (2014).

35. Wescombe, P.A., Heng, N.C., Burton, J.P., Chilcott, C.N. & Tagg, J.R. Streptococcal bacteriocins and the case for *Streptococcus salivarius* as model oral probiotics. *Future Microbiol* **4**, 819-835 (2009).

FIGURE LEGENDS

Figure 1. Population structure and pangenome of 1,579 globally distributed GAS strains. (**a**) Maximum-likelihood phylogenetic tree of 33,917 SNPs generated from an alignment of 484

core genes. Branch colours indicate bootstrap support according to the legend. Distinct genetic lineages (n=250) 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 blue). Colour coded around the outside of the phylogenetic tree is the country of isolation for each isolate. (**b**) Pangenome accumulation curve of 1,579 GAS genomes based on clustering of protein sequence at 70% homology.

Figure 2. Antigenic variation within vaccine targets from 1,579 GAS genomes. (**a**) Gene carriage (presence/absence) of vaccine antigens. (**b**) Amino acid sequence variation within 25 protein antigens for each of the GAS1579 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[28] and C5a peptidase[29]. (**a**) Frequency of amino acid variations within 1,579 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 1,579 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 7 and 8). Active sites for each enzyme are indicated (cyan).

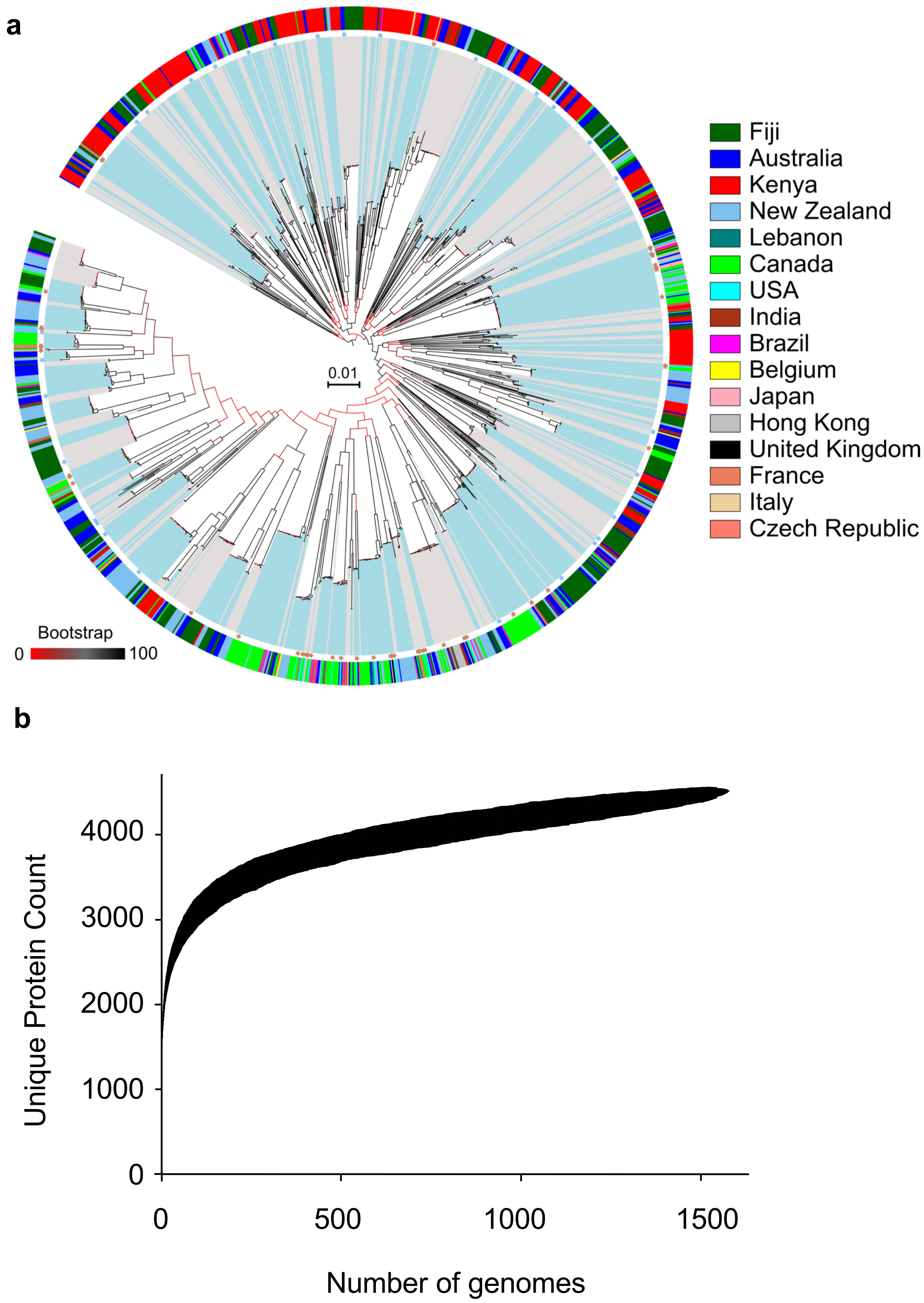
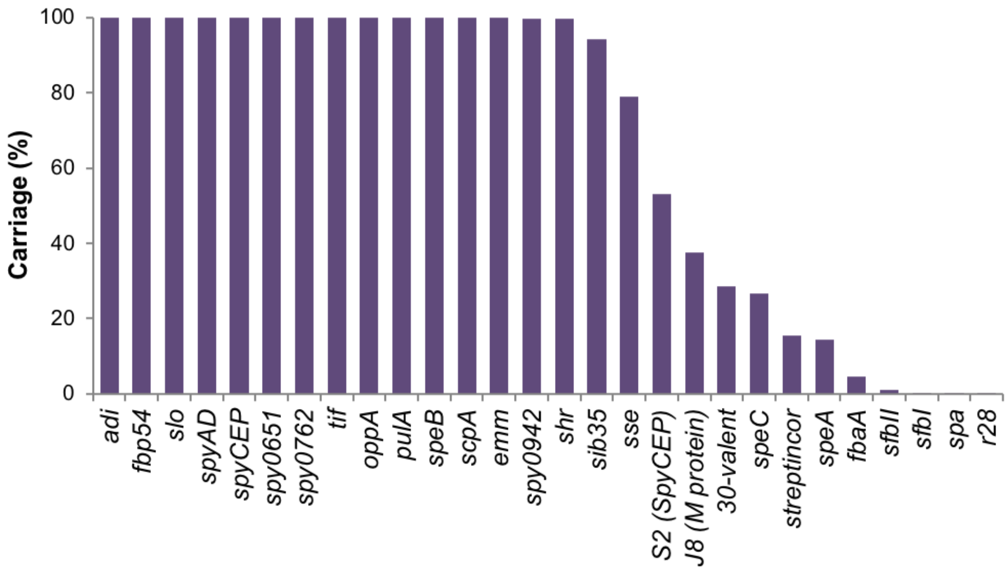

a
Fiji
Australia
Kenya
New Zealand
Lebanon
Canada
USA
India
Brazil
Belgium
Japan
Hong Kong
United Kingdom
France
Italy
Czech Republic
0.01
Bootstrap
0
100
b
Unique Protein Count
4000
3000
2000
1000
0
0
500
1000
1500
Number of genomes

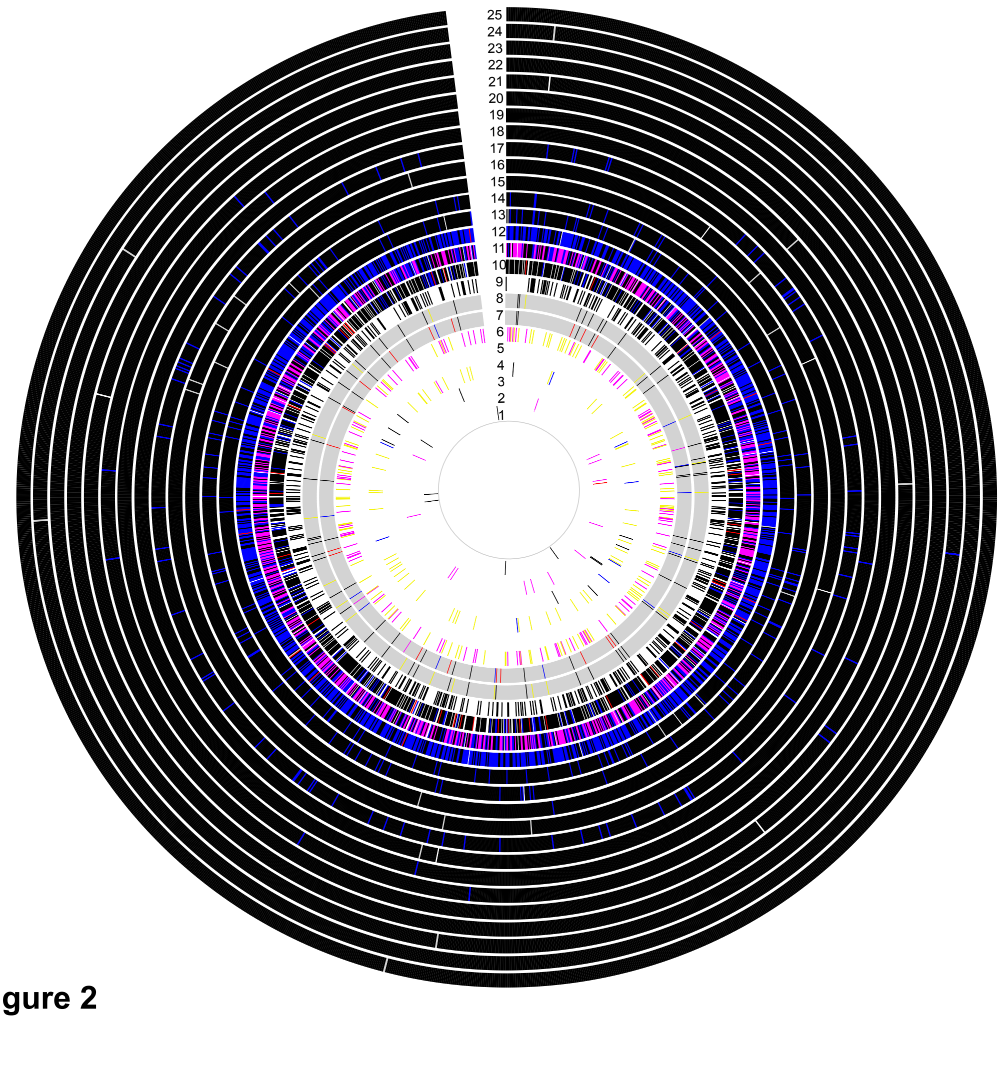
Figure 1

a



Carriage (%)
100
80
60
40
20
0
adi
fbp54
slo
spyAD
spyCEP
spy0651
spy0762
tif
oppA
pulA
speB
scpA
emm
spy0942
shr
sib35
sse
S2 (SpyCEP)
J8 (M protein)
30-valent
speC
streptincor
speA
fbaA
sfbII
sfbI
spa
r28

b



25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1

Figure 2

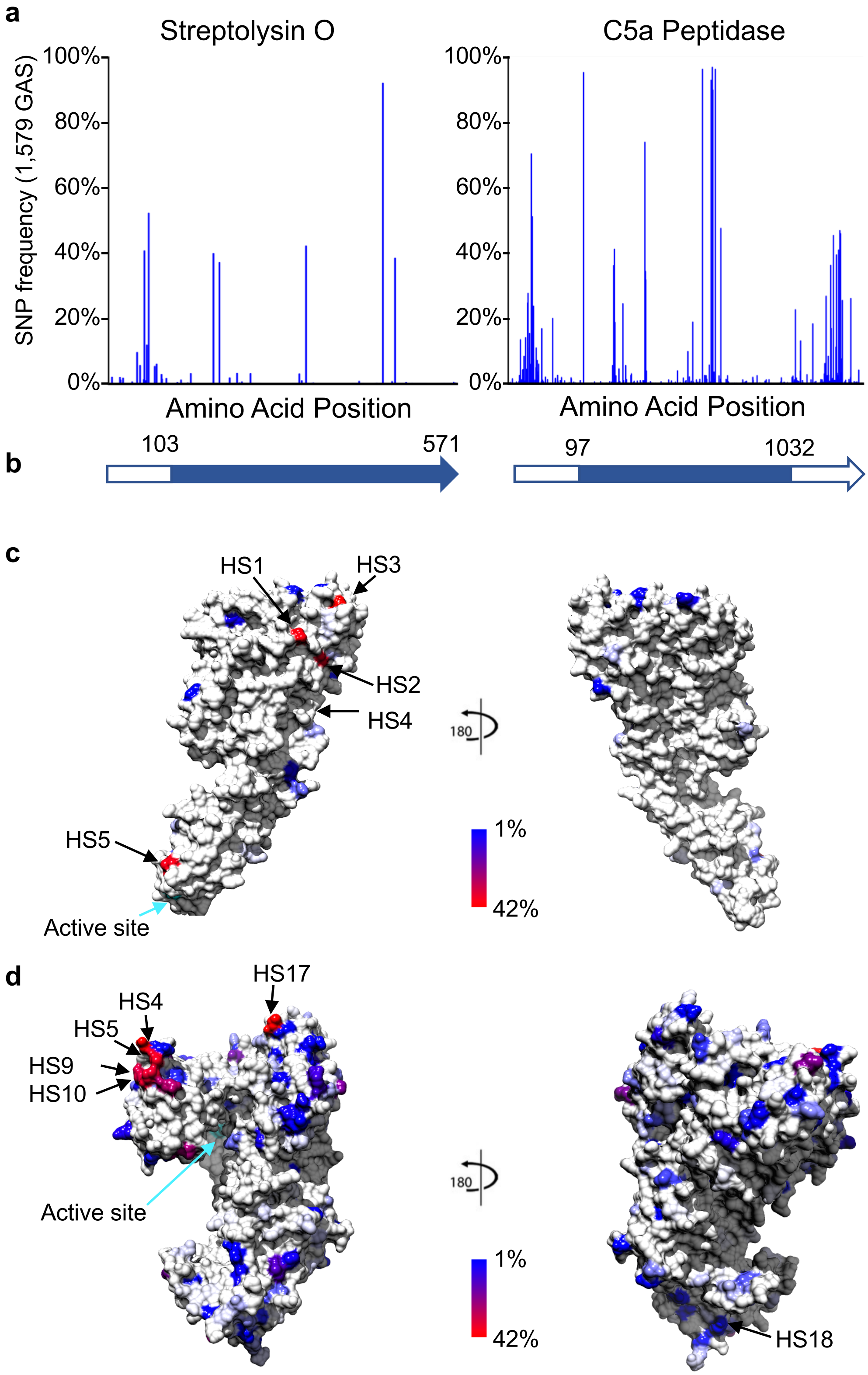

a
Streptolysin O
C5a Peptidase
SNP frequency (1,579 GAS)
100%
80%
60%
40%
20%
0%
Amino Acid Position
Amino Acid Position
b
103
571
97
1032
c
HS1
HS3
HS2
HS4
HS5
Active site
180
1%
42%
d
HS17
HS4
HS5
HS9
HS10
Active site
180
1%
42%
HS18

Figure 3