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: implication for the development of a regional multisubunit vaccine

Vivek Sagar[1]*, René Bergmann[1]*, Andreas Nerlich[1]*§, David J. McMillan[2], D. Patric Nitsche-Schmitz[1] and Gursharan S. Chhatwal[1#]

[1]Department of Medical Microbiology, Helmholtz Centre for Infection Research, Inhoffenstrasse 7, 38124 Braunschweig,

[2]Queensland Institute for Medical Research, Brisbane, Australia

*equal contribution

[#]correspondence: Gursharan S. Chhatwal, PhD

E-mail: gsc@helmholtz-hzi.de

Running Title: Genotyping of Indian GAS isolates

[§]present address: University of Veterinary Medicine Hannover Foundation, Institute of Microbiology, Centre for Infection Medicine, Bischofsholer Damm15, 30173 Hannover

Abstract

Streptococcus pyogenes causes a wide variety of human diseases and is a significant cause of morbidity and mortality. Attempts to develop a vaccine were hampered by genetic diversity of *S. pyogenes* across different regions of the world. This study aims to identify streptococcal antigens suitable for a region-specific vaccine in India. We used a two step approach, firstly performing epidemiological analysis to identify the conserved antigens among Indian isolates. The second step consisted of validating those identified antigens by serological analysis. The 201 streptococcal clinical isolates from India used in this study represented 69 different *emm* types with *emm*12 as the most prevalent. Virulence profiling of the North and South Indian *S. pyogenes* isolates with a custom-designed streptococcal virulence microarray identified 7 conserved putative vaccine candidates. Collagen-like surface protein (SCI), putative secreted 5 prime nucleotidase (PSNT) and C5a peptidase were found in 100 percent of the isolates, while R28, a putative surface antigen (PSA) and a hypothetical protein (HYP) were found in 90 percent of the isolates. A fibronectin binding protein, SfbI, was present in only 78 percent of the isolates. In order to validate the identified potential vaccine candidates, 185 serum samples obtained from patients with different clinical manifestations were tested for antibodies. Irrespective of clinical manifestations, serum samples showed high antibody titers to all proteins except for SCI and R28. Thus, the data indicates that PSNT, C5a peptidase, PSA, HYP and SfbI are promising candidates for a region-specific streptococcal vaccine for the different parts of India.

Introduction

Streptococcus pyogenes (group A streptococcus, GAS) is exclusively a human pathogen and the etiological agent of a wide variety of diseases that vary in clinical severity, while also being a significant cause of morbidity and mortality . These diseases include pharyngitis, impetigo, scarlet fever, post-streptococcal glomerulonephritis, invasive diseases, rheumatic fever (RF) and rheumatic heart disease (RHD) . While rheumatic fever and rheumatic heart disease are the greatest cause of mortality in developing nations, deaths in developed nations are mainly attributable to invasive diseases .

The differences in the prevalence and molecular epidemiology of GAS isolates reflect the differences in the importance of RF/RHD and invasive diseases in these populations. GAS carriage and infection are prevalent in many developing nations, with a large number of different *emm* types circulating at one time, with no one type being dominant. In contrast, a limited number of specific *emm* types are predominant in developed nations, and are often associated with specific clinical manifestations .

Despite the majority of GAS associated deaths occurring in developing nations, the majority of comparative genetic and genomic studies have focussed on isolates from developed nations . These studies report a high degree of genetic diversity between isolates with divergent *emm* types, as well as diversity within an *emm* type. In regions with high streptococcal disease burden, where GAS isolates are more likely to come into direct contact, the probability of lateral gene transfer (LGT) is increased. Greater inter- and intra-*emm*-type genetic diversity is likely to exist. Such diversity and potential changes in population structure are important considerations when designing vaccine candidates that should provide coverage against the entire GAS

population.

In this study, we used a two step approach in order to identify promising candidates for a region-specific vaccine. In the first step we conducted epidemiological analysis in order to identify antigens conserved in different parts of India, in the second step we purified these antigens and performed serological analysis using convalescent serum samples. We used virulence gene profiling to assess genetic diversity and population structure of GAS in India, a country where streptococcal disease burden is high . We found a high degree of genetic diversity between isolates of different *emm* types, but relatively conserved genotypes within an *emm* type. We assessed the distribution of seven genes conserved in Indian isolates encoding current vaccine targets in this population, and report on serological responses against each of these antigens.

Experimental procedures

Bacterial strains and human sera

Bacterial isolates and sera used in this study were collected as part of ASSIST program, funded by the European Commission, with three Indian and three European partners (http://www.helmholtz-hzi.de/en/research/research_projects/view/projekt/projekt/assist/). The samples were collected during 2007 - 2010 in two defined areas in Chandigarh (North India) and Vellore (South India), which are about 3000 km apart, have high streptococcal disease burden and different climatic conditions. Besides hospital patients, 3000 and 2400 school children were screened in Chandigarh and Vellore, respectively. Sixty-five isolates from Chandigarh and 136 from Vellore were included in this study. The isolates were also classified on the basis of recovery from the throat of asymptomatic carriers (n=44), or patients presenting with pharyngitis (n=20). Another 32 isolates were collected from the skin in these surveys. Thirty-four isolates were collected from patients presenting with invasive disease at clinics in Chandigarh and Vellore. The isolation site of 71 isolates was unknown. Serum samples were collected from individuals in North (n=110) and South India (n=75). Eighty-two of these samples were obtained from patients with RHD, 24 collected from patients with RF and 9 sourced from patients with invasive disease. Another 15 were obtained from patients presenting with symptoms of pharyngitis, 30 were asymptomatic patients positive for GAS and 25 sourced from healthy people from the survey areas with no current signs of streptococcal infection. Because of the possibility of subclinical exposure in the survey areas due to high disease burden, we included 10 serum samples collected from healthy people from Germany as controls. The serum samples and streptococcal isolates were from the same patients in case of invasive diseases. For

all other clinical manifestations the isolates and serum samples were from the defined survey areas, but not from the same patients.

DNA extraction and *emm* typing

Genomic DNA was isolated using zirconium beads in combination with the Qiagen DNeasy kit (Qiagen, Hilden, Germany). The *emm* type of individual strains was determined by amplification and sequencing the 5'-end region of the *emm* gene described by Beall and Facklam . The resulting nucleotide sequences were blasted against the *emm* gene database (http://www.cdc.gov/ncidod/biotech/strep/strepblast.htm) to determine the *emm* type of individual isolates.

Microarray, hybridization and data analysis

The microarray used is this study has been described in detail in a previous study . Sample preparation, array hybridization and data processing were performed as described previously . Briefly, isolated genomic DNA was digested with *Alu*I, labelled with biotin/streptavidin-Cy5 and fluorescence signals were quantified after hybridization using ImaGene software (BioDiscovery). A two component mixture model was fitted to the background-corrected and log2-transformed data by a maximum likelihood method. A discriminant function was used to represent the propensity of a gene for being present or absent in a particular isolate. Discriminant values were stored in a signal probability matrix and used to construct dendrograms using Bayesian agglomerative hierarchical clustering . All routines for statistical calculations were implemented in the R statistics package (www.r-project.org).

Cloning and expression of cell surface proteins

Out of 219 genes on the array, 7 genes were selected which showed 75% or more frequency in 201 strains and encode for cell wall proteins. These proteins, which included collagen like-surface protein (SCI), putative secreted 5'-nucleotidase (PSNT), C5a peptidase (C5a), R28 (R28), putative surface antigen (PSA) and a hypothetical transposase (HYP) were amplified (Table 1) and cloned into the pGEX-6P-1 vector (GE health care). Cloning of streptococcal fibronectin binding protein I (SfbI) was described by Talay *et al.* . The expression of the recombinant GST-fusion proteins in *E. coli* BL21(DE3) was induced by addition of IPTG to 0.5 mM. Cultures were harvested and lysed using a French Press (SLM instruments inc.). Cellular debris was removed by centrifugation and the presence of recombinant proteins in supernatants confirmed by SDS-PAGE.

Antibody responses to putative vaccine candidates

ELISA was performed using glutathione coated microtitre plates (Thermo Fisher Scientific) with *E. coli* lysates (0.25 µg/µl) containing GST-tagged recombinant proteins following the method of Sehr *et al*. . PBS and *E. coli* lysate containing GST alone served as controls. A peptide representing the IgG binding motif from FOG, a streptococcal protein with IgG-Fc fragment binding capacity was used as the normalization control. HRP conjugated goat antihuman IgG antibody and ABTS (2,2'-azino-bis(3-ethylbenzthiazoline-6-sulphonic acid) were used for color development that was measured at 405 nm. Absorbance values were subsequently normalized against values obtained from wells containing PBS, GST and the FOG IgG binding peptide. Differences in absorbance between groups were examined using one-way ANOVA. Because of the high number of serum samples and 8 proteins, the experiment was designed in such a way that maximum number of serum samples

could be covered by one experiment in order to avoid experiment to experiment variations. Therefore we chose to monitor the immunogenicity of proteins in all serum samples at the dilution of 1:200 and expressed the result as OD.

Results

emm sequence type distribution of GAS isolates from North and South India

The 201 isolates collected represent 69 different *emm* sequence types (Supplementary Table S1) and 83 different *emm* sequence subtypes (data not shown). The 20 most prevalent *emm* types (Figure 1) were *emm*12 (6.5%), *emm*11 (5.5%), *emm*49 (4.5%), *emm*28, *emm*80, st1389 (3.5% each), *emm*3 (3.0%), *emm*4, *emm*44, *emm*75 and *emm*112 (2.5% each). The types *emm*2, *emm*22, *emm*69, *emm*74, *emm*77, *emm*93, *emm*110, *emm*104 and *emm*108 accounted for 2% of all strains recovered. The 20 prevalent *emm* types represented 57.7% of all GAS isolated in both regions.

Of these twenty prevalent *emm* types, ten (*emm*4, *emm*11, *emm*12, *emm*44, *emm*49, *emm*74, *emm*80, *emm*110, *emm*112, and st1389) were found both in Northern and Southern India. *Emm*3, *emm*28, *emm*69, *emm*77, *emm*93, *emm*104, and *emm*108 were only found in Southern India, whereas *emm*2, *emm*22 *emm*75 were specific for Northern India. With 55 different *emm* sequence types the South Indian strains showed a greater heterogeneity as compared to 32 different *emm* sequence types found within the North Indian strains. This is reflected in the Simpsons Index of Diversity (*D*) for *emm* types in Southern India (*D*= 0.981; 95% Confidence interval, 0.976-0.986) compared to Northern India (*D*= 0.957; 95% Confidence interval, 0.936-0.977).

Overview of gene distribution in Indian Isolates

Of the 219 streptococcal virulence factors and extracellular surface proteins (VF/ECP) represented on the array, 91 (41.5%) were found in all isolates (Supplementary Table S2), and 150 (68.5%) were found in more than 80% of the isolates. When we analyzed the distribution of the VF/ECP genes with respect to geographic region, 102 genes (46.6%) and 100 genes (45.7%) were present in all North Indian and South Indian isolates, respectively. In order to identify region-specific virulence genes we statistically analyzed the distribution of VF/ECP genes in both regions. We identified one gene in the North Indian isolates and 15 genes in the South Indian isolates with a statistically significant difference in distribution (Fisher exact test, $P < 0.05$, Table 2), reflecting the higher *emm* type heterogeneity in South India. However, none of the 16 genes was found exclusively in one of the regions.

Genetic relationships amongst isolates

Emm typing is still largely accepted as sufficient for defining related strains in GAS research. However, epidemiological and molecular observations also demonstrate that LGT involving the *emm* gene occurs, suggesting that isolates of the same *emm* type that are temporally or geographically displaced may be genetically diverse, and be more closely related to isolates of a different *emm* type . To investigate the relationship between all isolates we used a Bayesian agglomerative hierarchical cluster algorithm (BHC) to construct a dendrogram of all isolates that were represented by ≥ 3 isolates per *emm* type (n = 154 isolates), based on the presence or absence of all genes represented on the array. Using this approach we found isolates of the same *emm* type to predominantly cluster together (Figure 2). However, there were instances where this did not occur, suggesting that LGT of the

emm gene may have occurred in these instances. We next analyzed the association of the clusters with regard to disease type. As shown in Fig 2, there was only one distinct sub-cluster solely composed of invasive isolates consisting of one *emm*4, one *emm*28 and three *emm*49 strains. Using BHC clustering we did not find any significant clustering of virulence factors with invasive and non-invasive isolates. However, when we analyzed the association of virulence factors with invasiveness using the Fischer´s Exact Test ($P<0.05$), we found 13 genes positively associated with invasiveness (Supplementary Table S3). Six genes were negatively associated with invasiveness (odds ratio < 1).

Serological response to vaccine antigens

We examined the distribution of the shortlisted seven GAS vaccine candidates in the India population, and found them to be highly conserved across the population (Table 3). Collagen-like surface protein (SCl), putative secreted 5'nucleotidase (PSNT) and C5a peptidase were found in 100% of isolates. R28, the putative surface antigen (PSA) and the hypothetical protein (HYP) were found in more than 90% of isolates. Streptococcal fibronectin-binding protein I (SfbI) was found in 78.5% of isolates.

Serum antibodies indicate that epitopes of a particular protein are visible to the host immune system, and these proteins are likely to be recognized by antibodies generated against corresponding vaccine candidates. Therefore, we next investigated serological responses to these seven identified GAS vaccine candidates. We expressed these proteins as fusion with GST and used them in ELISA with human sera (Figure 3).

Control healthy sera had very low immune response against all of the proteins. The majority of serum samples from India had elevated responses to PSNT, C5a peptidase and PSA, demonstrating the high incidence of GAS infection in these populations. PSA-specific IgG responses were statistically more frequent in the pharyngitis group than in healthy and RHD groups ($P < 0.05$). Lower antigen specific antibody responses were observed with HYP and SfbI. For these two proteins the highest responses were observed with sera collected from pharyngitis patients and patients with GAS throat colonization at the time of sera collection. For HYP, IgG responses in the pharyngitis group were statistically significantly higher than the healthy from non-survey and survey areas and RF groups ($P < 0.05$). The SfbI response for both the carrier group and pharyngitis group was statistically significantly higher than control and RHD groups ($P < 0.05$) (Supplementary Table S4). The majority of sera did not respond, or responded weakly to R28 and SCl. No differences in responses were observed in sera collected from North and South India. For 4/7 of the proteins used here the response in pharyngitis sera was significantly higher than non-survey area healthy sera (Supplementary Table S4) indicating the possibility that these proteins are most likely expressed in the early phase of infection.

Discussion

This study is the most comprehensive assessment of the genetic repertoire of GAS carried out in a country with high streptococcal disease burden. The twenty most common *emm* types accounted for 57.7% of the isolates collected, reinforcing the high degree of diversity in the population based on *emm* typing and is consistent with

results of other studies of GAS population structure carried out in developing nations . Within the Indian population, 91 genes were conserved in 100% of isolates. Using the same array, we previously reported that 129 and 125 genes were 100% conserved in GAS isolates from the Netherlands and United States respectively. The majority (80 genes) of conserved genes in the isolates from Europe and India were similar, indicating that they are part of the GAS core genome and encode proteins that are critical for virulence or confer biological functions essential to the fitness of the organism, and are therefore promising vaccine candidates.

The M-protein remains a favored GAS vaccine candidate. The predominant bactericidal antibody response raised after GAS infection targets the amino-terminal of the M-protein, the same region that is used for *emm*-typing. However antibodies raised against this region have traditionally been thought to be type specific i.e. antibodies raised against the amino terminus of one M-protein will not recognise the amino terminus of other M-proteins. Candidate vaccines that target this region therefore must include amino-termini from multiple M-proteins. Of the 69 *emm* sequences reported in this study, 14 are represented in the 26-valent amino-terminal GAS vaccine candidate and represent 31.8% of isolates recovered. The highest proportion of isolates included in the vaccine was found in South Indian GAS isolates (17.4%) and included 10 vaccine related *emm* types. Of the North Indian Isolates 14.4% were covered by the vaccine, and included 6 vaccine related *emm* types. Recently a similar 30-valent vaccine has also been reported . Nineteen of the *emm* types present in the 30mer vaccine were present in isolates in this study. The same study also reported that cross-reactivity was observed with another 24 M-proteins not present in the vaccine. Fourteen of these cross-reactive *emm* types were present in this study. In theory, the 30-valent vaccine would induce antibodies that are effective

against 58% of the isolates recovered in this study. The *emm* types present in the 26-valent vaccine, and 30-valent vaccine were chosen based on their importance in North American and European contexts. A re-tailored vaccine, containing amino-termini from M-types common in India or other regions where streptococcal disease is prevalent, may increase vaccine coverage by this approach. However, given the large numbers of *emm* types circulating, and the differences in M-types present in North and South India, such a vaccine may prove difficult to design.

Both traditional and reverse vaccinology approaches have been used to identify alternatives to the M-protein . Rodriguez-Ortega *et al.* used a proteomic approach for identifying streptococcal surface exposed proteins for their use as vaccine candidates. They however used only three *emm* types that were not prevalent in India. Reverse vaccinology, a genome-based approach to vaccine development, has also been used to identify streptococcal vaccine candidates . This approach, however, requires the whole genome sequence of the isolates and to date no prevalent *emm type* strain in India has been sequenced. Proteins that confer essential virulence and biological functions, and are therefore encoded in the core genome, are attractive vaccine targets. In addition to functionality through bactericidal or neutralizing activity, antibodies raised against critical proteins may contribute to prevention of infection by inhibiting protein function. Multi-subunit vaccines containing several proteins have added appeal, as abolition of multiple biological activities may attenuate virulence of the organism further. The targeting of several proteins also potentially has an additive effect with respect to opsonic or neutralizing antibodies. A multi-subunit approach to vaccine can also reduce the probability of vaccine escape that may occur after loss of epitopes targeted by the vaccine that may occur through mutation or lateral gene transfer .

The seven proteins we examined serologically have all been proposed as potential vaccine candidates, and were chosen here because they represent both well characterized and relatively new targets. All these proteins are also predicted to be surface associated. Our array data support this selection because six of these proteins were conserved in all of the invasive isolates tested in this study. R28 is a cell surface virulence protein which has many repetitive sequence and has homology with a protein found in group B streptococcus . SCI is a collagen like protein which is known to have collagen like sequence . HYP has homology with FbaA proteins , suggesting it to have fibronectin binding ability. PSNT is a putative surface nucleotidase protein likely to be involved in nutrient acquisition . The corresponding PNST of *Haemophillus influenzae* has been reported to have protective efficacy in a rat infection model . PSA is a homologue to spy0843 of *S. pyogenes* MGAS5005 which contributed to a protective host immune response in a mouse infection model .

The identified antigens were validated for their vaccine potential by serological analysis. For invasive disease we had the matching isolates and serum samples, because these were acute manifestations that required hospitalization. For other manifestations, which were not acute, it was not possible to have the matching isolate serum samples, so we had to rely on the samples from defined survey areas. The immunogenicity of the identified proteins were tested using ELISA in order to determine their suitability as vaccine candidates. For simplicity, antibodies were determined using a single dilution and results were expressed as OD values. For further studies on these antigens, however, the end point titers or the use of a reference standard serum will be required . Our data demonstrates that antibodies to five of the seven GAS proteins are present in sera in individuals living in a region

where streptococcal disease is common. In general, there was no difference between antibody levels between individuals, irrespective of disease status of individuals. However two proteins, HYP and SfbI, had higher antibody responses in sera from pharyngitis and asymptomatic throat carriers when compared to sera from RF/RHD and asymptomatic non-carriers. Fibronectin binding proteins have been shown to be involved in adherence to epithelial cells and may therefore be important in colonization and early infection. Responses to PSA, which binds to the surface of epithelial cells were also significantly higher in sera from pharyngitis patients. Given the increased antibody response to these three antigens (HYP, SfbI and PSA) in carriers and patients with pharyngitis, they may be excellent candidates for inclusion in a vaccine designed to prevent initial colonization. In this study, generally a lower immune response was detected against all the proteins with serum samples from invasive diseases. This could be due to the fact that the invasive disease is an acute manifestation where the time is too short for the generation of sufficient antibodies to give a high response.

In conclusion, PSNT, C5a peptidase, PSA, HYP and SfbI are promising candidates for a region-specific streptococcal vaccine for the Indian continent. Among these five candidates PSA and PSNT are not well studied and can be further explored in future studies. The two-step approach used in this study to identify vaccine candidates with regional specificity seems to be very promising. However, more isolates and more defined serum samples should be analyzed in further work to support this study.

Acknowledgments

The authors would like to thank N. Janze for excellent technical assistance. This work

was supported by the European Community's Sixth Framework Programme ASSIST under contract number 032390. We are grateful to ASSIST epidemiology team members, especially A. Kumar, K.N. Brahmadathan, V. Abraham and Y. Sharma for providing streptococcal isolates and serum samples. We thank A.P. Oxley for carefully reading the manuscript.

References

1. **Abdissa, A., D. Asrat, G. Kronvall, B. Shittu, D. Achiko, M. Zeidan, L. K. Yamuah, and A. Aseffa.** 2006. High diversity of group A streptococcal emm types among healthy schoolchildren in Ethiopia. Clin Infect Dis **42:**1362-1367.
2. **Agarwal, A. K., M. Yunus, J. Ahmad, and A. Khan.** 1995. Rheumatic heart disease in India. J R Soc Health **115:**303-304, 309.
3. **Banks, D. J., S. F. Porcella, K. D. Barbian, S. B. Beres, L. E. Philips, J. M. Voyich, F. R. DeLeo, J. M. Martin, G. A. Somerville, and J. M. Musser.** 2004. Progress toward characterization of the group A Streptococcus metagenome: complete genome sequence of a macrolide-resistant serotype M6 strain. J Infect Dis **190:**727-738.
4. **Beall, B., R. Facklam, and T. Thompson.** 1996. Sequencing emm-specific PCR products for routine and accurate typing of group A streptococci. J Clin Microbiol **34:**953-958.
5. **Beres, S. B., E. W. Richter, M. J. Nagiec, P. Sumby, S. F. Porcella, F. R. DeLeo, and J. M. Musser.** 2006. Molecular genetic anatomy of inter- and intraserotype variation in the human bacterial pathogen group A Streptococcus. Proc Natl Acad Sci U S A **103:**7059-7064.
6. **Bessen, D. E., and S. K. Hollingshead.** 1995. Horizontal transfer and mosaic-like emm gene structures in group A streptococci. Dev Biol Stand **85:**169-173.
7. **Carapetis, J. R., A. C. Steer, E. K. Mulholland, and M. Weber.** 2005. The global burden of group A streptococcal diseases. Lancet Infect Dis **5:**685-694.
8. **Cole, J. N., A. Henningham, C. M. Gillen, V. Ramachandran, and M. J. Walker.** 2008. Human pathogenic streptococcal proteomics and vaccine development. Proteomics Clin Appl **2:**387-410.
9. **Cunningham, M. W.** 2000. Pathogenesis of group A streptococcal infections. Clin Microbiol Rev **13:**470-511.
10. **Dale, J. B., T. A. Penfound, E. Y. Chiang, and W. J. Walton.** 2011. New 30-valent M protein-based vaccine evokes cross-opsonic antibodies against non-vaccine serotypes of group A streptococci. Vaccine **29:**8175-8178.
11. **Dey, N., D. J. McMillan, P. J. Yarwood, R. M. Joshi, R. Kumar, M. F. Good, K. S. Sriprakash, and H. Vohra.** 2005. High diversity of group A Streptococcal emm types in an Indian community: the need to tailor multivalent vaccines. Clin Infect Dis **40:**46-51.
12. **Eshaghi, M., A. M. Ali, F. Jamal, and K. Yusoff.** 2002. Existence of two emm-like "mrp" and "emm" genes in the mga regulon of the Streptococcus pyogenes strain ST4547. J Biochem Mol Biol Biophys **6:**23-28.
13. **Ferretti, J. J., W. M. McShan, D. Ajdic, D. J. Savic, G. Savic, K. Lyon, C. Primeaux, S. Sezate, A. N. Suvorov, S. Kenton, H. S. Lai, S. P. Lin, Y. Qian, H. G. Jia, F. Z. Najar, Q. Ren, H. Zhu, L. Song, J. White, X. Yuan, S. W. Clifton, B. A. Roe, and R. McLaughlin.** 2001. Complete genome sequence of an M1 strain of Streptococcus pyogenes. Proc Natl Acad Sci U S A **98:**4658-4663.
14. **Hu, M. C., M. A. Walls, S. D. Stroop, M. A. Reddish, B. Beall, and J. B. Dale.** 2002. Immunogenicity of a 26-valent group A streptococcal vaccine. Infect Immun **70:**2171-2177.
15. **Loimaranta, V., J. Hytonen, A. T. Pulliainen, A. Sharma, J. Tenovuo, N. Stromberg, and J. Finne.** 2009. Leucine-rich repeats of bacterial surface proteins serve as common pattern recognition motifs of human scavenger receptor gp340. J

Biol Chem **284:**18614-18623.

16. **Lukomski, S., K. Nakashima, I. Abdi, V. J. Cipriano, R. M. Ireland, S. D. Reid, G. G. Adams, and J. M. Musser.** 2000. Identification and characterization of the scl gene encoding a group A Streptococcus extracellular protein virulence factor with similarity to human collagen. Infect Immun **68:**6542-6553.

17. **Maione, D., I. Margarit, C. D. Rinaudo, V. Masignani, M. Mora, M. Scarselli, H. Tettelin, C. Brettoni, E. T. Iacobini, R. Rosini, N. D'Agostino, L. Miorin, S. Buccato, M. Mariani, G. Galli, R. Nogarotto, V. Nardi Dei, F. Vegni, C. Fraser, G. Mancuso, G. Teti, L. C. Madoff, L. C. Paoletti, R. Rappuoli, D. L. Kasper, J. L. Telford, and G. Grandi.** 2005. Identification of a universal Group B streptococcus vaccine by multiple genome screen. Science **309:**148-150.

18. **McMillan, D. J., R. G. Beiko, R. Geffers, J. Buer, L. M. Schouls, B. J. Vlaminckx, W. J. Wannet, K. S. Sriprakash, and G. S. Chhatwal.** 2006. Genes for the majority of group a streptococcal virulence factors and extracellular surface proteins do not confer an increased propensity to cause invasive disease. Clin Infect Dis **43:**884-891.

19. **McMillan, D. J., R. Geffers, J. Buer, B. J. Vlaminckx, K. S. Sriprakash, and G. S. Chhatwal.** 2007. Variations in the distribution of genes encoding virulence and extracellular proteins in group A streptococcus are largely restricted to 11 genomic loci. Microbes Infect **9:**259-270.

20. **McShan, W. M., J. J. Ferretti, T. Karasawa, A. N. Suvorov, S. Lin, B. Qin, H. Jia, S. Kenton, F. Najar, H. Wu, J. Scott, B. A. Roe, and D. J. Savic.** 2008. Genome sequence of a nephritogenic and highly transformable M49 strain of Streptococcus pyogenes. J Bacteriol **190:**7773-7785.

21. **Miura, K., A. C. Orcutt, O. V. Muratova, L. H. Miller, A. Saul, and C. A. Long.** 2008. Development and characterization of a standardized ELISA including a reference serum on each plate to detect antibodies induced by experimental malaria vaccines. Vaccine **26:**193-200.

22. **Musser, J. M., V. Kapur, J. Szeto, X. Pan, D. S. Swanson, and D. R. Martin.** 1995. Genetic diversity and relationships among Streptococcus pyogenes strains expressing serotype M1 protein: recent intercontinental spread of a subclone causing episodes of invasive disease. Infect Immun **63:**994-1003.

23. **Nitsche-Schmitz, D. P., H. M. Johansson, I. Sastalla, S. Reissmann, I. M. Frick, and G. S. Chhatwal.** 2007. Group G streptococcal IgG binding molecules FOG and protein G have different impacts on opsonization by C1q. J Biol Chem **282:**17530-17536.

24. **Pai, R., M. R. Moore, T. Pilishvili, R. E. Gertz, C. G. Whitney, and B. Beall.** 2005. Postvaccine genetic structure of Streptococcus pneumoniae serotype 19A from children in the United States. J Infect Dis **192:**1988-1995.

25. **Panchaud, A., L. Guy, F. Collyn, M. Haenni, M. Nakata, A. Podbielski, P. Moreillon, and C. A. Roten.** 2009. M-protein and other intrinsic virulence factors of Streptococcus pyogenes are encoded on an ancient pathogenicity island. BMC Genomics **10:**198.

26. **Pfoh, E., M. R. Wessels, D. Goldmann, and G. M. Lee.** 2008. Burden and economic cost of group A streptococcal pharyngitis. Pediatrics **121:**229-234.

27. **Rato, M. G., A. Nerlich, R. Bergmann, R. Bexiga, S. F. Nunes, C. L. Vilela, I. Santos-Sanches, and G. S. Chhatwal.** 2011. Virulence gene pool detected in bovine group C Streptococcus dysgalactiae subsp. dysgalactiae isolates by use of a group A S. pyogenes virulence microarray. J Clin Microbiol **49:**2470-2479.

28. **Reid, S. D., N. M. Green, G. L. Sylva, J. M. Voyich, E. T. Stenseth, F. R. DeLeo, T. Palzkill, D. E. Low, H. R. Hill, and J. M. Musser.** 2002. Postgenomic analysis of four novel antigens of group a streptococcus: growth phase-dependent gene

transcription and human serologic response. J Bacteriol **184:**6316-6324.

29. **Rodriguez-Ortega, M. J., N. Norais, G. Bensi, S. Liberatori, S. Capo, M. Mora, M. Scarselli, F. Doro, G. Ferrari, I. Garaguso, T. Maggi, A. Neumann, A. Covre, J. L. Telford, and G. Grandi.** 2006. Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome. Nat Biotechnol **24:**191-197.

30. **Rohde, M., R. M. Graham, K. Branitzki-Heinemann, P. Borchers, C. Preuss, I. Schleicher, D. Zahner, S. R. Talay, M. Fulde, K. Dinkla, and G. S. Chhatwal.** 2011. Differences in the aromatic domain of homologous streptococcal fibronectin-binding proteins trigger different cell invasion mechanisms and survival rates. Cell Microbiol **13:**450-468.

31. **Sagar, V., D. K. Bakshi, S. Nandi, N. K. Ganguly, R. Kumar, and A. Chakraborti.** 2004. Molecular heterogeneity among north Indian isolates of Group A Streptococcus. Lett Appl Microbiol **39:**84-88.

32. **Sagar, V., R. Kumar, N. K. Ganguly, and A. Chakraborti.** 2008. Comparative analysis of emm type pattern of Group A Streptococcus throat and skin isolates from India and their association with closely related SIC, a streptococcal virulence factor. BMC Microbiol **8:**150.

33. **Sakota, V., A. M. Fry, T. M. Lietman, R. R. Facklam, Z. Li, and B. Beall.** 2006. Genetically diverse group A streptococci from children in far-western Nepal share high genetic relatedness with isolates from other countries. J Clin Microbiol **44:**2160-2166.

34. **Savage, R. S., K. Heller, Y. Xu, Z. Ghahramani, W. M. Truman, M. Grant, K. J. Denby, and D. L. Wild.** 2009. R/BHC: fast Bayesian hierarchical clustering for microarray data. BMC Bioinformatics **10:**242.

35. **Sehr, P., K. Zumbach, and M. Pawlita.** 2001. A generic capture ELISA for recombinant proteins fused to glutathione S-transferase: validation for HPV serology. J Immunol Methods **253:**153-162.

36. **Sette, A., and R. Rappuoli.** 2010. Reverse vaccinology: developing vaccines in the era of genomics. Immunity **33:**530-541.

37. **Smeesters, P. R., A. Vergison, D. Campos, E. de Aguiar, V. Y. Miendje Deyi, and L. Van Melderen.** 2006. Differences between Belgian and Brazilian group A Streptococcus epidemiologic landscape. PLoS One **1:**e10.

38. **Smoot, J. C., K. D. Barbian, J. J. Van Gompel, L. M. Smoot, M. S. Chaussee, G. L. Sylva, D. E. Sturdevant, S. M. Ricklefs, S. F. Porcella, L. D. Parkins, S. B. Beres, D. S. Campbell, T. M. Smith, Q. Zhang, V. Kapur, J. A. Daly, L. G. Veasy, and J. M. Musser.** 2002. 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:**4668-4673.

39. **Stalhammar-Carlemalm, M., T. Areschoug, C. Larsson, and G. Lindahl.** 2000. Cross-protection between group A and group B streptococci due to cross-reacting surface proteins. J Infect Dis **182:**142-149.

40. **Stalhammar-Carlemalm, M., T. Areschoug, C. Larsson, and G. Lindahl.** 1999. 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:**208-219.

41. **Steer, A. C., I. Law, L. Matatolu, B. W. Beall, and J. R. Carapetis.** 2009. Global emm type distribution of group A streptococci: systematic review and implications for vaccine development. Lancet Infect Dis **9:**611-616.

42. **Talay, S. R., A. Zock, M. Rohde, G. Molinari, M. Oggioni, G. Pozzi, C. A. Guzman, and G. S. Chhatwal.** 2000. Co-operative binding of human fibronectin to

Sfbl protein triggers streptococcal invasion into respiratory epithelial cells. Cell Microbiol **2:**521-535.

43. **Terao, Y., S. Kawabata, E. Kunitomo, J. Murakami, I. Nakagawa, and S. Hamada.** 2001. Fba, a novel fibronectin-binding protein from Streptococcus pyogenes, promotes bacterial entry into epithelial cells, and the fba gene is positively transcribed under the Mga regulator. Mol Microbiol **42:**75-86.

44. **Zagursky, R. J., P. Ooi, K. F. Jones, M. J. Fiske, R. P. Smith, and B. A. Green.** 2000. Identification of a Haemophilus influenzae 5'-nucleotidase protein: cloning of the nucA gene and immunogenicity and characterization of the NucA protein. Infect Immun **68:**2525-2534.

Tables

Table 1 Oligonucleotides used in this study.

primer name	sequence 5´- 3´	Description	Annealing temperature (°C)
Scl F	GC**GAATTC**GAGGTTTCTTCTACGACTATGA	***Eco*RI** restriction site	65
Scl R	GC**GTCGAC**ACGTCTGTGGTTGTTGGCTA	***Sal*I** restriction site	
PSNT F	GC**GAATTC**GATCAAGTTGATGTGCAATTCC	***Eco*RI** restriction site	65
PSNT R	GC**GTCGAC**AGTGGAACTAGAGATAGTATTT	***Sal*I** restriction site	
HYP F	GC**GAATTC**GTAGATGGCATCCCTCCAAT	***Eco*RI** restriction site	65
HYP R	GC**GTCGAC**CTGACTCATGGGCCCTAA	***Sal*I** restriction site	
PSA F	GC**GAATTC**GTCAAAGAGCCGATTCTTAAACA	***Eco*RI** restriction site	65
PSA R	GC**GTCGAC**TATTGCAGAGTGTCGTCCT	***Sal*I** restriction site	
R28 F	GC**GAATTC**TCTACAATTCCAGGGAGTGC	***Eco*RI** restriction site	65
R28 R	GC**GTCGAC**CCTTTGACTTGCTGATTTTTACC	***Sal*I** restriction site	
Scp F	GC**GGATCC**AATACTGTGACAGAAGACACTCC	***Bam*HI** restriction site	55
Scp R	GCT**GTCGAC**TTATTACGCTCCTGCTCCTTGTTGGCG	***Sal*I** restriction site	

1**Table 2** Genes differentially conserved in the North and South Indian GAS 2populations (Fishers exact test, $P < 0.05$).

GeneID	Gene name	Conservation (%)	
		North India	South India
SPy0116	hypothetical protein	56.9	91.9
SPy0317	conserved hypothetical protein	55.4	92.6
SPy1006	put. lysin - phage associated	96.9	85.3
SPy2009	hypothetical protein (transposase)	83.1	94.9
SpyM3_0130	streptolysin O	63.1	94.1
SpyM3_0304	conserved hypothetical protein	78.5	94.9
SpyM3_0343	hypothetical protein	93.8	100.0
SpyM3_0653	put. ABC transporter substrate-binding protein	87.7	98.5
SpyM3_0815	put. hemolysin III	70.8	90.4
SpyM3_0823	hypothetical protein	83.1	97.1
SpyM3_0833	put. citrate lyase beta subunit	86.2	97.1
SpyM3_0862	put. DNA/pantothenate metabolism flavoprotein	40.0	91.9
SpyM3_0999	hypothetical protein	93.8	100.0
SpyM3_1390	put. penicillin-binding protein 1A	92.3	99.3
SpyM3_1718	surface lipoprotein DppA	67.7	91.2
SpyM3_1762	hypothetical protein	93.8	99.3

Table 3 Distribution of genes encoding putative vaccine antigens in the Indian GAS population and comparison to isolates from the Netherlands (McMillan, 2006).

Protein	GeneID	Distribution India %	Distribution The Netherlands %
Collagen like surface protein (SCl)	spy1983	100	100
Putative secreted 5'-nucleotidase (PSNT)	spyM3_0591	100	100
C5a peptidase	SpyM3_1726	100	100
R28	AF091393	97.6	55
Putative surface antigen (PSA)	spyM3_0569	99.5	100
Hypothetical protein (HyP)	Spy2009	91.2	70
SfbI	X67947	78.5	68

Figures

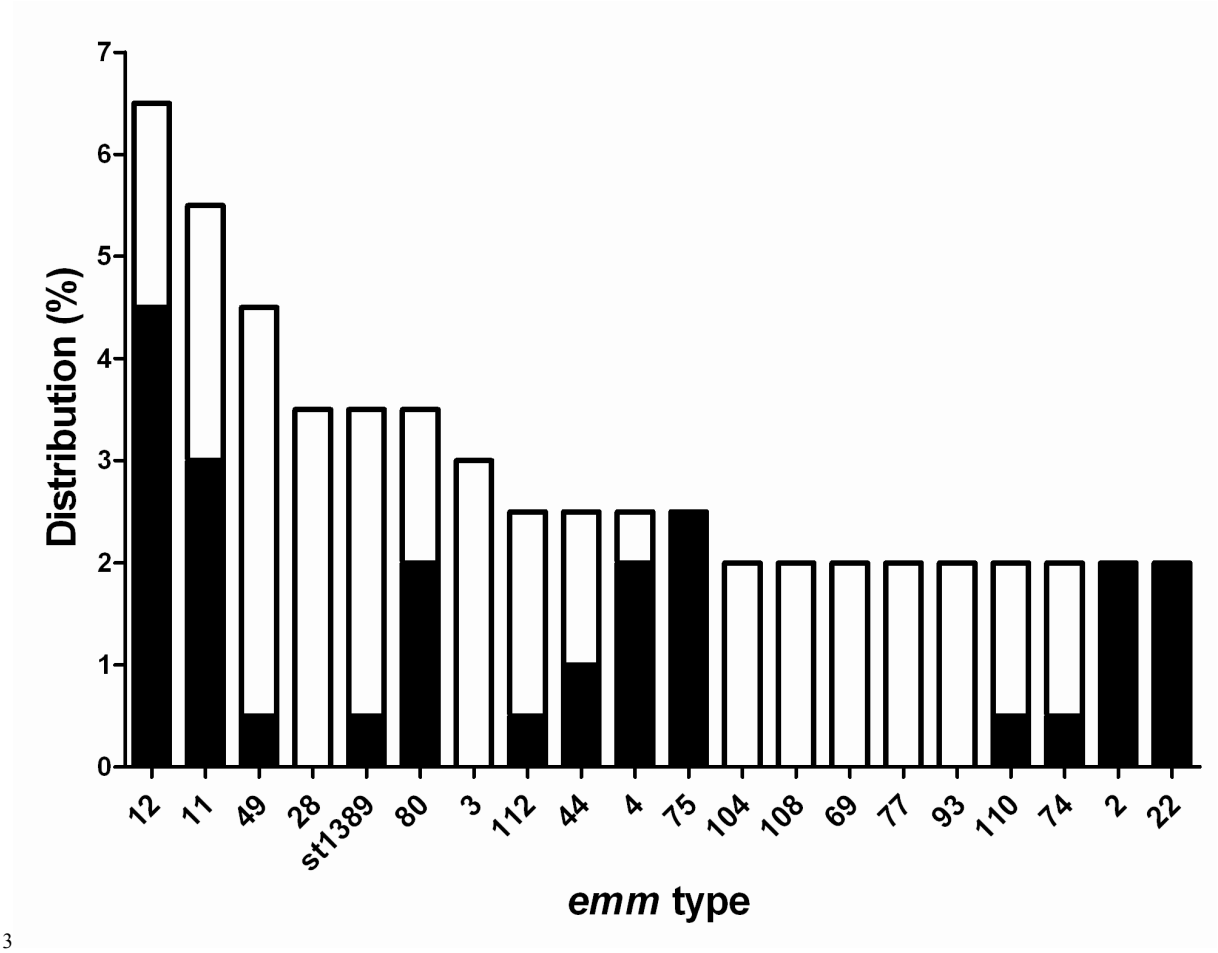

Figure 1 Distribution of the 20 most common GAS *emm* types. Black bars represent *emm* types found in North India, while open bars represent isolates found in South India. The 20 prevalent *emm* types only represented 58% of all GAS isolated in both regions.

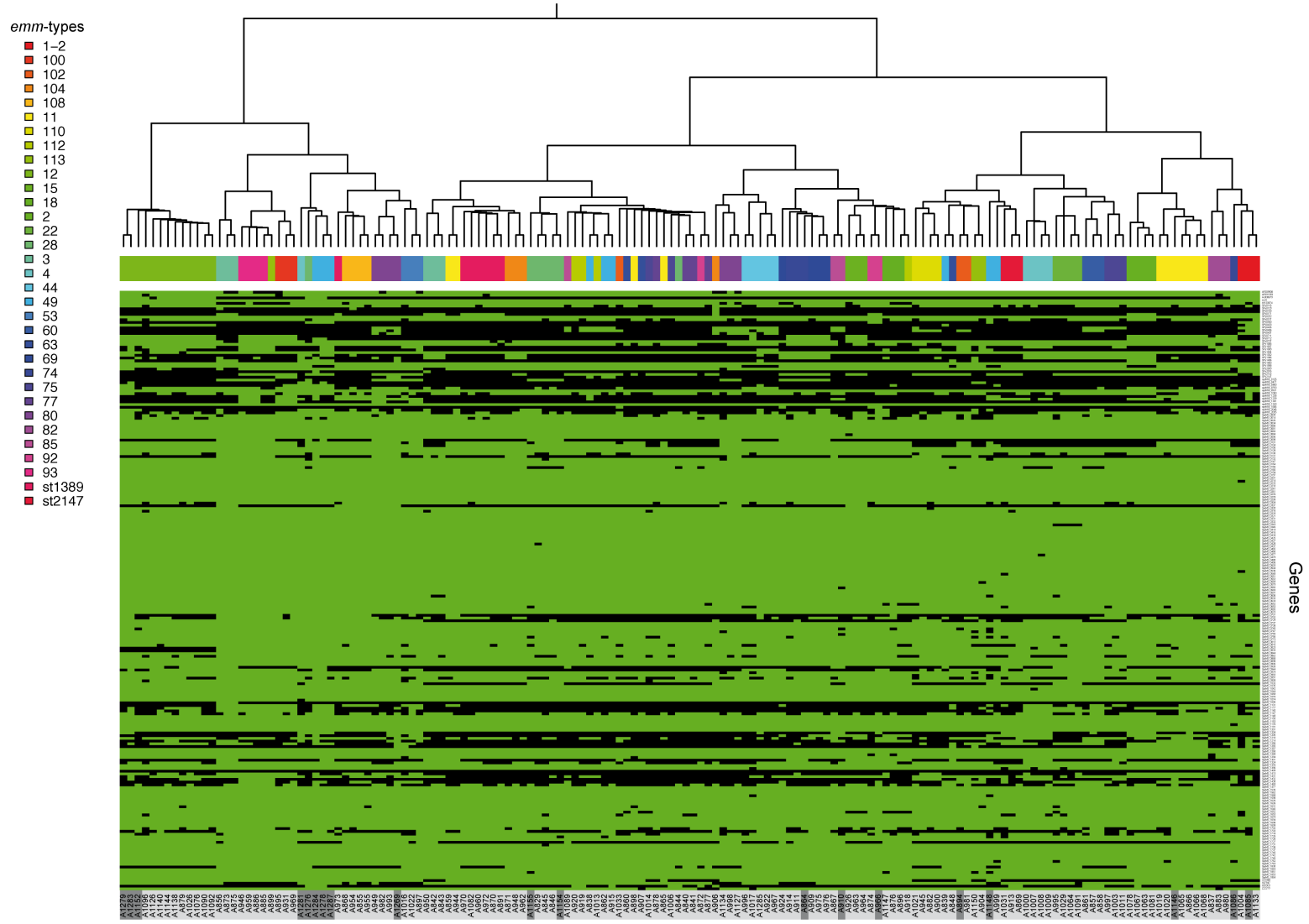
emm-types
1–2
100
102
104
108
11
110
112
113
12
15
18
2
22
28
3
4
44
49
53
60
63
69
74
75
77
80
82
85
92
93
st1389
st2147
Genes
*

Figure 2 Bayesian agglomerative hierarchical clustering of 154 GAS isolates collected in North and South India. The dendrogram shows the clustering of the isolates into groups with similar genetic profiles. Rows represent VF/ECP genes. Genes that are present in specific isolates are shown in green and genes that are absent or highly divergent are shown in black. Columns represent isolates and the colour bar below the dendrogram represents the different emm types. Strain numbers of invasive isolates are shaded in dark grey whereas non-invasive isolates are shaded in light grey. The asterisk indicates a sub-cluster consisting solely of invasive isolates.

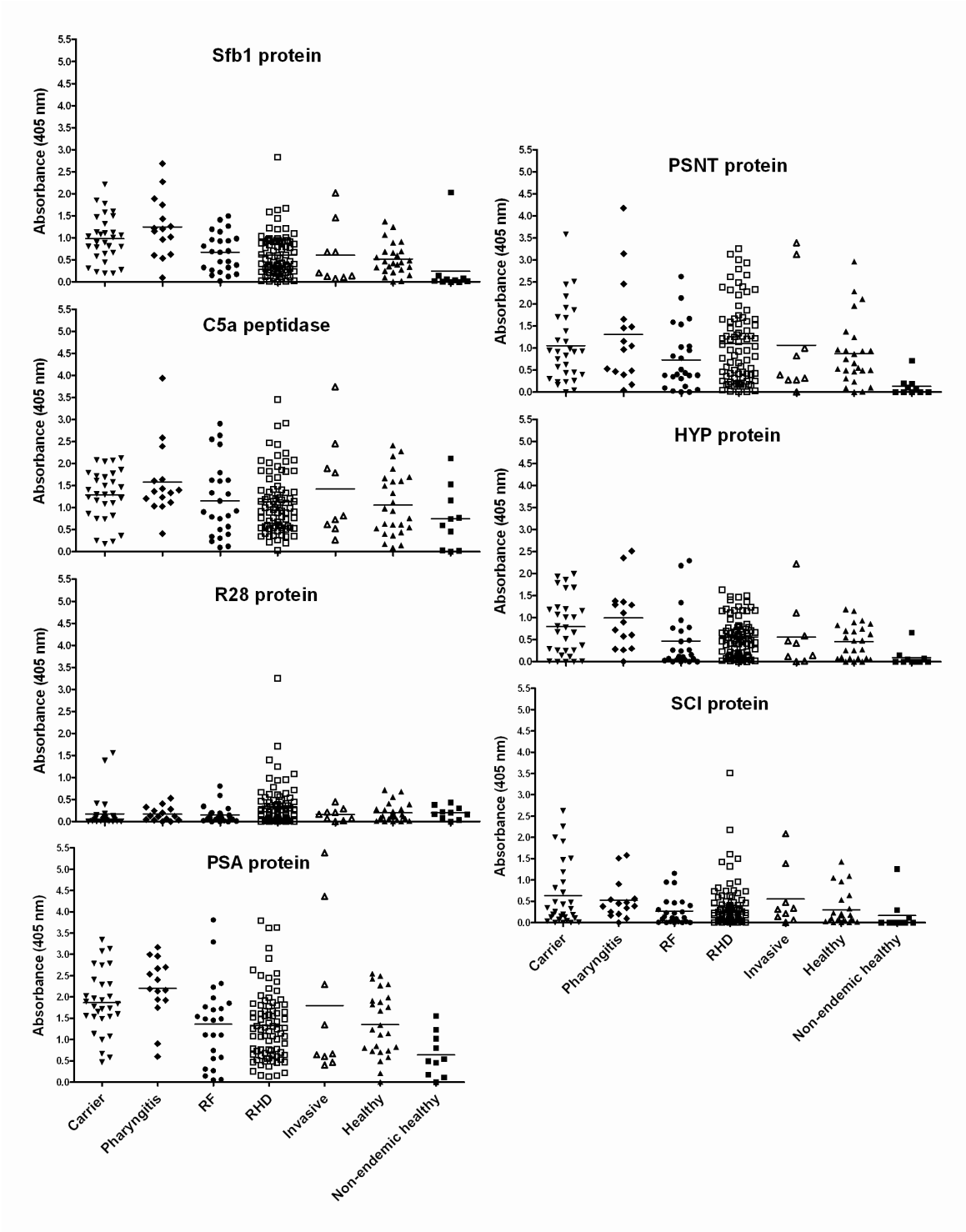

Figure 3 Serological responses to streptococcal antigens in sera. Samples were collected from healthy individuals, asymptomatic carriers and patients presenting with pharyngitis, RF, RHD and invasive diseases. Control sera were collected from patients living in Germany. Statistical differences between the mean absorbance for each group were determined using one-way ANOVA and presented in supplementary table S4.