Clinical Infectious Diseases 2006;43:884-891


1058-4838/2006/4307-0013

# Genes for the Majority Of Group A Streptococcal Virulence Factors and Extracellular Surface Proteins Do Not Confer an Increased Propensity to Cause Invasive Disease

**David J. McMillan,[1,2] R. G. Beiko,[3,4] R. Geffers,[1] Jan Buer,[1] L. M. Schouls,[5] B. J. M. Vlaminckx,[6] W. J. B. Wannet,[5] K. S. Sriprakash,[2] and G. S. Chhatwal[1]**

[1]German Research Centre for Biotechnology, Braunschweig, Germany; [2]Bacterial Pathogenesis Laboratory, Queensland Institute of Medical Research, [3]Institute for Molecular Bioscience, The University of Queensland, and [4]Australian Research Council Centre in Bioinformatics, Brisbane, Australia; and [5]National Institute of Public Health and the Environment, Bilthoven, and [6]University Medical Centre, Utrecht, The Netherlands

***Background.*** The factors behind the reemergence of severe, invasive group A streptococcal (GAS) diseases are unclear, but it could be caused by altered genetic endowment in these organisms. However, data from previous studies assessing the association between single genetic factors and invasive disease are often conflicting, suggesting that other, as-yet unidentified factors are necessary for the development of this class of disease.

***Methods.*** In this study, we used a targeted GAS virulence microarray containing 226 GAS genes to determine the virulence gene repertoires of 68 GAS isolates (42 associated with invasive disease and 28 associated with noninvasive disease) collected in a defined geographic location during a contiguous time period. We then employed 3 advanced machine learning methods (genetic algorithm neural network, support vector machines, and classification trees) to identify genes with an increased association with invasive disease.

***Results.*** Virulence gene profiles of individual GAS isolates varied extensively among these geographically and temporally related strains. Using genetic algorithm neural network analysis, we identified 3 genes with a marginal overrepresentation in invasive disease isolates. Significantly, 2 of these genes, *ssa* and *mf4,* encoded superantigens but were only present in a restricted set of GAS M-types. The third gene, *spa,* was found in variable distributions in all M-types in the study.

***Conclusions.*** Our comprehensive analysis of GAS virulence profiles provides strong evidence for the incongruent relationships among any of the 226 genes represented on the array and the overall propensity of GAS to cause invasive disease, underscoring the pathogenic complexity of these diseases, as well as the importance of multiple bacteria and/or host factors.

Reprints or correspondence: Dr. David J. McMillan, Bacterial Pathogenesis Laboratory, Queensland Institute of Medical Research, PO Royal Brisbane Hospital, Queensland 4029, Australia (david.mcmillan@qimr.edu.au).

Invasive diseases associated with group A streptococcal (GAS) infection have reemerged in the past 25 years and are occurring with increased incidence [1–3]. Although several different M-types have the capacity to cause invasive disease, 2 serotypes, M1 and M3, are most commonly recovered from severe disease cases [4, 5]. Because the propensity to cause different diseases appears to vary from M-type to M-type (and even from strain to strain), it is generally believed a relationship between the possession of a particular virulence gene(s) or allele(s)

and streptococcal disease outcome exists. It has also been postulated that changes in genomic content are also responsible for the observed increase in invasive disease [6–8]. However, studies assessing the association between specific genes and disease have failed to reach a clear consensus [5, 9–11]. For instance, the association between invasive diseases and the genes encoding streptococcal exotoxins, such as *speA* and *speC,* which are thought to be important in invasive disease, have been variably reported [11–13].

For a study addressing associations between virulence genes and diseases, it is meaningful to recover strains from a defined region of low GAS endemicity and within a specified period of time to minimize the effects of panmixia with recently introduced strains. Given the conflicting data relating to individual genes and streptococcal disease, molecular epidemiological studies focusing on only a few of the virulence genes are of limited value. In addition, when examining the entire set of known or well-characterized virulence genes, it is important to understand the variations in the distribution of other as-yet uninvestigated genes and proteins to derive a global perspective of genomic makeup and disease propensity.

In this study, we performed extensive virulence profiling on a cross-section of GAS isolates recovered during a defined time period from patients who presented at hospitals and health care centers with invasive and noninvasive disease in The Netherlands. Profiling, combined with sophisticated machine learning techniques, allowed us to identify 3 genes that showed a marginal overrepresentation in invasive disease isolates.

## METHODS

***Bacterial isolates.*** A surveillance program for invasive GAS disease was conducted in The Netherlands from 1992 to 2003 [14]. From this program, 42 isolates were selected from patients presenting with invasive disease. These isolates comprised the most common invasive M-types (i.e., M1, M3, M6, M28, and M89). A total of 28 of these isolates were associated with streptococcal toxic shock syndrome (STSS), and the remainder were from individuals presenting with other invasive diseases (primarily bacteremia). Twenty-six isolates of matching M-type collected from patients presenting with superficial noninvasive disease (e.g., pharyngitis or otitis/sinusitis) during the same time period were defined as the noninvasive control set. By including the multiple M-types, overrepresentation due to clonal expansion of individual strains was minimized.

***Microarray design and hybridization.*** To assess the complete virulence gene content of isolates, we designed a targeted GAS virulence array. The array includes 74 genes encoding classical GAS virulence factors or genes which are homologous to virulence genes of other species, as well as 144 open-reading frames identified in the M1, M3, and M18 genomes that encoded proteins containing motifs (i.e., a signal sequence and the presence of a single transmembrane domain [15–17]) consistent with extracellular localization, 9 positive control genes, and 9 negative control sequences. Genomic DNA extracted from each GAS strain underwent restriction fragment digestion, labeled through incorporation of dCTP-biotin, and hybridized to the array. After hybridization, arrays were washed, incubated with streptavidin-Cy5, and scanned. Raw fluorogenic intensities for each gene were determined and subsequently normalized against the positive and negative control signals.

***Statistical analysis.*** Initially, *t* tests were used to assess the differences in gene frequencies between invasive and noninvasive groups. Subsequently, 3 machine learning methods were also employed to compare the invasive and

noninvasive isolate datasets. In every instance, the classification strategy always defined noninvasive *Streptococcus* isolates as the "negative" set, and, in separate experiments, either all invasive isolates or only those associated with STSS were defined as the "positive" set. In separate trials, the set of genes that was considered was limited either to those that were over- or underrepresented in the positive set or to only those that were overrepresented. Assignment of isolates to the correct category by different methods was assessed using a classification accuracy score [18].

Genetic algorithm neural network (GANN) [19] is an artificial neural network–based approach that uses a genetic algorithm to select for independent variables (in this instance, particular genes) that are good classifiers, either individually or in combination with other variables. In each iteration of the genetic algorithm, 1000 combinations of 4 genes each were used to train a set of neural networks. The combinations that yielded the best classification accuracy on the test set of isolates (see below) were selected, recombined, and retained in the following iteration. One hundred iterations were performed, and the effectiveness of different genes in classification was assessed via their contribution to high classification accuracy on the test set and by their frequency in the final iteration of the genetic algorithm, because genes that are good classifiers are likely to spread through the genetic algorithm population over time. In addition, 2 other common classification techniques, support vector regression and classification trees, were used to analyze the data [20–23].

## RESULTS

***Overview of gene distribution.*** Of the 226 genes represented in the virulence array, 129 (57%) were found in all isolates irrespective of disease status or M-type, and 160 (71%) were found in >80% of isolates. We next divided the genes on the array into those representing well-characterized GAS virulence factors and those that were identified through the bioinformatic analysis as hypothetical virulence factors. Of the 73 well-characterized GAS virulence factors analyzed here, 17 were known to be phage associated, and the remaining 56 were chromosomally encoded genes. Our data show that two-thirds of the chromosomally encoded genes are ubiquitously present (found in >99% of all isolates) with the frequency of the remaining genes ranging from 11% to 94% (table 1). Genes for extracellular proteases and chromosomally encoded toxins as a group are highly ubiquitous. By contrast, only 2 (5%) of 37 phage-associated genes are found in >99% of all isolates. These observations show that, although variation in the repertoire of virulence characteristics in GAS is contributed to by both phage-associated and chromosomally encoded genes, the former has overwhelming influence.

Clinical Infectious Diseases 2006;43:884-891

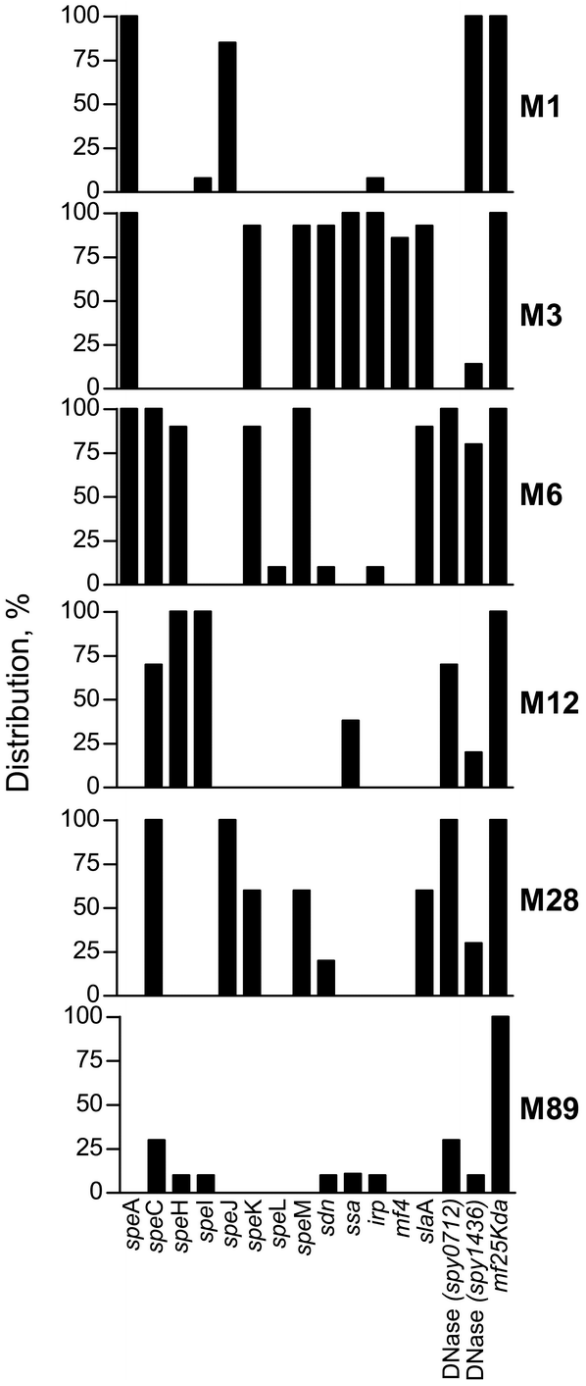
Distribution of well-characterized group A streptococcal virulence genes.

| Virulence class | Distribution, % |
|---|---|
| Adhesins | |
| Putative extracellular matrix binding protein | 61 |
| Putative collagen-like protein (spyM3_1702) | 100 |
| Putative collagen-binding protein (spyM18_0126) | 37 |
| Putative collagen-binding protein (spyM3_0093) | 35 |
| Glyceraldehyde-3-phosphate dehydrogenase | 100 |
| Putative enolase | 100 |
| Laminin-binding protein | 99 |
| PAM | 94 |
| Putative collagen-like protein (spyM3_1703) | 99 |
| Putative collagen-like protein (spyM3_1783) | 92 |
| Collagen-like surface protein | 100 |
| Collagen-like protein SclB | 100 |
| Putative internalin A precursor | 100 |
| Putative pullulanase | 100 |
| R28 | 55 |
| Fibronectin-binding proteins | |
| prtF15 | 76 |
| sfbX | 79 |
| Putative fibronectin-binding protein-like protein A | 100 |
| prtF2 | 63 |
| sfbI | 68 |
| sfbII | 11 |
| Proteases | |
| N-terminal fragment of Sib38 homolgue | 100 |
| Putative exfoliative toxin | 100 |
| Putative C3-degrading proteinase | 99 |
| Streptokinase A precursor | 100 |
| C5A peptidase precursor | 100 |
| Pyrogenic exotoxin B | 100 |
| Putative serine protease | 100 |
| Toxins | |
| Streptolysin O | 99 |
| Putative hemolysin (spyM3_0276) | 100 |
| Streptolysin S–associated protein | 100 |
| Putative hemolysin III | 100 |
| CAMP factor | 92 |
| Putative hemolysin (spyM3_1153) | 100 |
| Other chromosomal virulence factors | |
| Mitogenic exotoxin Z | 100 |
| Pyrogenic exotoxin G precursor | 100 |
| Extracellular hyaluronate lyase | 100 |
| Immunogenic secreted protein precursor homologue | 100 |
| Putative acid phosphatase | 100 |
| Immunogenic secreted protein precursor | 90 |
| DRS | 32 |
| Arginine deaminase | 100 |
| Putative cyclomaltodextrin glucanotransferase | 63 |
| Inhibitor of complement-mediated lysis | 21 |
| Streptococcal protective antigen | 30 |
| Nicotine-adenine-dinucleotide-glycohydrolase precursor | 100 |
| Putative metal-binding protein of ABC transporter | 100 |
| 67-kDa myosin-crossreactive streptococcal antigen | 100 |
| Putative glutathione peroxidase | 100 |
| Putative secreted 5′-nucleotidase | 100 |
| Putative GTP-binding protein LepA | 100 |
| Putative acid phosphatase (class B) | 100 |
| Protein G–related α2M-binding protein | 99 |

| | |
|---|---|
| Putative hyaluronidase | 100 |
| Lantibiotic salivaricin A precursor | 100 |
| Putative exotoxin *spe*J | 30 |
| Bacteriophage-encoded virulence factors | |
| Streptococcal exotoxinspe C | 45 |
| Putative deoxyribonuclease (spy0712) | 44 |
| Streptococcal exotoxin *spe*I | 17 |
| Streptococcal exotoxin *spe*H | 30 |
| Streptococcal exotoxin *spe*L | 1 |
| Streptococcal exotoxin *spe*M | 44 |
| Streptococcal superantigen *ssa* | 28 |
| Putative mitogenic factor *mf4* | 17 |
| Streptococcal exotoxin *spe*K | 42 |
| Streptococcal exotoxin *spe*A | 55 |
| Streptodornase | 24 |
| Mitogenic factor *mf25K* precursor | 100 |
| Putative lysin | 65 |
| Putative deoxyribonuclease (spy1436) | 45 |
| Putative endolysin | 100 |
| Streptococcal phospholipase A2 | 42 |
| Hyaluronidase-phage | 52 |

A greater heterogeneity was observed in the distribution of the adhesins. As a group, the fibronectin-binding genes were much more variable in their distribution. Only 1, the putative fibronectin-binding, protein-like protein A, was present in all isolates. Of the non–fibronectin-binding proteins, 2 putative collagen-binding proteins were found in <40% of isolates. R28, found in 55% of the isolates, including all M28 isolates, is a homologue of the Rib protein in *Streptococcus agalactiae* [24–26]. Of the remaining 20 chromosomally encoded virulence factors, 17 (85%) were found in >90% of isolates.

***The associations between individual genes and invasive disease.*** The observation that some GAS strains are isolated from patients with invasive disease more often than other strains gives weight to the hypothesis that specific genetic differences in GAS strains account for their disease-causing propensity. These genetic differences may be at the nucleotide level or in the gene content, or they may reflect differences in expression patterns of certain virulence factors. An alternative hypothesis is that the increased isolation of these strains may reflect the increased presence of these strains in the community [27]. Both the increased "virulence" and increased "fitness" hypotheses can be explained by changes in the genetic content of invasive isolates when compared with other isolates. To test these hypotheses, we first used *t* tests to compare the frequency of every gene in the invasive set, compared with the noninvasive set. However, no statistically significant differences (defined as $P < .05$) were observed. A similar comparison between noninvasive isolates and isolates collected from patients with STSS, representing "severe" invasive disease, also failed to find any differences in gene distribution.




Because exotoxins may have a role in STSS and may be important in invasive GAS disease, we next analyzed their distribution among the isolates recovered from invasive and noninvasive cases of illness (figure 1). The 3 chromosomally encoded exotoxins (*sme*Z, *mf25* and *spe*G) were found in all isolates. In contrast, all bacteriophage-encoded exotoxins, apart from *spe*A, were found in <50% of the isolates. The most abundant phage-encoded exotoxin gene, *spe*A, was found in 54% of isolates. *spe*C and *spe*M were found in 44% and 43% of isolates, respectively. The least common exotoxin, *spe*L, was found in only a single M6 isolate. Once again, *t* tests found no statistically significant differences between the distribution of any single exotoxin in "invasive" and "noninvasive" isolates. Closer examination of the distribution found the majority of these genes had an M-type–specific distribution, reflecting the distribution of the respective bacteriophages carrying these genes. Comparisons made within M-types also failed to find an association between any specific exotoxin and propensity to cause disease ($P < .05$). Although differing in the distribution, invasive and noninvasive isolates had similar numbers of exotoxin genes in their chromosomes (6.8 vs. 6.7) (table 2).

**Figure 1.** Distribution of bacteriophage-encoded virulence factors among M1, M3, M6, M12, M28, and M89 group A streptococcal isolates.

**Table 2.** Exotoxin genes in invasive and noninvasive group A streptoccocal isolates.

| No. of exotoxin genes | No. (%) of genes in invasive isolates (*n* = 42) | No. (%) of genes in noninvasive isolates (*n* = 26) |
|---|---|---|
| 3 | 4 (10) | 4 (15) |
| 4 | 2 (5) | 0 (0) |
| 5 | 10 (24) | 2 (8) |
| 6 | 3 (7) | 4 (15) |
| 7 | 4 (10) | 5 (19) |
| 8 | 4 (10) | 6 (23) |
| 9 | 14 (22) | 5 (19) |
| 10 | 1 (2) | 0 (0) |

**NOTE.** Genes include *speA, speC, speG, speH, speI, speJ, speK, speL, speM, smeZ, mf25K,* putative DNAse, putative *mf4, ssa,* and *sdn*.

***Machine learning analysis.*** Given the complexity of invasive disease, it is likely that multiple genes are required for progression from initial infection to full invasive disease. The bacteria have to successfully colonize, cross mucosal surfaces, and survive and proliferate in normally sterile body sites. We, therefore, used 3 different machine learning methods in an attempt to identify genes and combinations of genes that are more prevalent in invasive isolates (all invasive isolates or STSS isolates), compared with noninvasive isolates. The classification accuracy values obtained using all invasive isolates as the positive set are summarized in table 3 for all combinations of method and set of genes used. For each trial, the maximum test set prediction accuracy obtained from each of the 10 experimental runs is compared with the same quantity from negative control runs in which positive and negative set members were randomly reassigned. The associated *P* value was calculated using a 1-tailed Wilcoxon signed rank test and expresses the probability that the increase in predictive accuracy associated with the experimental set is because of chance. The GANN scores are particularly high, because they represent the maximum value from a large set of such test scores. The only *P* value <.01 that was observed is associated with the GANN runs performed on genes overrepresented in invasive *Streptococcus* isolates; however, the *P* values from the corresponding support vector machines and classification tree runs are both close to the statistically significant value of *P* = .05. Although the difference between experimental and negative control sets is small (5.4% in the GANN runs), the consistently low *P* values may indicate a small but statistically significant effect of 1 or a few genes on classification accuracy. Only 3 genes were retained in at least 5 of the 10 experimental replicate runs in the GANN trial comparing all genes from "invasive" versus "noninvasive" isolates: streptococcal protective antigen *(spa;* 9 of 10 replicates), streptococcal superantigen *(ssa;* 5 of 10), and a bacteriophage-associated mitogenic factor *(mf4;* 5 of 10). These genes were also retained in multiple replicates of runs investigating only those genes that were overrepresented in invasive isolates *(spa* [6 of 10 replicates], *mf4* [4 of 10], and *ssa* [2 of 10]). When the analysis was repeated comparing isolates collected from patients with STSS with noninvasive isolates, the only *P* value <.01 was associated with the GANN runs on the "overrepresented" data set. The only gene that was retained by a majority of experimental replicates in the "STSS-only overrepresented genes" run was *spa,* which was still present in 9 of 10 replicates. This was also the only gene retained by most replicate runs (7 of 10) on the STSS over- and underrepresented trial data set.

**Table 3.** Accuracy of different methods of classification on a set of invasive and noninvasive isolates.

| Method[a] | Genes | Prediction accuracy, % | | Difference, % | *P* |
|---|---|---|---|---|---|
| | | Experimental set | Negative control set | | |
| GANN | Over- and underrepresented | 91.8 | 89.3 | 2.5 | .216 |
| GANN | Overrepresented | 92.6 | 87.2 | 5.4 | .006 |
| SVM | Over- and underrepresented | 51.1 | 52.8 | -1.7 | |
| SVM | Overrepresented | 60.0 | 48.9 | 11.1 | .035 |
| Tree | Over- and underrepresented | 56.1 | 54.4 | 1.7 | .786 |
| Tree | Overrepresented | 59.5 | 52.7 | 6.8 | .069 |

**NOTE.** GANN, genetic algorithm neural network; SVM, support vector machines; tree, classification tree.

[a] Prediction accuracies are mean values across 10 replicates of the corresponding run. For GANN, the maximum test set accuracy from the entire run of a given replicate contributes to the mean. *P* values shown were calculated using a 1-tailed Wilcoxon signed rank test; consequently, there is no *P* value reported when the prediction accuracy for the negative control set was higher than that for the the experimental set.

## DISCUSSION

This is, to our knowledge, the first study involving a very large set of purported GAS virulence genes and strains recovered from a small geographical region purported to have low GAS endemicity. Our data clearly demonstrate that some GAS virulence genes are unevenly distributed among isolates and that the class of gene can be used as a predictor of gene conservation. Genes encoding fibronectin-binding proteins and virulence genes present on bacteriophage sequences demonstrate the greatest variation in distribution. In contrast, other chromosomally encoded toxins and proteases, most other adhesins, and proteins of putative or unknown function are found in virtually all GAS isolates. Genes encoding proteins of unknown function with variable distributions were invariably found to be associated with bacteriophage sequences.

Despite the large number of virulence gene studies and the large number of isolates analyzed, we were unable to detect a statistically significant difference in the distribution of any gene in the invasive and noninvasive strain sets using traditional statistical methods. Machine learning methods were, therefore, employed in an attempt to identify genes or groups of genes that were overrepresented in invasive isolates. Machine learning methods are becoming powerful tools for the complex analysis of large datasets and have been applied to diverse problems ranging from epidemiology to detection of motifs in DNA sequences [19, 28, 29]. Their advantage lies in their ability to identify patterns, classifications, or associations that, either because of complexity or the large dataset size, are not evident to the investigator. Our study was the first to employ these methods to comparative genomic studies of streptococci. Using GANN, we were able to identify 3 genes that were slightly overrepresented in isolates collected from patients with invasive disease; all 3 learning methods utilized here also appeared to provide accurate statistically significant classification when overrepresented genes were used to distinguish between invasive and noninvasive isolates. Both *ssa* and *mf4* encode superantigens and are carried on bacteriophages, which themselves show restricted M-type distributions. In several other studies, superantigens have been linked to invasive disease [30]. In the present study, unlike *mf4* and *ssa,* genes encoding other superantigens did not segregate with invasive or noninvasive isolates. In this study, *mf4* was found in

only M3 isolates. *ssa* Was found in all M3 isolates, but in only a restricted number of M12 isolates and a single M89 isolate. Thus, although they show an overrepresentation with invasive disease, *mf4* and *ssa* can only account for increases in the propensity of M3 and M12 isolates to cause invasive disease and can not account for the invasive capacity of strains of different M-types. In contrast, *spa* was found in 38% of invasive isolates but in only 19% of noninvasive isolates and is found in isolates of all M-types but not in all isolates within any specific M-type. Although demonstrated to be important in virulence, a functional role for *spa* has yet to be ascertained [31].

Invasive streptococcal disease can differ in both presentation and pathogenic features [32]. This, in turn, suggests that strains associated with different invasive diseases may possess unique genetic factors that may provide an impediment to the comparative nature of the present study. For all invasive diseases, GAS must first cross mucosal barriers to establish infection in sterile environments, suggesting the presence of common factors involved in bacterial transmission and survival. The combined reemergence of these invasive diseases in the past 25 years is a further argument for the presence of shared virulence factors in invasive isolates. Both the invasive and noninvasive isolates used in this study were collected from patients with some type of GAS-associated disease, and therefore, they represent a potential strain bias towards disease-causing isolates. It is also conceivable that under different conditions, some of the noninvasive isolates may be capable of initiating invasive disease and, therefore, obfuscate the results of comparative studies. However, there is no evidence in the literature to suggest that strains associated with noninvasive diseases, such as pharyngitis and otitis, are more or less likely to be associated with invasive disease than isolates collected in community studies which represent "nondisease" strains [33]. Given that a large proportion of individuals may carry GAS asymptomatically [34] and that the incidence of noninvasive diseases far outweigh that of invasive diseases [35], we also believe that there is a minimal chance that a statistically significant number of the strains in our control have the capacity to cause invasive disease.

Our data also provide valuable new insights into the minimum genomic content of GAS required for virulence. The conserved genes present in all isolates are likely to be employed in core virulence or housekeeping functions, such as immune evasion, bacterial transmission, or metabolism, and are essential for the bacteria to survive and multiply within the host. The dispensable nature of genes encoding fibronectin-binding proteins and exotoxins suggest the functionality afforded by these proteins is not critical to the overall fitness of the organisms but may impart new virulence characteristics. With respect to the presence of a core set of genes that may be supplemented with dispensable genes, our study has similar findings to that of the study by Tettelin et al. [36], which undertook comparative genomic comparison on 8 group B streptococcal genomes. In contrast to the closed bacterial genomes of other species, the group B streptococcal pangenome is open. Although it is not a full genomic comparison, our data, as well as the polylysogenic of the GAS genome, suggest it may be open. The influx and efflux of genes from GAS and other streptococci would enable the continuous mixing of genes, some combinations of which may result in increased virulence.

The lack of any strong correlation between any gene and invasive disease further emphasizes the complexity of streptococcal disease. Although this study has focussed of differences at the gene content level, differences in the expression of virulence genes [37] or allelic differences may also be important in the development of invasive disease [38]. Host susceptibility, particularly in relation to responses to individual strains, may be another factor that determines individual disease outcomes [39]. In this regard, the interaction between host HLA

molecules and specific superantigens may well be a critical factor in determining the outcome of invasive disease [30]. Extending this study to investigate expression profiles of these and other GAS virulence factors in appropriate models of infections is a logical next step in the identification of genes that contribute to invasive disease.

## Acknowledgments

***Financial support.*** German Federal Ministry of Education and Research, Network Pathogenomik; Alexander von Humboldt Research Fellowship (to D.J.M.); and the Australian National Health and Medical Research Council.

***Potential conflicts of interest.*** All authors: no conflicts.

## References

1. Kiska DL, Thiede B, Caracciolo J, et al. Invasive group A streptococcal infections in North Carolina: epidemiology, clinical features, and genetic and serotype analysis of causative organisms. J Infect Dis 1997; 176:992–1000.

2. Carapetis J, Robins-Browne R, Martin D, Shelby-James T, Hogg G. Increasing severity of invasive group A streptococcal disease in Australia: clinical and molecular epidemiological features and identification of a new virulent M-nontypeable clone. Clin Infect Dis 1995; 21:1220–7.

3. Vlaminckx B, van Pelt W, Schouls L, et al. Epidemiological features of invasive and noninvasive group A streptococcal disease in The Netherlands, 1992–1996. Eur J Clin Microbiol Infect Dis 2004;23:434–44.

4. Cleary PP, Kaplan EL, Handley JP, et al. Clonal basis for resurgence of serious Streptococcus pyogenes disease in the 1980s. Lancet 1992; 339:518–21.

5. Chatellier S, Ihendyane N, Kansal RG, et al. Genetic relatedness and superantigen expression in group A Streptococcus serotype M1 isolates from patients with severe and nonsevere invasive diseases. Infect Immun 2000; 68:3523–34.

6. Musser JM, Kapur V, Szeto J, Pan X, Swanson DS, Martin DR. 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 1995;63:994–1003.

7. Aziz RK, Edwards RA, Taylor WW, Low DE, McGeer A, Kotb M. Mosaic prophages with horizontally acquired genes account for the emergence and diversification of the globally disseminated M1T1 clone of Streptococcus pyogenes. J Bacteriol 2005; 187:3311–8.

8. Sumby P, Porcella SF, Madrigal AG, 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 2005; 192:771–82.

9. Schmitz FJ, Beyer A, Charpentier E, et al. Toxin-gene profile heterogeneity among endemic invasive European group A streptococcal isolates. J Infect Dis 2003; 188:1578–86.

10. Delvecchio A, Currie BJ, McArthur JD, Walker MJ, Sriprakash KS. Streptococcus pyogenes prtFII, but not sfbI, sfbII or fbp54, is represented more frequently among invasive-disease isolates of tropical Australia. Epidemiol Infect 2002; 128:391–6.

11. Musser JM, Kapur V, Kanjilal S, et al. Geographic and temporal distribution and molecular characterization of two highly pathogenic clones of Streptococcus pyogenes expressing allelic variants of pyrogenic exotoxin A (Scarlet fever toxin). J Infect Dis 1993; 167:337–46.

12. Descheemaeker P, Van Loock F, Hauchecorne M, Vandamme P, Goossens H. Molecular characterisation of group A streptococci from invasive and noninvasive disease episodes in Belgium during 1993–1994. J Med Microbiol 2000; 49:467–71.

13. Hauser AR, Stevens DL, Kaplan EL, Schlievert PM. Molecular analysis of pyrogenic exotoxins from Streptococcus pyogenes isolates associated with toxic shock–like syndrome. J Clin Microbiol 1991; 29:1562–7.

14. Vlaminckx BJ, van Pelt W, Schellekens JF. Epidemiological considerations following long-term surveillance of invasive group A streptococcal disease in The Netherlands, 1992–2003. Clin Microbiol Infect 2005; 11:564–8.

15. Nielsen H, Engelbrecht J, Bunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997; 10:1–6.

16. Nielsen H, Krogh A. Prediction of signal peptides and signal anchors by a hidden Markov model. Proc Int Conf Intell Syst Mol Biol 1998; 6:122–30.

17. Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 1998; 14:378–9.

18. Chou PY, Fasman GD. Empirical predictions of protein conformation. Annu Rev Biochem 1978; 47:251–76.

19. Beiko RG, Charlebois RL. GANN: genetic algorithm neural networks for the detection of conserved combinations of features in DNA. BMC Bioinformatics 2005; 6:36.

20. Joachims T. Making large-scale SVM learning practical. In: Schölkopf B, Burges CJC, Smola AJ, eds. Advances in kernel methods—support vector learning. Cambridge, MA: MIT Press, 1999:169–184.

21. Quinlan JR. C4.5: programs for machine learning. San Francisco: Morgan Kaufmann Publishers, 1993.

22. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002; 46:389–422.

23. Byatov E, Schneider G. Support vector machine applications in bioinformatics. Applied Bioinformatics 2003; 2:67–77.

24. Areschoug T, Carlsson F, Stalhammar-Carlemalm M, Lindahl G. Host-pathogen interactions in Streptococcus pyogenes infections, with special reference to puerperal fever and a comment on vaccine development. Vaccine 2004; 22(Suppl 1):S9–14.

25. Lachenauer CS, Creti R, Michel JL, Madoff LC. Mosaicism in the -like protein genes of group B streptococci. Proc Natl Acad Sci U S A 2000; 97:9630–5.

26. Green NM, Zhang S, Porcella SF, 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 2005; 192:760–70.

27. Johnson DR, Wotton JT, Shet A, Kaplan EL. A comparison of group A streptococci from invasive and uncomplicated infections: are virulent clones responsible for serious streptococcal infections? J Infect Dis 2002; 185:1586–95.

28. Flouris AD, Duffy J. Applications of artificial intelligence systems in the analysis of epidemiological data. Eur J Epidemiol 2006; 21:167–70.

29. Sajda P. Machine learning for detection and diagnosis of disease. Annu Rev Biomed Eng 2006; 8:537–65.

30. Kotb M, Norrby-Teglund A, McGeer A, et al. An immunogenetic and molecular basis for differences in outcomes of invasive group A streptococcal infections. Nat Med 2002; 8:1398–404.

31. McLellan DG, Chiang EY, Courtney HS, et al. Spa contributes to the virulence of type 18 group A streptococci. Infect Immun 2001; 69:2943–9.

32. Stevens DL. Invasive group A streptococcal disease. Infect Agents Dis 1996; 5:157–66.

33. Ekelund K, Darenberg J, Norrby-Teglund A, et al. Variations in emm type among group A streptococcal isolates causing invasive or noninvasive infections in a nationwide study. J Clin Microbiol 2005; 43:3101–9.

34. Hoe NP, Fullerton KE, Liu M, et al. Molecular genetic analysis of 675 group A Streptococcus isolates collected in a carrier study at Lackland Air Force Base, San Antonio, Texas. J Infect Dis 2003; 188:818–27.

35. Carapetis JR, Steer AC, Mulholland EK, Weber M. The global burden of group A streptococcal diseases. Lancet Infect Dis 2005; 5:685–94.

36. Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." Proc Natl Acad Sci U S A 2005; 102:13950–5.

37. Sumby P, Whitney AR, Graviss EA, DeLeo FR, Musser JM. Genome-wide analysis of group A streptococci reveals a mutation that modulates global phenotype and disease specificity. PLoS Pathog 2006; 2:e5.

38. Mascini EM, Jansze M, Schouls LM, Fluit AC, Verhoef J, van Dijk H. Invasive and noninvasive group A streptococcal isolates with different speA alleles in The Netherlands: genetic relatedness and production of pyrogenic exotoxins A and B. J Clin Microbiol 1999; 37:3469–74.

39. Norrby-Teglund A, Chatellier S, Low DE, McGeer A, Green K, Kotb M. Host variation in cytokine responses to superantigens determine the severity of invasive group A streptococcal infection. Eur J Immunol 2000; 30:3247–55.