Evaluation of sclB gene variation in *Streptococcus pyogenes* (Lancefield group A *Streptococcus*) and potential for subtyping

Juliana M. Coelho, Steven Platt and Androulla Efstratiou

Microbiology Services Division, Health Protection Agency, 61 Colindale Avenue, London NW9 5EQ, UK

Variation of sclB gene sequences in group A *Streptococcus* (GAS) strains was evaluated for its potential use in subtyping the most common serotypes of *Streptococcus pyogenes* encountered in the UK. We sequenced a total of 188 strains, including randomly selected invasive GAS and outbreak-related strains. Variability was highest amongst M/emm 89 strains, whereas very little variation was observed amongst M/emm 1 and M/emm 28 GAS strains. Repeat patterns were identified in the collagen structure motif (CSM) of the M/emm 89 GAS strains. The sporadic strains were very diverse and encompassed most of the CSM patterns, whereas the outbreak-related strains were mainly clustered into two CSM groups. sclB gene cluster analysis distinguished outbreak strains from two different healthcare settings in the same geographical area. Sequence variations were assessed by the number of pentameric repeats (CAAAA) present at the 5’ region of the sclB gene. The determination of sclB polymorphisms amongst GAS serotype M/emm 89 strains could be used as an important epidemiological marker to inform clinicians and outbreak control teams during outbreak investigations.

**INTRODUCTION**

*Streptococcus pyogenes*, Lancefield group A *Streptococcus* (GAS), is an important pathogen that can affect both weak and immunocompromised patients, as well as healthy individuals. The most common presentation of GAS-related disease is pharyngitis (‘strep throat’), accounting for approximately 616 million infections per year worldwide (Carapetis et al., 2005), followed by impetigo, which is considered endemic in tropical countries with a prevalence of 12–25 % in children observed on the islands of Fiji (Steer et al., 2009a). Invasive and severe conditions caused by GAS, such as bacteraemia, necrotizing fasciitis and streptococcal toxic shock syndrome, are less frequent but are associated with high mortality rates. Severe GAS-related disease and post-infectious autoimmune sequelae, such as acute rheumatic fever, rheumatic heart disease and acute glomerulonephritis, collectively account for at least 517 000 deaths each year (Carapetis et al., 2005; Wahl et al., 2007).

A variety of virulence factors have been described in GAS strains. Amongst these are the surface-exposed M-protein, which is used to differentiate GAS strains in combination with the characterization of the T-protein complex, recently described as pili proteins of GAS (Mora et al., 2005), and streptococcal serum opacity factor production, a type-specific cell-surface enzyme present in strains of certain GAS M-types (Johnson et al., 2006).

The M protein is encoded by the emm gene, and sequence analysis of the 5’ hypervariable portion of this gene has largely replaced the serological typing method (Beall et al., 1996). The emm type distribution varies worldwide but emm types 1, 3, 28 and 89 are the most commonly found types in reports of GAS-related disease (Imo¨hl et al., 2010; Meisal et al., 2010a; Siljander et al., 2010).

Amongst the surface-exposed proteins more recently proposed as virulence factors in GAS are the collagen-like proteins Scl1 and Scl2, also known as ScIA and ScIB (Lukomski et al., 2000, 2001; Rasmussen et al., 2000; Whatmore, 2001). Both streptococcal and human collagens form triple helices (Mohs et al., 2007), leading to the hypothesis that streptococcal collagens could be involved in the production of antibodies that cross-react with host collagens and may play a role in streptococcal autoimmune disease (Xu et al., 2002). Furthermore, Caswell et al. (2008, 2010) proposed that the ScI protein interacts with human plasma factor H, contributing to the evasion of complement-mediated opsonization and phagocytosis, as well as acting as an adhesin by binding to fibronectin and laminin, facilitating bacterial adhesion to the human extracellular matrix and subsequent internalization by host cells. It has also been demonstrated that ScIB is expressed in higher quantities in human blood than in artificial media...
The sclB gene is composed of a 5’ sequence with varying number of pentameric repeats (CAAAA) followed by a sequence encoding a variable length of (GXY)_n amino acid repeats, constituting the collagen structure motif (CSM). The protein expressed is determined by the number of these repeats, yielding a full-length or truncated protein (Rasmussen & Björck, 2001). The CSM region is mostly organized in repeat tracts and different patterns of repeats have been observed amongst genetically diverse GAS strains (Lukomski et al., 2001; Rasmussen & Björck, 2001; Whatmore, 2001). A detailed analysis of the sclB gene amongst Norwegian M3 GAS strains indicated that polymorphisms in the CSM region could potentially be used as a marker for the differentiation and subtyping of GAS serotype M3 strains (Meisal et al., 2010b). Here, we describe sclB gene variability amongst superficial and invasive UK GAS strains and its potential for use as a subtyping tool during outbreak investigations of GAS-related disease in UK hospitals.

METHODS

Bacterial strains. GAS strains were collected from hospitals across England, Wales and Scotland and sent to the National Streptococcus Reference Laboratory for typing. A total of 188 strains were included in this study, 124 of which were iGAS strains, randomly selected using the ‘Uniform’ command in Stata Statistical Software 11.1 (StataCorp) from two periods of enhanced surveillance (2003–2004 and 2008–2009), 17 were superficial isolates selected over the same periods, 24 strains were selected from outbreaks at two elderly care homes in 2010, and 23 outbreak and sporadic strains were isolates from one hospital between 2004 and 2010.

Typing. All strains were characterized by emm gene typing, T-protein typing and determining opacity factor production (Neal et al., 2007). GAS strains were grown overnight at 30 °C in Difco Todd–Hewitt broth containing 0.6 % (v/v) trypsin for T-typing, and overnight at 37 °C in Difco Todd–Hewitt broth to determine opacity factor production according to standard protocols (Johnson et al., 1996, 2006).

The emm types were determined according to the protocol and guidelines available on the CDC website (http://www.cdc.gov/ncidod/biotech/strep/M-ProteinGene_typing.htm). When sequence data obtained using the CDC recommended primers were ambiguous, alternative primers (MFI, 5’-ATAAGGGAGCATAAAAATGCT-3’), and M1, 5’-AGCTTTAGTTTCTCTTGCG-3’) (Podbielski et al., 1991) (Sigma) were used for the amplification of the emm gene.

Chromosomal DNA preparation. GAS strains were streaked for single colonies on Columbia blood agar plates and incubated overnight at 37 °C. Crude DNA extracts were then prepared by resuspending a loopful of each test sample in 62 μl lysis buffer containing 50 μl TE, 10 μl mutanolysin (3000 U ml⁻¹) (Sigma–Aldrich), and 2 μl hyaluronidase (30 mg ml⁻¹) (Sigma–Aldrich). The bacterial suspensions were incubated for 30 min at 37 °C and heat inactivated for 10 min at 100 °C. DNA was then purified using the QuiAxtractor robot with the QuiAxtractor DX reagent and plastic ware (Qiagen) according to the manufacturer’s instructions. Purified DNA was used for amplification of the emm and sclB genes.

Nucleotide variation in the sclB gene. The complete sclB gene was amplified from GAS strains using the published primers SclBF1 (5’-GTGTTATCCTATTTAGTG-3’) (Meisal et al., 2010b) and SclBR (5’-ACTTTCCATGTTAGGTCG-3’) (Lukomski et al., 2001) using a HotStar Taq Mastermix kit (Qiagen). Amplification conditions were as follows: initial hot start at 95 °C for 15 min, followed by 30 cycles of 95 °C for 20 s, annealing at 50 °C for 1 min and extension at 72 °C for 2 min 30 s, with a final extension step at 72 °C for 10 min. PCR products were purified using the Agencourt Ampure or CleanSEQ purification systems (Beckman Coulter) and sequencing was performed with a BigDye Terminator v1.1 or v3.1 Cycle Sequencing kit (Applied Biosystems). Sequencing reactions were set up with PCR primers, and the following internal sequencing primers were used for the determination of the complete sclB gene sequence from isolates of GAS serotype M/emm 89: SclB89-A (5’-GAGAT-CGATTAGAAGACCTAAATG-3’), SclB89-B (5’-TGCGCAAATC-AAAAACGTGACAC-3’) and SclB89-B-RC (5’-GTGTCACGTCCTT-TTGATTGGCCA-3’).

Sequence data were analysed using BioNumerics software (Applied Maths). Repeat tracts on the CSM regions were identified using a diagonal plot method in the Dotlet program (Junier & Pagni, 2000) and cluster analysis of these patterns was performed using the minimum spanning tree method within the BioNumerics Software.

RESULTS

GAS emm gene typing, T-protein typing and opacity-factor production

GAS emm gene sequence analysis typed 50, 45 and 93 strains as M/1emm 1, M/1emm 28 and M/1emm 89, respectively. The M/1emm types, opacity factor production and T-types were in agreement with the most commonly found combinations in previous studies (Beall et al., 1998; Johnson et al., 2006; Neal et al., 2007). All M/1emm1 strains were opacity-factor-negative and of T-type 1; all M/1emm 28 strains were opacity-factor-positive and of T-type 28, except one strain that was T-type 11. All M/1emm 89 strains were opacity-factor-positive but varied in their T-type profile, the majority of which (n=83, 88 %) were T-type 3/13/B3264, with five (5 %) T-type 3 strains, one T-type 28, one T-type 11 strain, and five strains that were non-typable using the available T-sera.

sclB variation amongst randomly selected strains

The sclB gene shared an overall sequence similarity of 77 % amongst M/1emm 1, 28 and 89 GAS strains included in the study. Three main clusters were distinguished according to the nucleotide analysis, which paralleled the M/1emm type. M/1emm 1 strains showed the highest sclB gene sequence similarity (99.9 %), followed by types M/1emm 28 (92–99 %) and M/1emm 89, which showed most variation (85–99 %).

All M/1emm 1 strains had out-of-frame sclB alleles, whereas 33 % of M/1emm 28 and 47 % of M/1emm 89 strains were in-frame, dictated by the variable number of CAAAAA repeats
A CSM composed of a string of varying numbers of in-frame sequences. Interestingly, 93% (n=42) of the M/emm 89 strains had varying numbers of hexameric repeats amongst the pentamers, which affected the translational frame of 24 strains (53%). An analysis of the translational frame versus invasive/non-invasive infection did not find a clear correlation between invasive strains and in-frame sclB alleles. The majority of sclB alleles (53%, n=87) were from sterile sources of collection and were out-of-frame, and 23% (n=35) were from sterile sources of collection but were in-frame. Those from non-sterile sites (n=31) were mainly out-of-frame (16%), followed by six strains with in-frame sequences.

A CSM composed of a string of varying numbers of (GXY)$_n$ residues was identified in all strains included in this study. Dot plot analysis showed a series of parallel lines when comparing CSM sequences from M/emm 89 strains, indicating that repeat tracts (minisatellites) were present on that region, but no significant tracks of repeat patterns were observed amongst the GAS M/emm 1 and M/emm 28 strains. The motifs observed amongst the M/emm 89 strains were distinct from the patterns M3CSMR-1 and M3CSMR-2 described in Norwegian M3 strains (Meisal et al., 2010), as shown in Fig. 1. The pattern numbering system is based on counts of motif 1 (M89CSMR-1, AKGEPG), followed by motif 2 (M89CSMR-2, PKGEPG), motif 3 (M89CSMR-3, GKDQDGKDGLPGKD) and motif 4 (M89CSMR-4, GKDQNGKDGLPGKD), therefore a pattern 13/5/3/2 means that there are 13 × M89CSMR-1, 5 × M89CSMR-2, 3 × M89CSMR-3 and 2 × M89CSMR-4 motifs.

Twenty-seven repeat patterns were detected on the CSM region of M/emm 89 strains randomly selected from two periods of iGAS enhanced surveillance (Coelho et al., 2010; Lamagni et al., 2008, 2009). These motif patterns clustered into nine distinct groups of multiple strains and 18 unique patterns, the most common being 16/5/3/2, followed by 13/5/3/2, 7/1/5/3, and 15/5/3/2. sclB nucleotide cluster analysis distinguished all strains within each of the above CSM clusters, which confirmed good sclB sequence variability amongst the randomly selected strains of serotype M/emm 89.

**sclB variation amongst outbreak and non-outbreak M/emm 89 strains**

Further evaluation of sclB polymorphisms was performed with additional sets of M/emm 89 strains from two outbreaks in elderly care settings (setting 'A' and 'B') and from an outbreak in a maternity ward (setting 'C'), all occurring between July and October 2010. Fourteen sporadic strains identified by hospital C during 2004–2010 and two sporadic strains identified by hospital B in 2010 were also included for comparison. The strains were clustered into nine groups according to their CSM repeat patterns. The largest group (pattern 15/5/3/2) comprised 32 strains, followed by one group with four strains, one group with three strains, two groups with two strains each, and four unique patterns. Most of the outbreak strains from settings B (n=18) and C (n=9) were clustered in the largest group, which also included five of the sporadic strains from setting C. It is of note that settings B and C are geographically close, being in the north of the UK, 45 miles apart. The four outbreak-related strains from the elderly care setting A in Southeast England (385 miles from setting B) were clustered alone and separate from the group with strains from settings B and C. The remaining strains from sporadic referrals were diverse according to their CSM repeat patterns and constituted all the clusters with two or less.

An overall combined picture of sclB CSM clusters, including all M/emm 89 GAS strains from this study, confirmed that unrelated strains are very diverse and those with known outbreak links cluster very tightly according to sclB gene polymorphisms. A minimum spanning tree based on all CSM patterns observed in this study is shown in Fig. 2. Thirty-four repeat patterns were identified in total, with 14 distinct groups of multiple strains and 20 unique patterns. The unrelated strains comprised 87% of the CSM patterns, whereas the outbreak-related strains were predominantly clustered into two CSM groups, the biggest cluster, now with 36 strains, including the outbreak strains from setting B and C, and the setting A outbreak cluster with 4 strains (Fig. 2).
**sclB nucleotide sequence variations within CSM clusters**

A detailed analysis of *sclB* sequences from the largest CSM cluster (with 36 strains) subdivided this group into four multimember subclusters with 100% internal identity, and five strains with unique DNA sequences (see Fig. 3). The variations observed were due to the number of pentameric repeats (10 to 17), with the remainder nucleotide sequences being identical. The biggest subcluster (group 1, Fig. 3) included 15 of the outbreak-related superficial isolates from the elderly care ward in setting B, 14 of which were from patients and one from a staff member. The other strain that clustered within this subgroup was an iGAS strain isolated from a patient with septic shock symptoms at hospital C in February 2010. The second subgroup (group 2, Fig. 3) was composed of eight strains from the maternity ward outbreak in hospital B, including four strains from two mother/baby pairs, three strains from other post-partum mothers and one strain from a newborn baby who presented symptoms only after discharge from the maternity unit. This subgroup also included one strain (isolated in August 2010) from a GP patient with ear discharge, with no known links to the maternity ward, but who resides in the same geographical area.

A third subgroup (group 3, Fig. 3), containing strains with identical nucleotide sequences, comprised two strains from patients in the elderly care ward outbreak in hospital B as well as one randomly selected strain received in November 2009 (distance between hospitals of 188 miles) and a sporadic referral from hospital C. Finally, the smaller subcluster (group 4, Fig. 3) from the main CSM group included two strains, the first related to the elderly care ward outbreak in hospital B and the second a randomly selected strain from a hospital 291 miles apart from hospital B.

The four strains related to fatal iGAS infections from patients in elderly care ward A were indistinguishable from each other based on their *sclB* sequences and different from all *sclB* sequences of the GAS M/*emm* 89 strains included in this work, supporting the hypothesis that this strain was spread nosocomially.

**DISCUSSION**

GAS typing, initially performed by serological methods but now mainly by sequencing of the *emm* gene, has, over the years, contributed towards the identification of clones (Aziz & Kotb, 2008) during surveillance of GAS-related disease, and has provided valuable information during the management of GAS outbreaks, both in the community and in healthcare institutions (Aguero et al., 2008; Thigpen et al., 2007; Vikerfors et al., 2009). The replacement of the original serological method by sequencing methods means that resulting data are less ambiguous, more uniform and easier to exchange between laboratories (Neal et al., 2007), providing better information of GAS type distribution across the globe (Steer et al., 2009b). In the UK, 63% of GAS isolates received by the National Reference Laboratory during 2008–09 were of the types M/*emm* 1, M3/*emm*, M/*emm* 28 and M/*emm* 89; however, ascertaining whether a particular strain of the same M/*emm* type was related to another was hindered, requiring additional typing methods.
to add finer resolution to the current method. Pulsed-field gel electrophoresis (PFGE) has been shown to discriminate between GAS strains of the same M type (Bahnan et al., 2011; Carriço et al., 2006; Lacy & Horn, 2009) and has been used to characterize outbreak strains, such as those from a foodborne outbreak in a Japanese university, where GAS strains were isolated from kitchen workers, food products and affected individuals (Takayama et al., 2009). This method, however, is laborious and highly dependent on individual training and skills, making comparison of results between laboratories difficult, if not impossible.

Multi-locus sequence typing (MLST) is used for the delineation of clones or clonal complexes; by using this method, data can be easily exchanged between laboratories and has been shown to subdivide some M types (Enright et al., 2001). GAS epidemiological investigations have benefited from MLST data; Sakota et al. (2006) observed that genetic diversity amongst superficial GAS strains from Nepal was remarkably higher than that of strains from the USA, aiding the delineation of GAS clones and elucidating their distribution in different parts of the globe. In hospital settings it can add valuable information to determine the distribution of clones and any possible change in prevalent types, which might signify the appearance of more virulent and successful clones (McGregor & Spratt, 2005). Strain characterization by MLST is sometimes combined with PFGE. A study of the decrease of macrolide-resistant clones with phenotype MLSB in relation to phenotype M in Portugal suggested the possibility of a Europe-wide dissemination of only a few macrolide-resistant lineages, according to PFGE and MLST data (Silva-Costa et al., 2008). The applicability of these techniques for strain discrimination during local epidemiological investigations, however, is hindered by the intrinsic high cost and laborious nature of these tests.

Polymorphisms on the \(sclB\) gene of Norwegian M3 strains have been proposed as potential targets for fine typing (Meisal et al., 2010b); however, the repeat region of the collagen structure motif identified amongst M3 strains from the UK were extremely large (data not shown), hindering the collection of reliable sequence data and the evaluation of the method for robustness and usefulness for subtyping UK M3 strains. Subtyping of the GAS \(M/emm\) 1 strains included in this study was not achieved as their \(sclB\) sequences were highly homogeneous, with variation observed only at the 5’ region by the variable number of

---

Fig. 3. UPGMA tree of the \(sclB\) gene multiple alignments. Sequences are from all strains included in the largest CSM cluster shown in Fig. 2. The largest subgroups encompassed the outbreak strains from the elderly care home (group 1) and the maternity ward (group 2).
pentameric repeats. An alternative approach for fine typing M/emm 1 strains is the investigation of polymorphisms on the sic (streptococcal inhibitor of complement) gene. The sic gene is present only in M/emm 1 strains and has been shown to discriminate between outbreak- and non-outbreak-related M1 strains (Bidet et al., 2009; Hoe et al., 1999).

Some variability was observed amongst sclB alleles of GAS serotype M/emm 28 strains; however the discriminatory power was greater amongst M/emm 89 strains. A high number of distinct CSM patterns from the randomly selected M/emm 89 strains confirmed that there was significant gene variability among these strains, and the high similarity observed amongst outbreak-related strains confirmed their close relationship in comparison with the randomly selected strains. Even though the cluster containing the outbreak strains also contained four randomly selected strains, all were from the North of the UK, indicating a possible relationship between them. The number of pentameric repeats was useful as a fine marker within this typing system as two main outbreaks were defined based on nucleotide sequences; however, it is of note that the number of CAAAA repeats could change in passages through human host (Rasmussen & Björck, 2001) and additional epidemiological information should be taken into consideration during outbreak investigation. In conclusion we found that determination of sclB polymorphisms amongst GAS serotype M/emm 89 strains could be used during outbreak investigations in combination with the epidemiological information available. This is, to our knowledge, the first description of a subtyping marker for GAS M/emm 89 strains.

REFERENCES


**scIB gene for group A Streptococcus subtyping**


