Microbiology (1998), **144**, 629–637 Printed in Great Britain

Extensive genetic diversity among clinical isolates of *Streptococcus pyogenes* serotype M5

Meeta Desai,[1] Asha Tanna,[2] Androulla Efstratiou,[2] Robert George,[2] Jonathan Clewley[1] and John Stanley[1]

Author for correspondence: John Stanley. Tel: +44 181 200 4400. Fax: +44 181 200 1569.

Molecular Biology Unit, Virus Reference Division,[1] and Streptococcus and Diphtheria Reference Unit, Respiratory and Systemic Infection Laboratory,[2] Central Public Health Laboratory, 61 Colindale Avenue, London NW9 5HT, UK

The genetic diversity of clinical isolates of *Streptococcus pyogenes* serotype M5 has been characterized. Strain genotypes were defined by macrorestriction profile, 16S ribotype, *emm* gene subtype, insertion element IS*1239* profile, and exotoxin gene determinant. By these criteria, clinical isolates of M5 constituted a multiplicity of strain clusters rather than a homogeneous population as found for certain serotypes. Distance matrices and an unrooted tree were constructed from macrorestriction data with three rarely cutting endonucleases, determined by PFGE. A single IS*1239* profile was common to 85% of isolates but there was great diversity of both ribotype and macrorestriction profile, and 18 different *emm* gene subtypes were detected by PCR-RFLP. DNA sequence analysis of the antigen-coding 5′ (hypervariable) region of *emm* gene amplicons (about 240 bp) showed that 14/18 exhibited up to 6% divergence. Four amplicons had highly divergent sequences – corresponding to those previously determined for *emm*6, *emm*11, *emm*18 and *emm*77. Further serological and hybridization studies were used to analyse the discrepancy between the Lancefield serotype of these strains (M5) and their *emm* genotype. Overall, this study shows a high degree of genetic diversity in serotype M5, with implications for the Lancefield scheme itself, for the epidemiology of group A streptococci, and for recombinant DNA strategies for M protein-based vaccine development.

Keywords: group A *Streptococcus*, *emm* gene, polymorphism

INTRODUCTION

Streptococcus pyogenes group A streptococci (GAS) are causative agents of human diseases which range in severity from pharyngitis to toxic-shock-like syndrome and necrotizing fasciitis. *S. pyogenes* possesses a number of virulence factors, of which the best known is the M protein. The C-terminus of this fibrillar protein is anchored to the bacterial cell membrane, while the N-terminus protrudes outwards from the cell. The M protein confers ability to resist phagocytosis by polymorphonuclear leucocytes in the absence of type-specific antibodies (Fischetti, 1989). Anti-M protein immunity is type-specific (Lancefield, 1962), and more than 80 serotypes can be recognized by precipitin reactions constituting the M serotyping scheme. Type-specific antisera are thought to be directed primarily against the short hypervariable region which follows the conserved leader sequence at the N-terminus of M protein.

The *emm* gene encoding M protein can be amplified by PCR with conserved primers (Podbielski *et al*., 1991; Whatmore *et al*., 1994), and many *emm* gene sequences have been established (Beall *et al*., 1996, 1997; Upton *et al*., 1996; Whatmore *et al*., 1994). The sequences of the 5′ regions of *emm* genes from 79 M serotypes exhibited a phylogeny which was not congruent with that derived for the whole genome by multilocus enzyme electrophoresis (MLEE). This suggests that variation in *emm* gene sequences has been generated by horizontal gene transfer and recombination (Whatmore *et al*., 1994). Similarly, the sequence of the downstream gene, *enn*, from one strain of serotype M5 has a mosaic structure, consisting of segments of divergent origin (Whatmore & Kehoe, 1994).

The Lancefield scheme (Lancefield, 1962) for serotyping GAS isolates is based on the association of the T

Abbreviations: GAS, group A streptococci; OF, opacity factor.

agglutinins, the opacity factor (OF) reaction and certain groupings of M serotypes. For example, isolates which are T1, OF^- are invariably serotype M1. The *emm* gene sequences of most GAS isolates are in agreement with these classical M and T antigen associations, but there is recent evidence that a minority of isolates are T-nontypable or contain *emm* genes encoding M proteins which have no recorded T-antigen association (Beall *et al*., 1997).

There is evidence (Cleary *et al*., 1992; Musser *et al*., 1991, 1993) that in the case of serotype M1, a globally distributed (genotypic) clone is responsible for most GAS diseases. By contrast, examination of small sample sets by PFGE and ribotyping alone indicated that other serotypes might be greatly more diverse genetically than is M1 (Stanley *et al*., 1995).

In this study, we examined the genetic diversity of randomly chosen clinical isolates of serotype M5, a major serotype associated with a wide spectrum of GAS diseases from pharyngitis to severe invasive infections. We analysed strain phylogenies by carrying out genome macrorestriction with three (rather than one) rarely cutting endonucleases, by ribotyping with three endonucleases, and by restriction fragment length polymorphism (RFLP) in and around loci of the transposable insertion element IS*1239* (Kapur *et al*., 1994). We placed other genetic markers (*emm* gene polymorphism and exotoxin determinants) onto the phylogenetic tree based on macrorestriction. Polymorphism of the *emm* gene was investigated by nucleotide sequencing of the 5′ end of the gene and, where indicated, by further serological studies.

METHODS

Bacterial strains and growth conditions. The type strain (NCTC 8193; obtained from the National Collection of Type Cultures, Colindale, UK), a reference strain traditionally used for producing antisera to M type 5 (strain 100065), and 38 clinical isolates of *S. pyogenes* serotype M5 (Table 1) were analysed. Isolates from the following noninvasive diseases were included: scarlet fever, sore throat, infected skin, vaginitis, and ear and eye infection. Isolates from the following invasive diseases were included: septicaemia, cellulitis, pneumonia, bacteraemia, brain abscess, bursitis, lymphadenitis, chest infection, and renal disease. Streptococci were cultured aerobically at 37 °C for 18–24 h on horse blood agar plates, and preserved for reference in blood glycerol (16% v/v) broth (Oxoid) at −70 °C.

Serological analysis. Isolates were serotyped before and after genotype analysis according to standard methods (Johnson *et al*., 1996; Lancefield, 1962). Lancefield acid extracts were re-examined for precipitin lines of identity with type-specific antisera against the M type predicted from the relevant *emm* amplicon sequence. Standard Ouchterlony double-diffusion tests (Johnson *et al*., 1996), were used in a format designed to reveal one or more serotype identities.

***emm* gene polymorphism, 5′ sequence analysis and hybridization studies.** The 'all M' PCR primers and conditions of Podbielski *et al*. (1991) were used for amplification of the *emm* gene, and RFLP analysis of *emm* amplicons was carried out as previously described (Stanley *et al*., 1996). Amplicons were purified with Gene Clean (Bio101), and subjected to cycle sequencing using the 'all M' forward primer of Podbielski *et al*. (1991) or forward primer 1 of Whatmore & Kehoe (1994), with the PRISM Dye Terminator Cycle Sequencing kit. Analysis was performed on an ABI 373A DNA sequencer. For four amplicons which had highly divergent sequences from that of *emm5* (>6%; termed *emm**), the second strand was also sequenced, using primers designed from the first-strand sequence data (see below).

emm gene hybridization studies were carried out for the four M5 strains (R2223, R2357, R2160 and R2247) from which *emm** products were amplified as follows. Genomic DNA (5 μg) of these strains and the M5 type strain was blotted onto Hybond-N (Amersham) membrane (five replicate filters) using a slot-blot apparatus (Stratagene), and fixed by UV cross-linking (Stratagene). The 5′ regions of the four *emm** genes and that of the M5 type strain were amplified using forward primer 1 (Whatmore & Kehoe, 1994) and five new reverse primers designed from first-strand sequence data (see Results). PCR reactions were carried out under standard conditions, with 25 cycles of denaturation at 94 °C for 1 min, annealing at 45 °C for 1 min, and extension at 72 °C for 1 min in a Robocycler (Stratagene). Amplicons, ranging in size from 250 to 300 bp, were labelled with [α-^{32}P]dCTP by random priming (Feinberg & Vogelstein, 1983) using the Multiprime DNA labelling system (Amersham), and hybridized with individual filters.

Genomic DNA (5 μg) of the type strain and the four strains alone was digested with *Hin*dIII, electrophoresed in 1·0% agarose (55 V, 17 h), and Southern blotted (five replicate filters). The same 5′ *emm* PCR amplicons were labelled with biotin-16-dUTP (random-primed labelling kit; Boehringer Mannheim) and hybridized with these filters. After stringent washing (two washes, 30 min each, 63 °C, 0·16× SSC/0·1% SDS), they were developed as previously described (Stanley *et al*., 1996).

Macrorestriction, PFGE and data analysis. PFGE was carried out following *Sma*I macrorestriction as previously described (Stanley *et al*., 1996). For *Sfi*I and *Ngo*AIV macrorestriction, different ramping and electrophoresis conditions were used, as follows: 10–90 s at 5·5 V cm^{-1}, 24 h, 1·3% agarose for *Sfi*I; 0·1–30 s at 5·7 V cm^{-1}, 23 h, 1·1% agarose for *Ngo*AIV. Genetic relationship between isolates was estimated by using the equation of Nei & Li (1979) to calculate *D* values (distance matrices) for all three enzymes. Estimates of overall restriction site similarity were then used to construct an unrooted tree by the FITCH option of the PHYLIP computer package (Felsenstein, 1988).

Minipreparation of genomic DNA, 16S ribotyping, and exotoxin gene carriage. Genomic DNA was extracted from streptococcal plate cultures as previously described (Stanley *et al*., 1995). DNA was digested with *Xho*I, *Eco*RI or *Sac*I, electrophoresed, blotted, and hybridized with a 1500 bp *S. pyogenes* 16S rRNA gene probe as previously described (Stanley *et al*., 1995). Membrane filters were developed colorimetrically, and scanned directly with a ScanMaker IIG (Microtek Lab) into a Power Macintosh 6100/60 (Apple Computer). PCR primers and conditions used for detection of exotoxin genes were as previously described (Stanley *et al*., 1996).

IS*1239* profiling. An 865 bp fragment of IS*1239* was amplified by PCR from genomic DNA. The final reaction mixture (50 μl) contained 5 μl standard PCR buffer (Life Technologies), 3·0 mM magnesium chloride, 200 μmol (each) deoxynucleotide triphosphates, 0·6 μM of each primer and 2·5 units of *Taq*

Table 1. *Streptococcus pyogenes* M5 isolates and their genotypes

Isolate	Year/source (all UK) of isolation	Disease	*emm* gene PCR-RFLP*	Combined PFGE profile†	Combined 16S ribotype‡	*spe* genes§	IS*1239* profile‖
R0443	1991/Surrey	Scarlet fever	5.H2	#5.1	R-4	ABC	IP3
R0060	1995/Sussex	Pneumonia	5.H2	#5.1	R-4	ABC	IP3
R0065	1995/London	Cellulitis	5.H2	#5.1	R-4	ABC	IP3
R0439	1991/Surrey	Scarlet fever	5.H2	#5.1	R-4	ABC	IP3
R2275	1994/London	Brain abscess	5.H2	#5.1	R-4	ABC	IP3
R1508	1991/Edinburgh	Rheumatic fever	5.H11	#5.1	R-4	ABC	IP3
R0116	1995/London	Scarlet fever	5.H3	#5.1	R-4	ABC	IP3
R1349	1995/Salisbury	Bursitis	5.H8	#5.1	R-4	ABC	IP3
R1628	1995/Isle of Wight	Persistent discharge	5.H8	#5.1	R-4	ABC	IP3
R2454	1994/Llanelli	Conjunctivitis	5.H8	#5.2	R-4	ABC	IP3
R2577	1994/Hampton	Septicaemia	5.H5	#5.2	R-4	ABC	IP3
R2606	1995/Worcester	Unknown	5.H7	#5.3a	R-5	BC	IP3
R0126	1995/Blackburn	Fatal septicaemia	5.H7	#5.3	R-5	ABC	IP3
R0304	1995/Tyneside	Septicaemia	5.H2	#5.3	R-5	BC	IP3
R0428	1995/Truro	Postnatal pyrexia	5.H14	#5.3	R-5	AB	IP3
R2356	1994/Carmarthen	Cellulitis	5.H3	#5.4	R-5	ABC	IP3
R2644	1994/Llanelli	Pyrexia	5.H3	#5.4	R-5	ABC	IP3
R1635	1995/Carlisle	Septicaemia	5.H3	#5.5	R-4	ABC	IP3
R0379	1995/Barnsley	Sore throat	5.H3	#5.5	R-4	ABC	IP3
R0161	1995/Sheffield	Lymphadenitis	5.H6	#5.5	R-4	ABC	IP3
R2785	1994/Epsom	None	5.H9	#5.5	R-8	ABC	IP3
R1388	1995/Exeter	Ear infection	5.H3	#5.6	R-4	ABC	IP3
R1648	1970/Unknown	Unknown	5.H12	#5.7	R-9	AB	IP3
R0022	1995/Nottingham	Unknown	5.H2	#5.8	R-4	ABC	IP3
R0536	1991/London	Pharyngitis	5.H2	#5.8	R-4	ABC	IP3
R2919	1991/Exeter	Jaundice	5.H10	#5.8	R-4	ABC	IP3
R0009	1995/Llanelli	Vaginitis	5.H3	#5.8	R-4	ABC	IP3
R1973	1991/Luton	Septicaemia	5.H3	#5.8	R-4	ABC	IP3
R0701	1991/Luton	Septicaemia	5.H3	#5.8	R-4	ABC	IP3
R2264	1994/Isle of Wight	Post partum	5.H4	#5.8	R-4	ABC	IP3
R0574	1994/Stroud	Unknown	5.H18	#5.9	R-10	B	IP5
R2357	1994/London	Chest infection	5.H18	#5.10	R-7	BC	–
R2160	1994/London	Vaginitis	5.H15 (NC)	#5.11	R-6	BC	IP4
R2223	1994/Stoke-on-Trent	Pneumonia	5.H16 (NC)	#5.12	R-3	BC	IP2
R2247	1994/Dorchester	Bacteraemia	5.H17	#5.13	R-2	B	–
R0581	1995/Romford	Wound infection	5.H13	#5.14	R-5	BC	IP3
R2550	1994/Grimsby	Renal disease	5.H3	#5.14	R-5	BC	IP3
R0307	1995/Bristol	Septicaemia	5.H2	#5.14	R-5	BC	IP3
100065	1959/UK	Unknown	5.H1	#5.15	R-1	BC	IP1
NCTC 8193	1950/London	Puerperal fever	5.H1	#5.15	R-1	AB	IP1

* *emm* gene PCR amplification followed by RFLP analysis with *Hae*III; NC, not cut by *Hae*III.

† Numbers following the combined PFGE profile, #, indicate a unique RFLP obtained by combining three endonuclease digestions (*Sma*I, *Sfi*I, *Ngo*AIV).

‡ Numbers following the combined ribotype (with *Eco*RI and *Sac*I), R, indicate different patterns of hybridization with the probe.

§ Streptococcal pyrogenic exotoxin genes *A*, *B*, *C* determined by PCR.

‖ Numbers following the insertion sequence IS*1239* profile, IP, indicate different *Pvu*II patterns of hybridization with the probe; –, absence of the element.

polymerase. Samples overlaid with 100 μl of mineral oil were subjected to 25 cycles of denaturation at 94 °C for 1 min, annealing at 51 °C for 1 min, and extension at 72 °C for 1 min in a Robocycler (Stratagene). PCR products were labelled with biotin-16-dUTP using a random-primed labelling kit (Boehringer Mannheim). Ten micrograms of genomic DNA was digested with *Pvu*II, electrophoresed in 0·7% agarose (55 V, 16 h), Southern blotted and hybridized. Stringent filter wash-

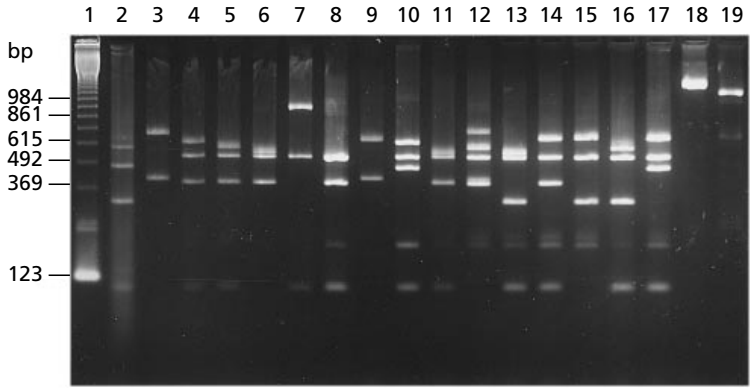

Fig. 1. PCR-RFLP subtypes (*Hae*III) of *emm* gene amplicons. Lane 1, 123 bp ladder (size markers); lane 2, H1; lane 3, H17; lanes 4–8, H2–H6; lane 9, H18; lanes 10–17, H7–H14; lanes 18 and 19, H15 and H16 (amplicons not digested by *Hae*III).

ing conditions (two washes, 30 min each, 64 °C, 0·16 × SSC/0·1 % SDS) were employed, and detection of hybridized probe was as previously described (Stanley *et al*., 1995).

RESULTS

emm gene polymorphism and sequence analysis

emm gene PCR amplicons from all M5 isolates varied in size from approximately 1050 to 2100 bp. Analysis for *Hae*III polymorphisms detected 18 RFLPs. Two amplicons (H17 and H18) were not digested (Fig. 1, lanes 18 and 19). Of the 40 isolates, 23 % and 25 %, respectively, belonged to two RFLP groups, H2 and H3 (Fig. 1, lanes 4 and 5). The 16 remaining RFLP groups consisted of one to three strains each (Table 1; Fig. 1).

We sequenced 240 bases from the 5′ region of the 18 amplicons (Fig. 1). These sequences were aligned using MegAlign module of Lasergene (DNASTAR). Fourteen of them (Fig. 2a) exhibited >94 % identity to that of the serotype M5 type strain, NCTC 8193, whose sequence was termed *emm*5.H1, and which was identical to that previously published (Whatmore & Kehoe, 1994). Four amplicons (termed *emm**) exhibited marked divergence (Fig. 2b). Nonetheless, all amplicons contained the nucleotide sequence 5′-TCGCTTAGAAAATTAA-3′, which is complementary to the forward primer designed to distinguish sequences in the 5′ conserved regions of *emm* genes from corresponding regions of *enn* or *mrp* genes (Whatmore *et al*., 1994; Whatmore & Kehoe, 1994). Sequences of the four *emm** amplicons were found to be highly similar (>98 %) to those of *emm* genes other than *emm*5 (GenBank data). The *emm** amplicons originated from strains R2247 (sequence identity with *emm*77), R2223 (*emm*6), R2160 (*emm*18) and R2357 (*emm*11).

Further analysis of M5 strains with *emm** sequences

PCR primers to amplify the 5′ region specific to a particular *emm** gene were designed for each strain, and also for the type strain, using first-strand sequence data. The reverse primers were: strain R2223, 5′-ATCTGTTAAGTTTTTATTC-3′; strain R2357, 5′-ATTTTTTGTCTCTTCATTTT-3′; strain R2160, 5′-AGTTTTTAAATCATCATTC-3′; strain R2247, 5′-TTCTGATTTTTTTTTCAAG-3′, and M5 type strain, 5′-AACTCAGCAGTCTTACGTTC-3′. When used with the forward primer 1 (Whatmore & Kehoe, 1994), products varying in size from 250 bp to 300 bp were amplified, and used as probes.

The resulting slot-blot and Southern blot hybridization data (Table 2) showed that these *emm**-specific amplicons hybridized with their genomic DNA of origin in all cases. The *emm*5-amplicon hybridized only with itself and conversely, none of the *emm** amplicons hybridized with the type strain. The *emm** amplicons from strains R2223 and R2160 hybridized only with their own genomic DNA. *emm** amplicons from strains R2247 and R2357 also cross-hybridized with genomic DNA of each other.

Further serological characterization was carried out for the four strains above, whose *emm** sequences corresponded to *emm*6 (strain R2223), *emm*11 (strain R2357), *emm*18 (strain R2160) and *emm*77 (strain R2247). As seen in Fig. 3, M5 precipitin lines of identity were seen with Lancefield extracts prepared from these four strains. However, lines of identity were also seen for extracts of these strains with anti-M6 (strain R2223), anti-M11 (strain R2357), anti-M18 (strain R2160) and anti-M77 (strain R2247).

Pyrogenic exotoxin determinants

A PCR amplicon from the *speB* gene was generated from all isolates. Five per cent of the isolates carried only this exotoxin determinant, whereas 65 % carried *speA*, *B* and *C* genes. Among the remaining isolates, three carried *speAB* and nine carried *speBC* (Table 1).

Macrorestriction (PFGE) profiles

Macrorestriction was carried out serially with *Sma*I, *Sfi*I or *Ngo*AIV, whose restriction sites are GC-rich. Of 14 *Sma*I profiles detected, 33 % of isolates shared one, 18 % another, and eight isolates had unique profiles. Of 13 *Sfi*I profiles, none was predominant: three profiles accounted for 28 %, 18 % and 15 % of isolates, and seven were strain-specific. Of 13 *Ngo*AIV profiles detected, one

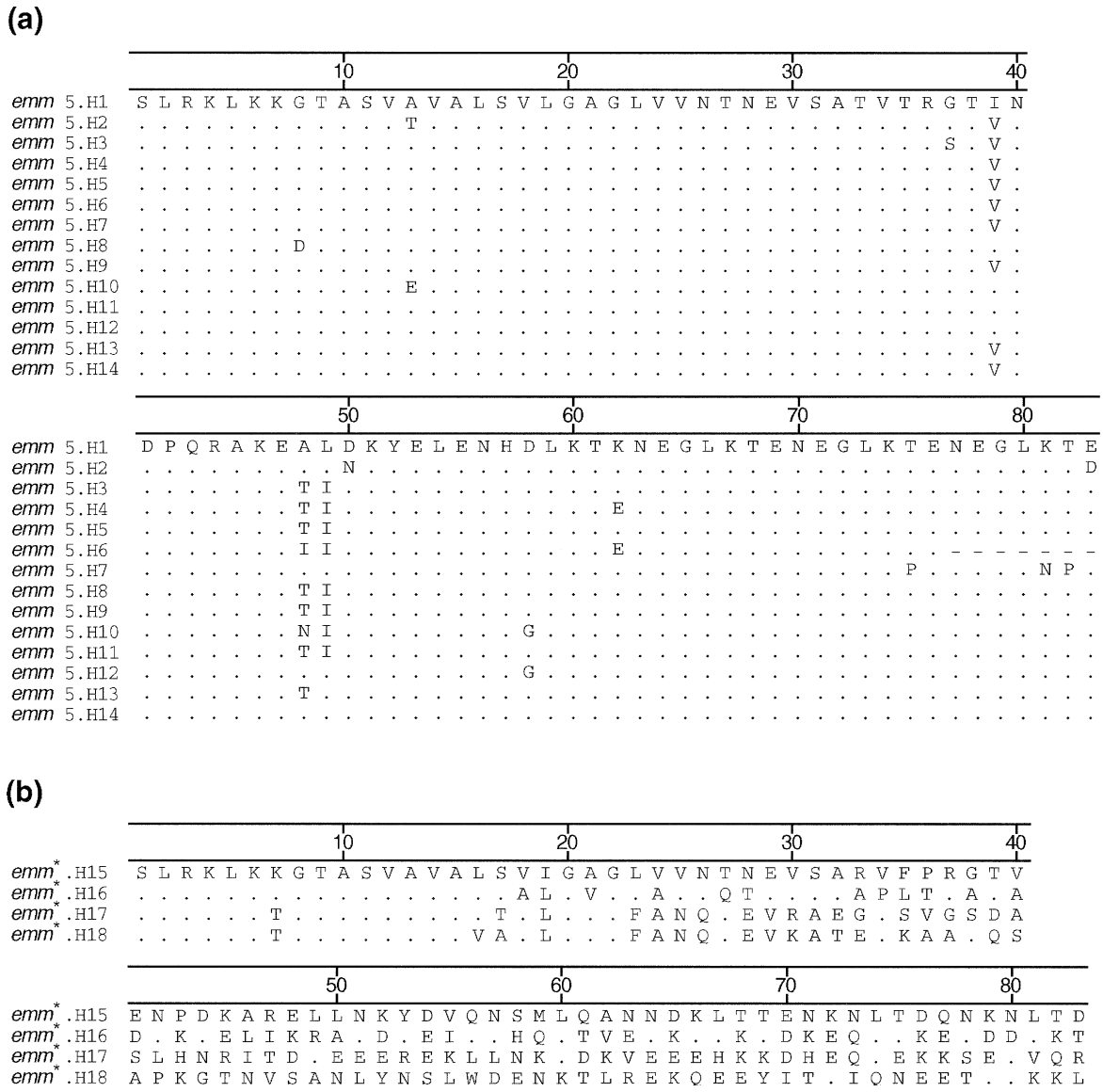

Fig. 2. Deduced amino acid sequences corresponding to the N-terminal regions of *emm* subtypes (cf. Fig. 1). The dots indicate positions with identity to the first sequence in each panel. (a) Sequences of 13 *emm*5 subtypes exhibiting >95% sequence identity to that of the type strain NCTC 8193 (*emm*5.H1; GenBank accession number SP02480). In *emm*5.H6 (GenBank accession number AJ223285), dashes represent probable deletion of a 7-amino-acid (A) repeat. (b) Sequences of four subtypes showing marked divergence, and consequently labelled *emm** (GenBank accession numbers AJ223286–AJ223289).

accounted for 53% of isolates, and seven were strain-specific (data not shown).

Individual profiles obtained with three endonucleases were combined to give 15 'combined profiles' (Table 1). For example, serotype M5 isolates with the *Sma*I profile 8, *Sfi*I profile 4 and *Ngo*AIV profile 4 were designated 'combined profile' #5.1. Similarly, combined profile #5.2 represents the combination *Sma*I profile 8, *Sfi*I profile 4, *Ngo*AIV profile 6 and so forth. Each designated 'combined profile' was differentiated by at least three band differences with at least one endonuclease. Less than three band differences with any endonuclease was taken as indicative of clonal relationship (Tenover *et al*., 1995) and designated by a suffix letter. Thus, #5.3 was related to #5.3a by less than three band differences in the *Sma*I profile.

16S ribotypes

Genomic DNA was digested with *Eco*RI, *Sac*I or *Xho*I. Among the 40 isolates, polymorphism was extensive with all enzymes. Seven RFLPs were detected in both *Xho*I digests and *Eco*RI digests, with 34/40 (85%) isolates sharing a single ribotype. Of the remaining six isolates, five belonged to four 'atypical' strains. The

Table 2. Summary of results of slot-blot and Southern blot hybridizations

emm probe†	Genomic DNA from strain:				
	NCTC 8193	R2247	R2223	R2160	R2357
emm5	+	−	−	−	−
*emm** (77)	−	+	−	−	+
*emm** (6)	−	−	+	−	−
*emm** (18)	−	−	−	+	−
*emm** (11)	−	+	−	−	+

† Approximately 250 bp amplicon representing the 5′ end of the *emm* gene amplified from the strain shown. The number in parentheses indicates the *emm* gene to which the respective *emm** amplicon corresponds.

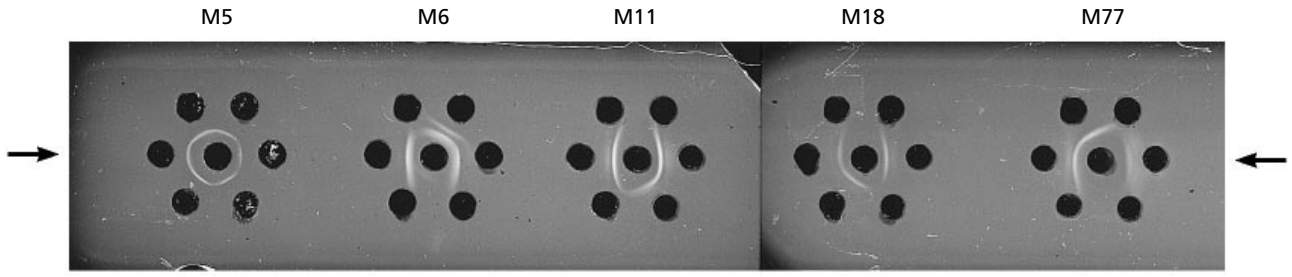

Fig. 3. M-precipitin reactions of strains with *emm** gene amplicons. The central well within each set contains antiserum to serotypes M5, M6, M11, M18 or M77 as indicated above the photograph. The outer wells of each set contained Lancefield acid extracts of test and control strains. Wells 1 and 4 of each set (arrowed) contained homologous control extract for the M type. Wells 2, 3, 5 and 6 (reading clockwise from the arrow on the left) contained acid extracts of strains R2247, R2223, R2160 and R2357. In each case there is reaction with two wells containing the homologous control extracts, and with one other well: M6 with well 3, M11 with well 5, M18 with well 6 and M77 with well 2.

most discriminatory enzyme, *Sac*I, detected ten ribotypes among the isolates. A 'combined ribotype' was derived by sequential addition of data from these enzymes. Ten 'combined ribotypes' were defined; seven of these occurred only in single strains (Table 1).

IS*1239* profiling

PCR primers were designed with Lasergene (DNASTAR) software, from the published sequence of IS*1239*. The forward primer 5′-ATGCAAGATTATTATACACC-3′ and reverse primer 5′-TCTTTAGGGGTCGTTGTTT-TGGTA-3′ generated an 865 bp product, corresponding to an internal fragment of IS*1239*. This amplicon was used to probe genomic Southern blots made with *Pvu*II, which has a single restriction site lying 146 bp downstream of the forward PCR primer. The rationale was that each genomic copy of IS*1239* should yield two hybridizing *Pvu*II fragments ('IS*1239* bands').

Thirty-eight of the 40 M5 isolates carried IS*1239*. Five IS*1239* profiles were found. The number of IS*1239* bands varied from 2 to 24, and their sizes varied from 0·7 to 12·0 kbp. The type and reference strains of serotype M5 shared profile IP1 (Fig. 4, lanes 2 and 5), while 33 clinical isolates shared a closely related IS*1239* profile, termed IP3 (Fig. 4, lane 1). Of the five remaining isolates (all containing *emm** genes – see above), three had unique profiles (Fig. 4, lanes 4, 6 and 7), and two lacked the element (Fig. 4, lane 3).

Analysis of genetic relationships between strains

Distance matrices for the three PFGE endonucleases were added, and used to construct a 'combined profile' dendrogram (Fig. 5). Ribotype data (Table 1) and IS profiling were concordant with this dendrogram, except for one cluster of strains (bracketed in Fig. 5) – those from which *emm** amplicons were generated. A striking feature of the tree is the very high level of genetic diversity within serotype M5. There were 15 strain clusters within only 40 randomly chosen clinical isolates. Some of these (e.g. R0443 to R1628 and R0022 to R0701, Fig. 5) shared the same ribotype and *spe* genes, others (e.g. R1026 to R0428) shared the same ribotype but were diverse in *spe* genes, while yet others were diverse with respect to both ribotype and *spe* genes (Table 1). Even if two isolates had identical PFGE profiles with all three endonucleases, they often differed at other genetic markers. For example, four *emm5* gene subtypes occurred among the otherwise identical strain cluster R0022 to R0701. Some strains within a PFGE cluster had the same *emm5* gene subtype, while others within the

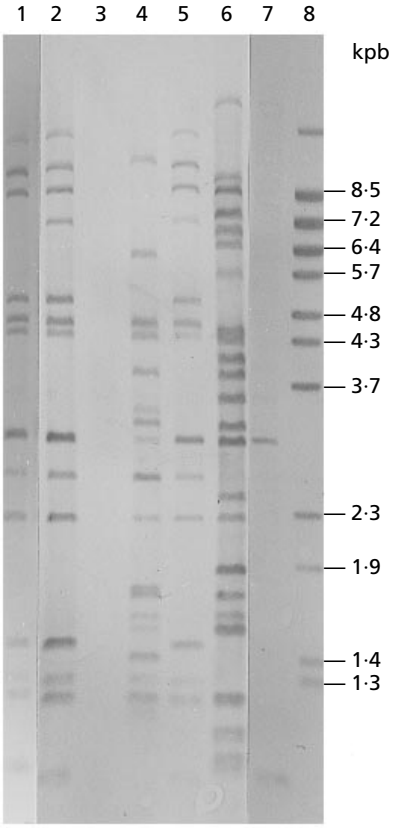

Fig. 4. IS*1239* profiles (IP) of serotype M5 strains. Genomic Southern blot made with *Pvu*II and hybridized with 865 bp internal fragment of IS*1239*. Lane 1, IP3; lanes 2 and 5, IP1; lane 3, isolate lacking the insertion element; lanes 4, 6 and 7, IP2, IP4 and IP5; lane 8, bacteriophage lambda DNA digested with *Bst*EII (size markers).

same PFGE cluster had different *emm5* gene subtypes (e.g. R0443 to R1628 and R0022 to R0701, Fig. 5). The type and reference strains, NCTC 8193 and 100065, were genetically dissimilar to contemporary M5 strains (consistent with their related but distinct IS*1239* profiles – see above).

The strain cluster bracketed in Fig. 5 was exceptional and is placed in the tree provisionally (dashed lines), only on the basis of macrorestriction. The strains in this cluster, which were diverse with respect to all genetic markers, did not share the IS*1239* profile common to other M5 strains: two lacked the element, while three had unique and unrelated IS*1239* profiles (Fig. 4, lanes 4, 6, 7). Their ribotypes were divergent from each other, and from other isolates in the study. Most notably, divergent *emm** amplicons identified by nucleotide sequencing (see above) were generated from them.

DISCUSSION

Among a relatively small set of serotype M5 isolates, the *emm* gene was found to be extensively polymorphic. Sequence analysis confirmed that 14 of the 18 subtypes detected by PCR-RFLP were readily recognizable as alleles of *emm5* (Fig. 2a). However, the four *emm** subtypes had highly divergent nucleotide sequences (Fig. 2b), and these corresponded to other known *emm* genes: *emm*6, *emm*11, *emm*18 and *emm*77.

With respect to molecular epidemiology, *emm* gene polymorphism alone was extensive enough to differentiate within PFGE profile. PCR-RFLP thus offers a straightforward method of subtyping, and it is shown here to be valid by parallel nucleotide sequence analysis. The genotyping methods reported showed a general hierarchical relationship: this is exemplified by congruence of the IS*1239* profile, ribotype and PFGE profile in defining strains and clones within serotype M5 (see Table 1). The *speA* and *speC* genes were unlinked to other markers, as expected for bacteriophage-encoded determinants (Yu & Ferretti, 1991). The type and reference M5 strains had an IS*1239* profile (Fig. 4, lanes 2 and 5) which differed by a single IS*1239* band from the predominant profile found in contemporary M5 strains (Fig. 4, lane 1). Despite the extensive diversity among other markers, a common ancestry for most M5 strains is indicated by these related IS*1239* profiles, IP1 and IP3. The utility of IS*1239* in the present study parallels that of other insertion sequences in establishing epidemiological clonality among clinical isolates of *Salmonella* and *Mycobacterium tuberculosis* (Stanley & Saunders, 1996).

Given that these IS*1239* profiles may be typical for serotype M5, it is also noteworthy that the four strains carrying *emm** genes exhibited unrelated IS profiles or did not even harbour the element. Furthermore, *Eco*RI ribotypes for strains R2223 and R2357 matched those previously described for strains of serotypes M6 and M11, corresponding to their *emm* gene sequences (Stanley *et al*., 1995) – no such comparative data were available for the other two strains carrying *emm** genes.

These four strains were originally (and repeatedly) serotyped as M5 according to classic serological methods (Johnson *et al*., 1996). According to the Lancefield scheme, they would not have been tested against individual M antisera for types M6, M11, M18 or M77. This is because the first stage of serotyping comprises the T agglutinin reaction; in this case the four strains are T5/27/44. The next stage is the opacity factor (OF) reaction; all four strains are OF^{-}. Isolates are then screened against M antisera defined by the scheme as related – in this case M5 and M12. The scheme does not recognize serotypes M6, M11, M18 or M77 to be related to isolates carrying the T5/27/44 complex antigens (Johnson *et al*., 1996).

Phenotypically, the Lancefield serological identity of these isolates is M5. However, sequence analysis of both strands of 5′ hypervariable region of their *emm** amplicons demonstrated identity to *emm* genes other than *emm5*. Hybridization studies indicated that the *emm5* gene was not present in these four strains; it is therefore possible that their M5 serodeterminant is not the hypervariable region of their *emm* gene. In summary,

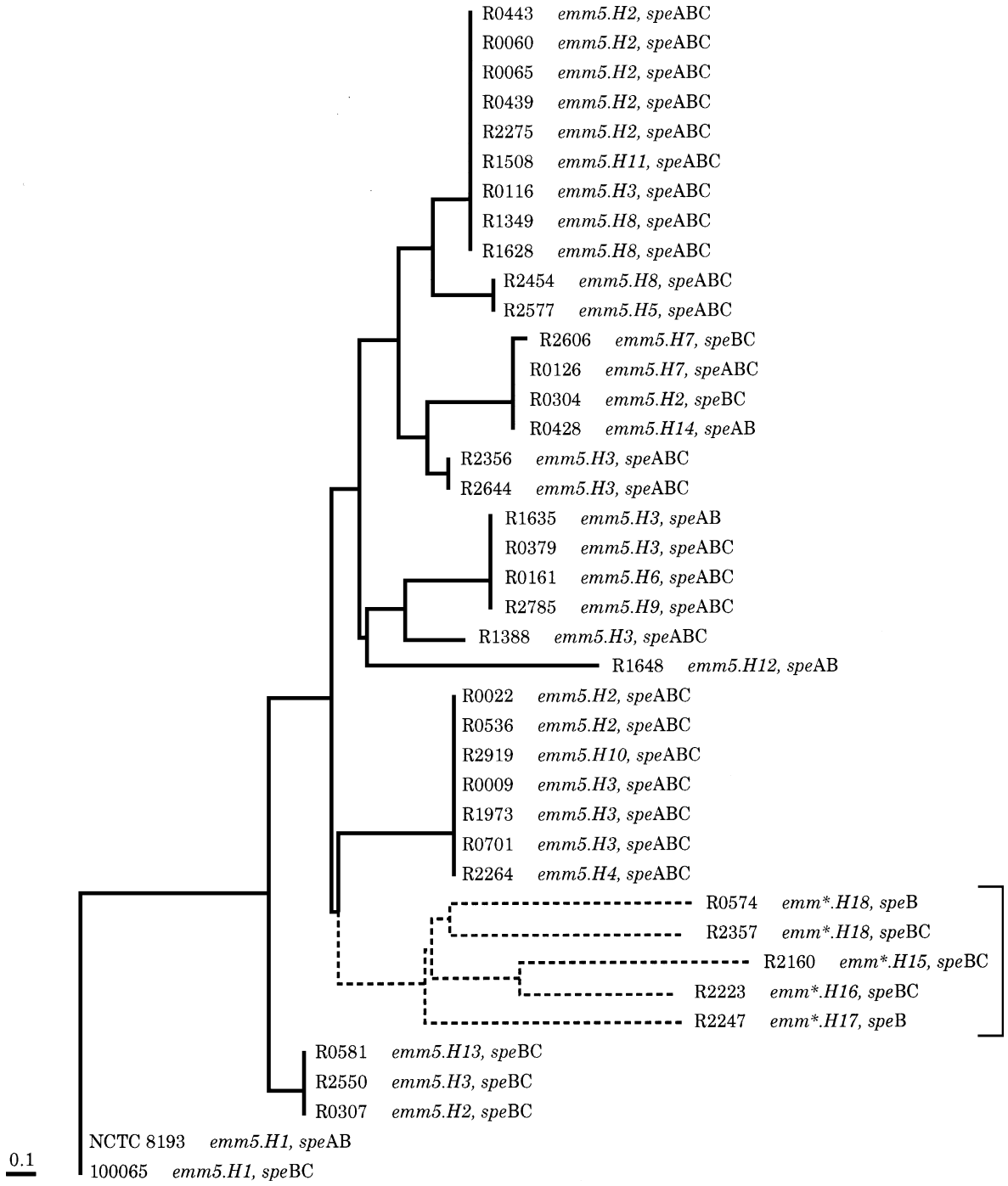

Fig. 5. Dendrogram showing relationships between serotype M5 isolates, inferred (following PFGE) from overall restriction site similarities estimated by the equation of Nei & Li (1979). These were used to construct an unrooted tree by the FITCH option of the PHYLIP package. Respective *emm* subtypes and *spe* determinants are shown next to the strain numbers. The cluster of strains (bracketed, with branching shown by dashed lines) with *emm** genes are placed here provisionally – other markers would suggest they are unrelated.

these strains show no collinearity between Lancefield serotype as determined by classical methods and *emm* gene hypervariable region sequence, or, for that matter, between Lancefield serotype and overall genotype (Fig. 5). Again, M77 and M11 are considered classically to be OF^+ serotypes, so that this phenotypic association is also previously unreported. These observations are important since the strains concerned represent over 10% of M5 isolates in this study. Our results further extend the evidence provided by Beall *et al*. (1997) that

classical M and T antigen associations are not always consistent with *emm* gene sequence data for a minority of GAS strains.

Data presented in this report will be relevant to the methodology and design of epidemiological studies, and to any debate about the most appropriate way of M typing group A streptococci. Beall *et al*. (1997) have commented that supplementation of Lancefield serotyping with *emm* gene sequencing greatly improves the efficiency and accuracy of epidemiological studies. Results presented in this report support and extend this concept. We have also shown that *emm* gene PCR-RFLP is a rapid and cost-effective way of identifying *emm* sequence type, and such data are congruent with chromosomal genotype as measured by ribotype, IS*1239* profile and PFGE profile. It should also be noted that recombinant DNA strategies for M protein vaccine development will need to take into account the extensive *emm* gene polymorphism found within important serotypes like M5.

ACKNOWLEDGEMENTS

We thank Jacqueline Xerry and Mark Broughton for technical assistance with ribotyping and radiolabelling experiments, and Philip Mortimer for critically reading the manuscript.

REFERENCES

Beall, B., Facklam, R. & Thompson, T. (1996). Sequencing *emm*-specific PCR products for routine and accurate typing of group A streptococci. *J Clin Microbiol* **34**, 953–958.

Beall, B., Facklam, R., Hoenes, T. & Schwartz, B. (1997). Survey of *emm* gene sequences and T-antigen types from systemic *Streptococcus pyogenes* infection isolates collected in San Francisco, California; Atlanta, Georgia; and Connecticut in 1994 and 1995. *J Clin Microbiol* **35**, 1231–1235.

Cleary, P. P., Kaplan, E. L., Handley, J. P., Wlazlo, A., Kim, M. H., Hauser, A. R. & Schlievert, P. M. (1992). Clonal basis for resurgence of serious *Streptococcus pyogenes* disease in the 1980s. *Lancet* **339**, 518–521.

Feinberg, A. P. & Vogelstein, B. (1983). A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal Biochem* **132**, 6–13.

Felsenstein, J. (1988). PHYLIP: phylogenetic inference package. Version 3.0. Seattle: University of Washington.

Fischetti, V. A. (1989). Streptococcal M protein: molecular design and biological behavior. *Clin Microbiol Rev* **2**, 285–314.

Johnson, D. R., Kaplan, E. L., Sramek, J., Bicova, R., Havlicek, J., Havlickova, H., Motlova, J. & Kriz, P. (1996). *Laboratory Diagnosis of Group A Streptococcal Infections*. Geneva: World Health Organization.

Kapur, V., Reda, K. B., Li, L. L., Ho, L. J., Rich, R. R. & Musser, J. M. (1994). Characterization and distribution of insertion sequence IS*1239* in *Streptococcus pyogenes*. *Gene* **150**, 135–140.

Lancefield, R. C. (1962). Current knowledge of type-specific M antigens of group A streptococci. *J Immunol* **89**, 307–313.

Musser, J. M., Hauser, A. R., Kim, M. H., Schlievert, P. M., Nelson, K. & Selander, R. K. (1991). *Streptococcus pyogenes* causing toxic-shock-like syndrome and other invasive diseases: clonal diversity and pyrogenic exotoxin expression. *Proc Natl Acad Sci USA* **88**, 2668–2672.

Musser, J. M., Kapur, V., Kanjilal, S. & 11 other authors (1993). 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* **167**, 337–346.

Nei, M. & Li, W. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci USA* **76**, 5269–5273.

Podbielski, A., Melzer, B. & Lutticken, R. (1991). Application of the polymerase chain reaction to study the M protein (-like) gene family in beta-hemolytic streptococci. *Med Microbiol Immunol* **180**, 213–227.

Stanley, J. & Saunders, N. (1996). DNA insertion sequences and the molecular epidemiology of *Salmonella* and *Mycobacterium*. *J Med Microbiol* **45**, 236–251.

Stanley, J., Linton, D., Desai, M., Efstratiou, A. & George, R. (1995). Molecular subtyping of prevalent M serotypes of *Streptococcus pyogenes* causing invasive disease. *J Clin Microbiol* **33**, 2850–2855.

Stanley, J., Desai, M., Xerry, J., Tanna, A., Efstratiou, A. & George, R. (1996). High-resolution genotyping elucidates the epidemiology of Group A streptococcal outbreaks. *J Infect Dis* **174**, 500–506.

Tenover, F. C., Arbeit, R. D., Goering, R. V., Mickelsen, P. A., Murray, B. E., Persing, D. H. & Swaminathan, B. (1995). Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol* **33**, 2233–2239.

Upton, M., Carter, P. E., Orange, G. & Pennington, T. H. (1996). Genetic heterogeneity of M type 3 group A streptococci causing severe infections in Tayside, Scotland. *J Clin Microbiol* **34**, 196–198.

Whatmore, A. M. & Kehoe, M. A. (1994). Horizontal gene transfer in the evolution of group A streptococcal *emm*-like genes: gene mosaics and variation in Vir regulons. *Mol Microbiol* **11**, 363–374.

Whatmore, A. M., Kapur, V., Sullivan, D. J., Musser, J. M. & Kehoe, M. A. (1994). Non-congruent relationships between variation in *emm* gene sequences and the population genetic structure of group A streptococci. *Mol Microbiol* **14**, 619–631.

Yu, C. E. & Ferretti, J. J. (1991). Molecular characterization of new group A streptococcal bacteriophages containing the gene for streptococcal erythrogenic toxin A (*speA*). *Mol Gen Genet* **231**, 161–168.

Received 23 May 1997; revised 4 September 1997; accepted 12 November 1997.