Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 3280–3284, April 1994
Microbiology

Allelic polymorphism of *emm* loci provides evidence for horizontal gene spread in group A streptococci

(bacterial surface proteins/genetic recombination)

DEBRA E. BESSEN*† AND SUSAN K. HOLLINGSHEAD‡

*Department of Epidemiology and Public Health (Microbiology Section), Yale University School of Medicine, New Haven, CT 06510; and ‡Department of Microbiology, University of Alabama, Birmingham, AL 35294

Communicated by Dorothy M. Horstmann, December 13, 1993

ABSTRACT **Group A streptococci have a virulence regulon containing a single *emm* locus or two or three distinct and adjacent loci of structurally related *emm* family genes. The products of the *emm* gene cluster consist of fibrillar surface proteins, at least some of which are known to contain determinants of type specificity located in their NH_2-terminal regions, lying distal to the cell surface. The *emm* genes can be categorized into four major subfamilies (SFs), based on structural differences within their 3′ regions encoding the peptidoglycan-spanning domain. In this study, we investigate the polymorphism within the 5′ region of SF-4 and SF-3 *emm* genes (which occupy the first and last *emm* positions of the gene cluster, respectively) in 22 strains representing different serotypes. Our findings indicate that unlike the centrally positioned SF-1 or SF-2 genes, SF-3 and SF-4 genes each display only limited polymorphism in their 5′ regions, suggesting that their gene products may not be major contributors to type specificity. Two forms of the SF-3 gene ($SF3^a$, $SF3^b$) and two forms of the SF-4 gene ($SF4^a$, $SF4^b$) are found to exist in all four possible combinations ($SF3^aSF4^a$, $SF3^aSF4^b$, $SF3^bSF4^a$, $SF3^bSF4^b$), strongly suggesting that horizontal gene spread has contributed to the evolution of *emm* genes and to the generation of *emm* gene diversity in group A streptococci.**

Group A streptococci are important human pathogens capable of causing a wide variety of diseases, including pharyngits, impetigo, a toxic shock-like syndrome, rheumatic fever, and acute glomerulonephritis. The *emm* gene cluster lies within a virulence regulon of the streptococcal chromosome and consists of one to three distinct but structurally related genes arranged in tandem (1–5). At least some *emm* family genes are known to encode M protein, a fibrillar surface molecule that plays a key role in virulence by virtue of its antiphagocytic property (6). In addition, the 5′ region of at least some *emm* family genes is known to encode the determinants of serological type, of which >80 exist; only antibodies directed to the type-specific immunodeterminants neutralize the antiphagocytic effect of M protein and are thereby protective (6). Some *emm* family genes encode surface proteins having nonimmune binding activity for IgA and/or IgG; the role of immunoglobulin binding in pathogenesis remains unknown (3, 4, 7–10). Thus, *emm* genes and their products are structurally and functionally diverse, and it has been hypothesized that this heterogeneity defines, in part, the pathogenic potential of the individual strain.

emm family genes contain relatively conserved 3′ halves, encoding (*i*) either a surface-exposed A or C repeat region, which consists of tandemly arranged blocks of direct sequence repeats, (*ii*) an α-helical-rich cell wall-associated region, (*iii*) one of four distinct peptidoglycan-spanning domains, and (*iv*) a hydrophobic membrane-spanning domain at the 3′ terminus (3, 4, 6, 7, 11). In contrast, the extreme 5′ end of some *emm* genes encodes the NH_2-terminal type-specific immunodeterminants; also located within the NH_2-terminal half of *emm* gene products are semiconserved domains distributed among a limited number of *emm* genes derived from strains of distinct serotypes (3, 6, 12). Based on these structural parameters, *emm* genes are considered to constitute a gene family [note that the term "*emm*" is used herein for genes previously referred to as *emmL, fcrA, enn, arp, mrp*, and *sir* (1)]. *emm* genes can be categorized into four subfamilies (SFs: SF-1, SF-2, SF-3, and SF-4) on the basis of nucleotide sequence differences within their 3′ ends (1). The number and order of the SF *emm* genes reveal five major chromosomal patterns for the *emm* gene cluster (1, 13). For example, chromosomal pattern 4 consists of an SF-4, SF-1, and SF-3 *emm* gene (in that order), whereas pattern 5 is comprised of an SF-4, SF-2, and SF3 *emm* gene (Fig. 1) (1). Group A streptococci can also be divided into two major classes (I and II), defined in part by antigenic epitopes present within the surface-exposed C repeat regions of SF-1, SF-2, and SF-3 *emm* gene products (12, 14). The *emm* chromosomal patterns correlate with streptococcal class (1, 2), which in turn has several correlates with pathogenicity (10, 12).

The evolution of *emm* genes and the generation of *emm* gene diversity likely involve several different genetic mechanisms. Intragenic recombination between repetitive sequence blocks leads to deletion or duplication of blocks and can also introduce point mutations that may contribute to antigenic diversity and changes in M protein function (15, 16). Gene duplication and subsequent divergence have been proposed to explain the existence of the multiple *emm* gene SFs (7, 13). The nature of the *emm* chromosomal patterns observed in naturally occurring isolates suggests that recombinational exchange of SF-specific *emm* genes from one site within the *emm* chromosomal region to another site is rare or that if such recombinants do occur, they are at a selective disadvantage for survival (1). However, a role for homologous recombination within the *emm* gene cluster following horizontal transfer of DNA from an isolate of a different strain has not yet been reported.

Here we provide evidence for the horizontal spread of *emm* genes or portions of *emm* genes. At least one of the domains that may be involved in this recombinational exchange encodes an IgA-binding site (21). In addition, we demonstrate that there is only a limited amount of sequence diversity at the 5′ ends of SF-4 genes derived from different serotypes, suggesting that the majority of SF-4 genes do not contribute to type specificity. Similarly, many SF-3 *emm* genes display only limited polymorphism at their 5′ ends, although the degree of polymorphism for the 5′ end of SF-3 *emm* genes

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: SF, subfamily.
†To whom reprint requests should be addressed.

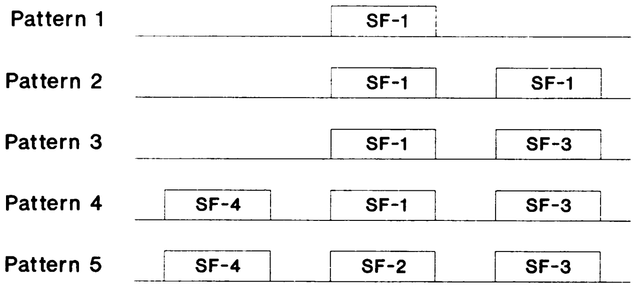

FIG. 1. Arrangement of SF-4 and SF-3 *emm* family genes within the *emm* chromosomal region. Chromosomal patterns for *emm* gene clusters containing SF-4 and/or SF-3 genes are shown (patterns 3–5); chromosomal patterns are established based on SF-specific sequences, which encode the peptidoglycan-spanning domains located near the 3′ terminus of each *emm* family gene (1). Chromosomal patterns 1 and 2 are shown for comparison. Class I isolates are represented by patterns 1–4, whereas pattern 5 isolates are class II (1). For most isolates, the *emm* family gene coding regions are ≈1.1–1.2 kb in length, and the intergenic noncoding regions are ≈0.2 kb (1).

appears to be intermediate to that of the SF-4 *emm* genes and the SF-1 and SF-2 *emm* genes.

MATERIALS AND METHODS

Bacterial Strains. The strains, sources, class, and *emm* chromosomal patterns of all group A streptococci used in this study have been described (1) and are designated according to their M protein serotype. The *emm* chromosomal patterns (Fig. 1) for individual isolates are as follows: M types 3, 6, 19, and 24 (pattern 1); M types 1 and 5 (pattern 2); M types 18 and 23 (pattern 3); M types 30, 32, 33, 36, 52, 53, and 56 (pattern 4); M types 2, 4, 11, 13, 15, 22, 25, 46, 49, 58, 60, 65, and 76 (pattern 5).

DNA Sequencing. Portions of *emm* genes were generated by PCR amplification and cloned into pT7Blue(U) T vectors for nucleotide sequence determination. For SF-4 genes, oligonucleotide primers corresponded to positions within the *fcrA76* gene (7): nucleotides 224–249 and nucleotides 1182–1160. For the SF-3 gene derived from strain D691 (type 11), the forward primer corresponded to a conserved region located within class II SF-2 and SF-3 gene leader peptides (5′-ACAGGTACAGCATCCGTAGCAGTCGCT-3′) (3, 5, 9), and the reverse primer corresponded to a (Glu-Gln)-rich domain (5′-GTTCTTGATAACGTTTTTTCTACTTCTCG-3′) (3). For the SF-3 gene derived from strain 29487 (type 33), the forward primer corresponded to sequences conserved within an upstream noncoding region (5′-CTGGCCTTTACTCCTTTTGATTAACC-3′) (3), and the reverse primer corresponded to the SF-3-specific region (5′-TTGAGCAGCTCTACC-3′) (1). The chromosomal positions of the priming sites used for generating the *emm* gene fragments were confirmed by PCR-based chromosomal mapping (1).

DNA Sequence Analysis. The percent sequence divergence (as reported in text) is measured based on nucleotide sequence alignments that exhibit no gaps; the nucleotide positions reported are based on position 1 being located at the 5′ end of the sequence encoding the predicted mature form of the gene product, as established based on alignment of putative signal peptidase cleavage sites (4, 7). A phylogenetic tree was constructed based on nucleotide sequence alignments between the 5′ regions of 12 SF-4 *emm* genes, using the program PILEUP in the GCG package with a gap penalty of 3.0 and a gap-length penalty of 0.1. Distances were calculated using the Kimura two-parameter method (17) and the relationships were estimated by the neighbor-joining method in the program NJBOOT2 (18).

Southern Hybridization. Southern hybridization was performed as described (1, 3) at 65°C. Double-stranded DNA probes were generated by PCR amplification using the primer pairs and chromosomal DNA templates listed in Table 1; oligonucleotides were synthesized by the Keck Biotechnology Resource Laboratory at Yale University and The Rockefeller University Sequencing Facility.

RESULTS

Structure of the 5′ Region of SF-4 *emm* Genes. SF-4 genes occupy the first position within the *emm* gene cluster and are found in all pattern 4 (class I) and pattern 5 (class II) isolates (Fig. 1) (1). SF-4 genes whose complete sequences are known include *fcrA76* and *mrp4* (derived from type 76 and 4 isolates, respectively), which encode IgG- and fibrinogen-binding proteins (4, 7). Because *fcrA76* and *mrp4* differ markedly in their 5′ region nucleotide sequence (and in their predicted amino acid sequence), it has been suggested that these genes encode proteins that contribute to the type specificity of group A streptococci. In an effort to better understand the *emm* gene products of a strain under intensive study in our laboratory (strain T2/MR, type 2, class II), its SF-4 gene was partially sequenced and found to be only 1% divergent from *fcrA76* in the first 231 nucleotides encoding the NH_2 terminus of the mature FcRA76 protein (data not shown). By comparison with the nucleotide sequence reported for *fcrA76* (7), only one of the two nucleotide substitutions is nonsynonomous (encoding amino acid residue 45). This finding sug-

Table 1. Oligonucleotides used as primers for the generation of DNA probes

Probe	Nucleotide sequence of primer	Nucleotide position
SF4^{a}	5′-GAGACCGTAGGTCGCTTTAGTGATG-3′	1–25 (F)
	5′-TTCTCAATAGTGTGCGTAAGAGCTT-3′	234–209 (R)
SF4^{b}	5′-GAGAGTCGTCGTTATCAGGCACCT-3′	1–24 (F)
	5′-TGTGACATGTGGTTAATAGTGTCAC-3′	131–109 (R)
SF4^{c}	5′-GACTTAAGTACTCAGGAACATCCTAGAG-3′	1–28 (F)
	5′-TTACGATAAGAGCCTGCCAAAGCAGC-3′	123–106 (R)
SF3^{a}	5′-GATGAAGCTAAAATGGAAGTA-3′	1–24 (F)
	5′-TAGTGCATCAATGCCATCCTGTAAT-3′	213–188 (R)
SF3^{b}	5′-GAAGGGGTAAAAGCGACTACGAACTTGCCA-3′	1–30 (F)
	5′-CTGAGTTTCAATTACATCATGAAAGTTAAG-3′	200–171 (R)
SF3^{c}	5′-GATGATGCTACTACGCAGGGGACT-3′	1–24 (F)
	5′-TAATTGATCTACTTTATCAAGTAATTC-3′	114–88 (R)
SF3^{d}	5′-GAAGAAGCAAAAAGAACAGCACCATAT-3′	1–27 (F)
	5′-GTAGTGACGTTCTACATCTTCTGA-3′	150–127 (R)

The nucleotide positions indicated are based on position 1 being located at the 5′ end of the sequence encoding the predicted mature form of the gene product. Template DNA was derived from strains of the following serotypes: type 2 (SF4^{a}, SF3^{a}), type 4 (SF4^{b}, SF3^{b}), type 49 (SF4^{c}), type 11 (SF3^{c}), type 33 (SF3^{d}). F, forward; R, reverse.

gested that SF-4 genes do not contribute to type specificity and prompted our further study of the 5′ sequence of nine additional SF-4 genes, derived from both class I and II group A streptococcal isolates.

All 12 SF-4 genes exhibited a highly conserved leader-encoding region, displaying <8% nucleotide sequence divergence frǫm *fcrA76* (for the 28 amino acids immediately preceding the predicted signal peptidase cleavage site; data not shown). Based on the strong homology in the leader region, the first nucleotide encoding the mature SF-4 gene product could be predicted.

The extent of 5′ region nucleotide sequence divergence, beginning at the nucleotide encoding the predicted mature form of each protein, was determined for the 12 SF-4 *emm* genes under study. The SF-4 genes from strains of serotypes 2, 15, and 65 exhibit <3% sequence divergence from *fcrA76* in the first 231 nucleotides corresponding to the mature FcRA76 protein. In contrast, the SF-4 gene from a serotype 4 strain (*mrp4*) displays high levels of sequence divergence from the *fcrA76*-like genes in the 5′ region (4, 7). However, the SF-4 genes from strains of serotypes 22, 33, and 52 exhibit only 3% nucleotide sequence divergence from *mrp4* in the first 225 nucleotides of the mature Mrp4 protein. The type 25 strain SF-4 gene is <2% divergent from the 5′ end of *mrp4* beginning 34 nucleotides beyond its predicted signal peptidase cleavage site and extending to base pair 144; however, the 5′ end of the gene encoding the first 11 amino acid residues of the mature protein exhibits >65% nucleotide sequence divergence from *mrp4*; the data suggest that the type 25 strain SF-4 gene is highly homologous to *mrp4*, except at its extreme 5′ end. A third major form of the SF-4 gene 5′ domain is represented by a strain of serotype 49. The type 49 strain SF-4 gene exhibits ≈10% nucleotide sequence divergence from SF-4 genes of serotypes 11 and 13 for the first 63 bp of the 5′ end; however, the level of sequence divergence increases significantly beyond nucleotide 63 (>50% divergence for nucleotides 64–147). The SF-4 genes derived from type 11 and 13 strains are closely related to one another, displaying >95% nucleotide sequence identity throughout their 5′ 147 bp, suggesting that the type 11 and 13 strain genes represent a divergent form of the type 49 strain SF-4 gene. Our nucleotide sequence for the SF-4 gene derived from a type 49 strain is largely in agreement with the recently reported sequence for *fcrA49* (19). A phylogenetic tree was constructed based on nucleotide sequence alignments between the 5′ ends of the SF-4 *emm* genes (Fig. 2). The phylogenetic tree is consistent with the percent nucleotide sequence divergence calculations presented above.

The 5′ end nucleotide sequences of the SF-4 genes described above were used as a basis for designing oligonucleotide primers (listed in Table 1), which in turn were used for PCR amplification of chromosomal DNA in order to generate DNA probes for Southern hybridization studies. Three gene-specific DNA probes corresponding to the 5′ end encoding mature SF-4 *emm* gene products of serotype 2 (SF4^a), 4 (SF4^b), and 49 (SF4^c) strains, respectively, were tested for hybridization to chromosomal DNA derived from strains of 28 different serotypes, under highly stringent conditions (Fig. 3). The three SF-4-specific gene probes are represented within each of the three major branches of the phylogenetic tree shown in Fig. 2. The SF-4 gene probe derived from type 2 organisms (SF4^a) hybridized to DNA from four isolates tested (types 2, 15, 65, and 76) (Fig. 3*A*). The SF-4 gene probe derived from type 49 streptococci (SF4^c) hybridized to DNA from three isolates tested (types 49, 11, and 13) (Fig. 3*C*). The SF-4 gene probe derived from type 4 organisms (SF4^b) hybridized to DNA from 6 of the 13 pattern 5 isolates tested (types 4, 22, 25, 46, 58, and 60) and to all 7 of the pattern 4 isolates tested (types 30, 32, 33, 36, 52, 53, and 56) (Fig. 3*B*). All 20 isolates containing SF-4 genes within their chromosomes hybridized with one of the three 5′ end probes (SF4^a, SF4^b, SF4^c), representing three different forms of the SF-4 *emm* allele. The data strongly suggest that only a limited number of distinct domains comprise the 5′ ends of SF-4

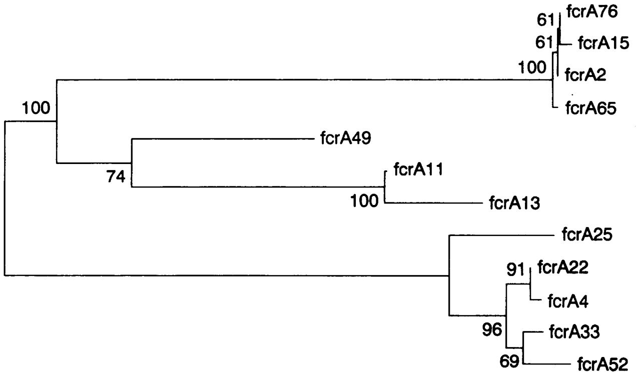

FIG. 2. Phylogenetic tree indicating genetic relationships that define three SF-4 *emm* gene (or *fcrA* or *mrp*) alleles. The 355 nucleotides used for alignment correspond to the region encoding the 28 amino acid residues preceding the predicted signal peptidase cleavage site and extend through the region encoding the NH$_2$-terminal 90 amino acid residues of the mature protein. The numbers that distinguish each SF-4 *emm* (or *fcrA*) gene refer to the M serotype of the strain from which it was derived (1); note that "*fcrA4*" has been designated "*mrp4*" (4). The numbers within the tree indicate the percentage of time each branch was joined together when 1000 bootstrap replications were performed.

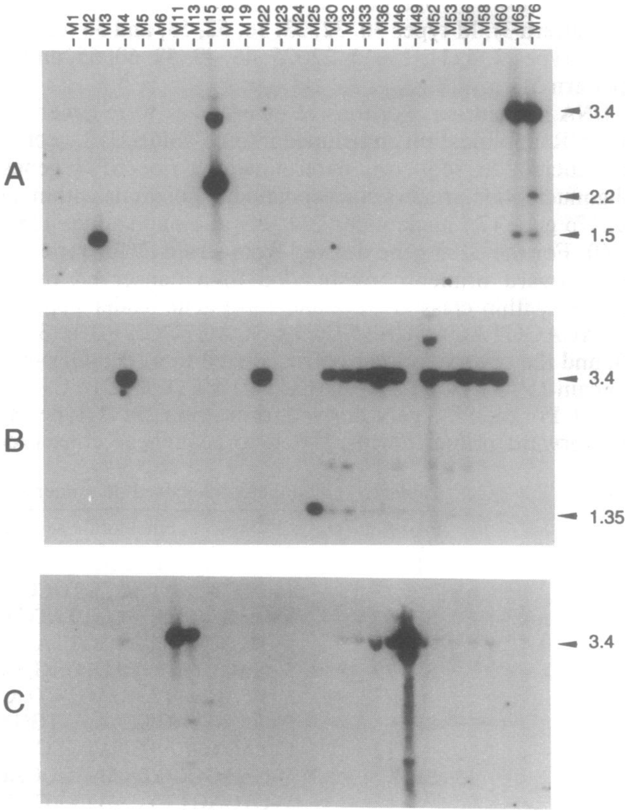

FIG. 3. Southern hybridization with DNA probes corresponding to the 5′ ends of SF-4 genes. Chromosomal DNA digested with a combination of *Pst* I, *Pvu* II, and *Cla* I was tested for hybridization with probes SF4^a (*A*), SF4^b (*B*), and SF4^c (*C*). The weak hybridization signals observed in the lanes containing types M36 and M46 DNA (*C*) are due to spillover during gel loading of type M49 DNA, which exhibits a very strong signal in this overexposed autoradiograph. Molecular size markers are in kb.

genes, and therefore it is unlikely that SF-4 genes are major contributors to type specificity.

Structure of the 5′ Region of SF-3 *emm* Genes. SF-3 genes occupy the last position within the *emm* gene cluster and are found in all pattern 3 and 4 (class I) and pattern 5 (class II) isolates (Fig. 1) (1). SF-3 genes whose complete sequences are known include *ennX*, *emm*L2.2, and *enn4*, derived from serotypes 49, 2, and 4 strains, respectively. *ennX* and *enn4* are nearly identical over their entire sequence (<4% divergent for the 262 nucleotides encoding the NH_2 terminus of the predicted, mature forms) and neither is known to be expressed (5, 20). *emm*L2.2 encodes an IgA-binding protein (3) and is ≈80% divergent from both *ennX* and *enn4* over its 5′ end 240 nucleotides (data not shown). The 5′ ends of the SF-3 genes are nonhomologous to the 5′ ends of the SF-4 genes. Gene-specific DNA probes (SF3^a and SF3^b) corresponding to the 5′ ends of the types 2- and 4-derived SF-3 genes, respectively, were tested for Southern hybridization to group A streptococcal isolates representing 28 different serotypes (Fig. 4). The gene-specific SF-3 DNA probe (SF3^a) derived from the type 2 strain hybridized to DNA from four isolates tested (types 2, 22, 25, and 65) (Fig. 4*A*). The SF-3 gene probe derived from type 4 organisms (SF3^b) hybridized to DNA from seven isolates tested (types 4, 15, 22, 46, 49, 58, and 76) (Fig. 4*B*).

There was a lack of concordance between hybridization with the 5′ end SF-3 and SF-4 gene probes derived from type 2 organisms (probes SF3^a and SF4^a) and with the 5′ end SF-3 and SF-4 gene probes derived from type 4 organisms (probes SF3^b and SF4^b) (summarized in Fig. 5). For example, the type 25 strain hybridized with the SF3^a and SF4^b probes, whereas the type 15 and 76 strains hybridized with the SF3^b and SF4^a probes. Since two forms of the SF-3 gene 5′ domain (SF3^a, SF3^b) and two forms of the SF-4 gene 5′ domain (SF4^a, SF4^b) are found to exist in all four possible combinations (SF3^a-SF4^a, SF3^{a}SF4^b, SF3^{b}SF4^a, SF3^{b}SF4^b), at some point in time a recombinational event involving DNA from a second streptococcal strain must have occurred.

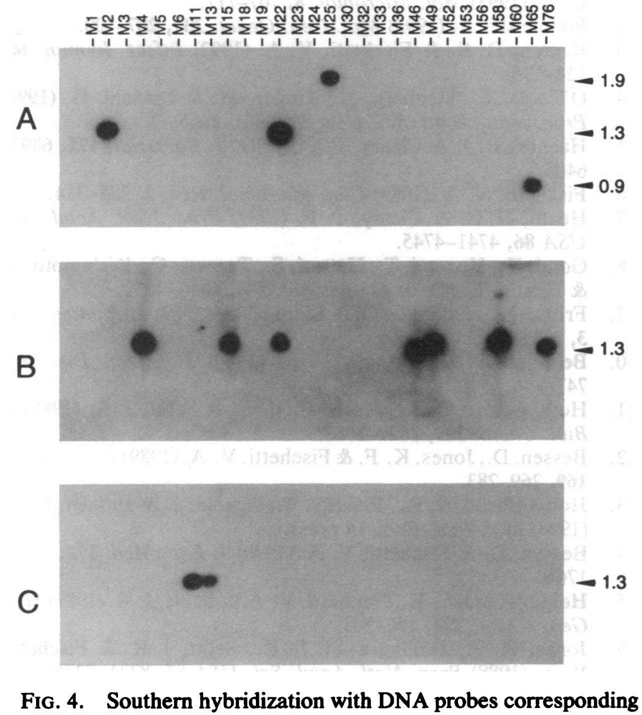

FIG. 4. Southern hybridization with DNA probes corresponding to the 5′ ends of SF-3 genes. Chromosomal DNA digested with a combination of *Pst* I, *Pvu* II, and *Cla* I was tested for hybridization with probes SF3^a (*A*), SF3^b (*B*), and SF3^c (*C*). Note that a single fragment of type 22 DNA hybridizes strongly with both the SF3^a and SF3^b probes; that this SF-3 gene may be a hybrid is supported by PCR-based mapping studies wherein priming is achieved with the SF3^a forward and SF3^b reverse primers but not with the SF3^b forward or SF3^a reverse primers (data not shown). Molecular size markers are in kb.

	SF4^a	SF4^b	SF4^c
SF3^a	M2, M65	M22, M25	
SF3^b	M15, M76	M4, M22 M46, M58	M49
SF3^c			M11, M13
SF3^d		M33	
ND		M30,M32,M36 M52,M53,M56 M60	

FIG. 5. Summary of combinations of 5′ end forms of SF-3 and SF-4 genes among 20 strains of group A streptococci. Hybridization with 5′ region SF-4 and SF-3 DNA probes is indicated for strains of each serotype examined. ND, not determined.

The SF3^a and SF3^b probes failed to hybridize with three class II strains (pattern 5) and with all nine class I strains having SF-3 *emm* genes (patterns 3 and 4). The 5′ region of the SF3 gene derived from the type 11 strain was partially sequenced and a gene-specific probe (SF3^c) was prepared based on its mature 5′ end sequence (Table 1). Probe SF3^c hybridized to two of the remaining three SF-3 genes derived from class II isolates under study (types 11 and 13) and failed to hybridize with class I organisms (Fig. 4*C*). Collectively, probes SF3^a, SF3^b, and SF3^c hybridized to 12 of the 13 pattern 5 (class II) isolates tested but to none of the pattern 3 or 4 (class I) isolates tested. To better define the 5′ ends of class I-derived SF-3 genes, the 5′ nucleotide sequence was determined for an SF-3 gene obtained from a pattern 4 isolate (type 33, strain 29487), and a gene-specific DNA probe (SF3^d) was constructed (Table 1) and tested for hybridization. Of the 22 isolates bearing SF-3 genes, probe SF3^d hybridized strongly only to the strain from which it was derived (data not shown); this finding was confirmed by PCR amplification of chromosomal DNA originating from the other class I isolates, using the 5′ region SF3^d forward primer (Table 1) paired with a reverse primer corresponding in sequence to the SF3-specific site that encodes the peptidoglycan-spanning domain (1). Thus, while one cannot rule out that some SF-3 genes may contain type-specific determinants encoded by their 5′ ends, it appears that many SF-3 genes are not contributors to type specificity.

PCR-Based Chromosomal Mapping. To confirm that the DNA probes corresponding to the 5′ ends of SF-4 and SF-3 genes do indeed hybridize to these genes and not to undefined fragments, a PCR-based chromosomal mapping technique was employed (1). The oligonucleotide forward primers used to generate the DNA probes for Southern hybridization (Table 1) were tested for PCR amplification of chromosomal DNA in combination with primers corresponding to the SF-specific regions of the *emm* genes, which encode the peptidoglycan-spanning region (1) (data not shown). For all strains, the PCR-based chromosomal mapping results demonstrated that the 5′ regions of SF-3 genes lie ≈800–900 bp upstream from the SF-3-specific site. Similarly, the 5′ regions of SF-4 genes lie ≈2.4 kb upstream from the SF-1-specific site (class I) or SF-2-specific site (class II).

With only two exceptions (types 25 and 58 strains), PCR amplification using the 5′ region SF-4 forward primers corroborated the Southern hybridization findings. The failure to amplify the type 25-derived SF-4 gene using the SF4^b forward primer is consistent with its lack of nucleotide sequence homology at the extreme 5′ end (Fig. 2). Likewise, all 5′ region SF-3 forward primers (with the single exception of the type 65 strain) amplified DNA from those strains exhibiting positive signals by Southern hybridization with the corresponding DNA probes. For all strains, PCR amplification was achieved with the 5′ region SF-3 and SF-4 reverse primers (Table 1), which match the DNA probes exhibiting positive Southern hybridization signals. Thus, the findings from PCR-based chromosomal mapping studies confirm the chromosomal positions of the 5′ regions of SF-3 and SF-4 genes and, furthermore, provide additional evidence in support of limited polymorphism at the extreme 5′ ends of SF-3 and SF-4 genes.

DISCUSSION

In this report, we provide evidence that strongly suggests that horizontal gene transfer is a mechanism that contributes to the generation of *emm* gene diversity. Our conclusions are based on the finding that all four possible combinations of two alleles (SF-3 and SF-4) located at different chromosomal positions, wherein each allele has two distinct forms, are found to exist among naturally occurring isolates of group A streptococci. An alternative hypothesis to explain these findings invokes the duplication of both SF-3 and SF-4 *emm* genes, followed by 5′ sequence divergence and, eventually, deletion of one SF-3 and one SF-4 *emm* gene. However, we believe that this latter scenario is unlikely because none of the 44 phenotypically diverse isolates examined to date display multiple SF-3 or SF-4 *emm* genes (1, 13), a chromosomal pattern characteristic of the hypothetical intermediate state. The vehicle by which horizontal spread occurs cannot be derived from the data; however, plasmids, bacteriophage, and conjugative transposons are all plausible candidates, whereas uptake of naked DNA is a less likely mechanism because of the poor natural transformability of this species.

One of the *emm* domains that is implicated in recombinational exchange following horizontal gene transfer is represented by the DNA probe SF3^a. Although expression vectors containing a cloned insert corresponding to the SF3^a probe yield IgA-binding activity (21), there is no strict correlation between hybridization with the SF3^a probe and IgA binding by whole streptococci (data not shown). IgA-binding activity depends on expression of the IgA-binding gene product, and there is evidence that not all SF-3 genes are expressed (5, 20). Second, IgA-binding activity can be associated with genes other than SF-3 genes, such as the SF-2 gene, *arp4* (9). The role of IgA binding in group A streptococcal pathogenesis is unclear; however, IgA-binding activity inversely correlates with isolates derived from nasopharyngeal sites and positively correlates with a disproportionately high percentage of isolates derived from deep tissue (10).

A second important finding of this study is the relative lack of 5′ sequence diversity among SF-3 and SF-4 *emm* genes. Collectively, three 5′ region gene probes (SF4^a, SF4^b, SF4^c) hybridized to all 20 of the SF-4 genes under study. The 5′ ends of SF-3 genes display greater polymorphism than the corresponding region of SF-4 genes, since three SF-3 gene probes (SF3^a, SF3^b, SF3^c) hybridized to only 12 of the 22 SF-3 genes examined. However, the extent of 5′ region polymorphism in SF-4 and SF-3 genes is in striking contrast to that observed for SF-1 and SF-2 genes. Of the >10 SF-1 and SF-2 genes whose complete or partial nucleotide sequences have been reported, all lack substantial homology to one another at their 5′ ends. Moreover, hybridization with an oligonucleotide probe corresponding to amino acid residues 30–38 of the mature, type 2-derived SF-2 gene product reveals positive signals with only type 2 isolates and no signal with any of the other 23 serotypes tested (3). Thus, our findings strongly suggest but do not directly prove that some, if not most, SF-3 and SF-4 genes are *not* major contributors to type specificity of group A streptococci. This is important because only type-specific serum IgG overcomes the antiphagocytic effect of M protein, resulting in opsonophagocytosis and protective immunity (6). It is of interest that all group A streptococci examined have either an SF-1 or SF-2 gene (1) and that for one pattern 5 isolate studied, RNA transcript levels of the SF-2 gene exceed that of the SF-3 gene by 30-fold (3). The data accumulated to date point to type-specific determinants lying within SF-1 and SF-2 gene products, whose levels of expression exceed that of other *emm* family genes. This hypothesis remains to be tested using a serological approach.

Some of the *emm* family gene products are presently recognized as important virulence determinants by virtue of their antiphagocytic properties (6), and those with immunoglobulin-binding activities are believed to influence pathogenicity since their expression correlates with disease (10). Understanding the mechanisms underlying the evolution of the structurally and functionally diverse family of *emm* genes and gene products should ultimately lead to insight on the epidemiology of group A streptococcal diseases.

We are grateful for the technical expertise of George Brough and also thank Drs. Barry Gray and Rebeca Rico-Hesse for critical review of the manuscript. This work was supported by U.S. Public Health Service Grant AI-28944 and an American Heart Association Grant-in-Aid (to D.E.B.) and by National Science Foundation Grant DMB 8903643 and an American Heart Association Grant-in-Aid (to S.K.H.).

1. Hollingshead, S. K., Readdy, T. L., Yung, D. L. & Bessen, D. E. (1993) *Mol. Microbiol.* **8,** 707–717.
2. Podbielski, A. (1993) *Mol. Gen. Genet.* **237,** 287–300.
3. Bessen, D. E. & Fischetti, V. A. (1992) *Infect. Immun.* **60,** 124–135.
4. O'Toole, P., Stenberg, L., Rissler, M. & Lindahl, G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 8661–8665.
5. Haanes, E. J. & Cleary, P. P. (1989) *J. Bacteriol.* **171,** 6397–6408.
6. Fischetti, V. A. (1989) *Clin. Microbiol. Rev.* **2,** 285–314.
7. Heath, D. G. & Cleary, P. P. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 4741–4745.
8. Gomi, H., Hozumi, T., Hattori, S., Tagawa, C., Kishimoto, F. & Bjorck, L. (1990) *J. Immunol.* **144,** 4046–4052.
9. Frithz, E., Heden, L.-O. & Lindahl, G. (1989) *Mol. Microbiol.* **3,** 1111–1119.
10. Bessen, D. & Fischetti, V. A. (1990) *J. Infect. Dis.* **161,** 747–754.
11. Hollingshead, S. K., Fischetti, V. A. & Scott, J. R. (1986) *J. Biol. Chem.* **261,** 1677–1686.
12. Bessen, D., Jones, K. F. & Fischetti, V. A. (1989) *J. Exp. Med.* **169,** 269–283.
13. Hollingshead, S. K., Readdy, T., Arnold, J. & Bessen, D. E. (1994) *Mol. Biol. Evol.* in press.
14. Bessen, D. & Fischetti, V. A. (1990) *J. Exp. Med.* **172,** 1757–1764.
15. Hollingshead, S. K., Fischetti, V. A. & Scott, J. R. (1987) *Mol. Gen. Genet.* **207,** 196–203.
16. Jones, K. F., Hollingshead, S. K., Scott, J. R. & Fischetti, V. A. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 8271–8275.
17. Kimura, M. (1980) *J. Mol. Evol.* **16,** 111–120.
18. Tamura, K. (1993) NJBOOT2: *Neighbor-Joining Analysis with Bootstrap Version 2.0* (Inst. Mol. Evol. Genet., Pennsylvania State Univ., College Park).
19. Podbielski, A., Kaufhold, A. & Cleary, P. P. (1993) *Immunomethods* **2,** 55–64.
20. Jeppson, H., Frithz, E. & Heden, L. O. (1992) *FEMS Microbiol. Lett.* **92,** 139–146.
21. Bessen, D. E. (1994) *Infec. Immun.*, in press.