JOURNAL OF BACTERIOLOGY, Aug. 1992, p. 4967–4976
0021-9193/92/154967-10$02.00/0

Vol. 174, No. 15

Architecture of the *vir* Regulons of Group A Streptococci Parallels Opacity Factor Phenotype and M Protein Class

ELIZABETH J. HAANES,† DAVID G. HEATH,‡ AND P. PATRICK CLEARY*

Department of Microbiology, University of Minnesota, Minneapolis, Minnesota 55455-0312

Received 17 January 1992/Accepted 22 May 1992

Group A streptococci have traditionally been categorized into two broad groups based on the presence or absence of serum opacity factor (OF). Recent studies show that these two groups vary in a number of properties in addition to the OF phenotype, including sequence variations in the constant region of the antiphagocytic M protein genes, the presence or absence of immunoglobulin G Fc receptor proteins, and the presence or absence of multiple M protein-like genes situated in a tandem array. The M protein genes (*emm*) in OF$^-$ streptococcal strains are known to be part of a regulon of virulence-related genes controlled by the *trans*-acting positive regulatory gene, *virR*, situated just upstream of *emm*. In OF$^+$ strains, however, the region adjacent to *virR* is occupied by an M protein-related, type IIa immunoglobulin G Fc receptor gene (*fcrA*), and the relative position of *emm* has not been determined. To further define the *vir* regulon in OF$^+$ streptococci, we used the polymerase chain reaction to show that *fcrA49* is situated immediately upstream of *emm49* in the OF$^+$ type 49 strain CS101. This result shows for the first time the separate identity and genetic linkage of these two genes in the *vir* regulon of an OF$^+$ group A streptococcal strain and confirms our previous hypothesis that *emm49* exists as the central gene in a trio of *emm*-like genes. Additionally, using DNA hybridizations, we found considerable sequence divergence between OF$^-$ and OF$^+$ group A streptococci in *virR* and in the noncoding sequences between *virR* and the *emm* or *fcrA* expression site. We found, however, a high degree of sequence conservation in this region within each of the two groups of strains.

Streptococcus pyogenes, or group A streptococcus (GAS), is a highly successful human pathogen. This gram-positive organism is responsible for a variety of diseases, and recently, highly invasive isolates are resulting in alarming increases in morbidity and mortality (35). Although the pathogenesis of GAS diseases is very complex, structures on the streptococcal surface clearly allow for establishment, survival, and proliferation of the organisms in the human host. Characterized surface structures include the antigenically variant M protein (24), the streptococcal C5a peptidase (SCP) (46), and various immunoglobulin receptors (6, 19, 28). Of these, the only structure presently proven to enhance streptococcal virulence is the M protein, which blocks phagocytosis via the alternative complement pathway. M protein and SCP are, nonetheless, coregulated by the *virR* (or *mry*) gene, situated just upstream of the M protein gene (7, 36, 40). The genes for M protein (*emm*) and the SCP (*scpA*), in addition to being coregulated, are genetically linked (9), thus constituting a *vir* regulon. Preliminary evidence also points to coregulation or genetic linkage of M protein genes and immunoglobulin G (IgG) or IgA Fc receptor genes in some streptococcal strains (17, 26, 28).

GAS have long been loosely categorized into two broad groups based on the presence or absence of the serum opacity factor (OF) (16, 21, 44). OF$^-$ and OF$^+$ streptococci also differ in a number of other properties, as reviewed by Maxted (29). For instance, OF$^+$ strains tend to be isolated from impetiginous skin lesions, and OF$^-$ strains are isolated from pharyngitis infections. Additionally, the M proteins from OF$^+$ strains appear to be less immunogenic than their OF$^-$ counterparts (45).

Recent studies have sparked new interest into the differences between OF$^-$ and OF$^+$ GAS. We recently showed that the nucleotide sequence of the gene encoding the type 49 M protein (*emm49*), the first cloned from an OF$^+$ serotype, differs considerably from sequences of the previously sequenced M protein genes of OF$^-$ serotypes not only in the region encoding the antigenically variant portion of the molecule but also in previously defined constant regions (15). We also showed that *emm49* was the upstream part of a tandem array of M protein-related genes, in contrast to the single-copy genes found in other serotypes. We designated this downstream gene *ennX*. Recently, the homolog of *ennX* in an M2, OF$^+$ GAS was shown to be an IgA Fc receptor protein (2).

In agreement with our findings of distinctions between OF$^+$ and OF$^-$ GAS at the DNA level, the M proteins (or M-like proteins) of OF$^-$ and OF$^+$ strains appear to have antigenically dissimilar constant regions, as determined by the binding of two monoclonal antibodies directed against the C repeat region of an M6, OF$^-$ GAS strain (3, 4). These investigators designated M proteins as being either class I (generally OF$^-$) or class II (generally OF$^+$), depending on whether or not the strain's M protein binds these antibodies.

OF$^+$ GAS also appear to differ from OF$^-$ GAS in the expression of another surface protein. *fcrA76*, a type IIa IgG Fc receptor gene, was cloned from the M76, OF$^+$ strain CS110 (19). This gene is situated in the same expression site as are the *emm* genes in OF$^-$ strains according to upstream flanking DNA similarities (including the *virR* region). Furthermore, the *fcrA76* sequence contains high DNA similarity to *emm* genes both in the leader sequence and in the sequences encoding the cell wall-spanning and membrane anchor regions (20). The FcrA protein is structurally similar to M proteins and is expressed on the surface of at least two OF$^+$ GAS, CS110 (M76) and CS101 (M49), according to specific antibody and nonspecific immunoglobulin (IgG1,

* Corresponding author.
† Present address: The Upjohn Company, Kalamazoo, MI 49001.
‡ Present address: Department of Biologic and Materials Sciences, University of Michigan, Ann Arbor, MI 48109.

IgG2, and IgG4) binding studies (6, 19, 25). In addition, Bessen and Fischetti (1) found high levels of IgG Fc binding in all OF^+ GAS and in some OF^- GAS isolated from skin infections, while they found much less IgG binding in OF^- GAS isolated from pharyngitis. The Fc receptor described by these authors, however, has not specifically been shown to be the FcrA protein.

To define more precisely the *vir* regulon in OF^+ GAS, we determined the location of the *emm* genes relative to the previously defined location of *fcrA*. Using the polymerase chain reaction (PCR) and nucleotide sequencing, we showed that *fcrA49* is situated immediately upstream of *emm49* in the OF^+ strain CS101, proving for the first time the separate identity and genetic linkage of these two genes in an OF^+ GAS. This finding confirmed our previous hypothesis that *emm49* existed as the central gene in a trio of *emm*-like genes.

Given the observed differences between OF^+ and OF^- GAS, and the multiple genes that we defined in the OF^+ M49 strain, we hypothesized that the two groups might have diverged considerably in noncoding regions within the *vir* regulon. Therefore, we used DNA fragments flanking the M protein gene *emm1* (14) and the Fc receptor gene *fcrA76* (19, 20) as hybridization probes to compare the *virR* region and the adjacent M protein (or Fc receptor) expression sites of a variety of M serotypes. In nearly all cases, we were able to differentiate highly conserved OF^+ and OF^- patterns on the basis of differential hybridization to the various probes.

MATERIALS AND METHODS

Preparation of DNA. GAS were protoplasted with mutanolysin, and DNA was purified from the lysed protoplasts as previously described (14, 15).

PCR. CS101 streptococcal DNA was subjected to PCR (34) using synthetic oligonucleotides fc (CCAAA ACGCT AGTCG TGAAA TCGC), the forward primer, and m49′ (GTCGC TCTTC CAGTT CTCCT ATCC), the reverse primer. The fc primer annealed to a point 22 bp upstream of a *Pvu*II site in *fcrA76* (20), and the m49′ primer annealed beginning 3 bp downstream of the first *Xba*I site in *emm49* (15). The reaction was performed in a buffer consisting of KCl (50 mM), Tris-HCl (pH 8.3) (10 mM), $MgCl_2$ (2.5 mM), and gelatin (0.01% [wt/vol]), to which the four deoxynucleoside triphosphates (25 μM each), the two primers (44 pM each), and 2.5 U of *Taq* polymerase (Perkin Elmer Cetus, Norwalk, Conn.) were added. After topping with light mineral oil, target DNAs were added to the tubes to a final volume of 100 μl. The PCR consisted of an initial 4-min denaturation at 94°C followed by 39 cycles of denaturation (94°C, 1 min), annealing (50°C, 2 min), and polymerization (72°C, 1.5 min). The 40th cycle concluded with a 6-min polymerization step followed by incubation at 4°C. PCR products were gel purified by using NA45 membranes (Schleicher & Schuell, Keene, N.H.).

PCR cloning. For hybridization analyses, the PCR product was digested with *Xba*I and *Pvu*II and cloned into pUC18 (47) digested with *Xba*I and *Sma*I. To detect expression, the product was cloned into the expression vector pKK233-2 exactly as described by Heath et al. (18). Briefly, the product was digested with *Pvu*II, and a 12-mer *Nco*I linker was ligated to the blunt end. The construct was then digested with *Nco*I and *Hin*dIII, the 487-bp fragment was gel purified and ligated into the *Nco*I- and *Hin*dIII-digested vector, and the construct was transformed into JM105. Three clones containing inserts, pK49-3, pK49-5, and pK49-9, were used in the expression studies. Basic enzymatic, cloning, and transformation operations were done by previously described methods (14, 15).

Detection of IgG1 binding. Subcultures of the three pK49 clones were grown to early log phase in M9 medium (31) with ampicillin (50 μg/ml), at which time they were induced with 0.1 M isopropyl-β-D-thiogalactopyranoside (IPTG) and grown for an additional 5 h. The cultures were sonicated as previously described (15) and analyzed in Ouchterlony double-diffusion analyses (19) against an IgG1 human myeloma protein (kindly provided by the World Health Organization).

DNA sequencing. The PCR product previously cloned into pUC18 was digested with *Hin*dIII and *Xba*I and cloned into the M13 vectors mp18 and mp19 (31, 47). Purification of single-stranded templates and DNA sequencing by the dideoxy-chain termination method (38) were performed exactly as previously described (15), using [α-^{35}S]dATP (NEN, Boston, Mass.) and Sequenase (U.S. Biochemical, Cleveland, Ohio).

DNA probes. Large DNA fragments (>499 bp) or whole plasmids were labelled with [α-^{32}P]dCTP or -dATP (NEN) by nick translation, using kit enzymes and reagents purchased from Bethesda Research Laboratories (Gaithersburg, Md.). Oligonucleotides were synthesized on a Biosearch 8600 DNA synthesizer and were purified by high-pressure liquid chromatography at the University of Minnesota Microchemical Facility. Oligonucleotides were labelled with [α-^{32}P]dATP, using terminal deoxynucleotidyltransferase (10).

DNA transfers and hybridizations. DNA restriction digests in agarose gels were transferred to nitrocellulose by vacuum transfer. Blots were hybridized by using heparin (Sigma Chemical Co., St. Louis, Mo.) as described by Singh and Jones (41). A modified hybridization solution containing 3× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate, pH 7.0), 0.2% sodium pyrophosphate, 0.2% sodium dodecyl sulfate, 100 mg of polyadenosine (Sigma) per ml, and 1 mg of heparin per ml was used in oligonucleotide hybridizations. Stringencies for both nick-translated and oligonucleotide probes were controlled by varying the temperature of hybridizations and washes according to standard formulas (30), assuming a G+C content for *S. pyogenes* DNA of 37% (27). Radiolabelled probes were stripped from blots by boiling in 0.1× SSC. Complete probe removal was then verified by extended autoradiography.

Nucleotide sequence accession number. The nucleotide sequence data presented in this paper were submitted to GenBank under accession number M86806.

RESULTS

Linkage of the *emm* and *fcrA* genes in GAS strain CS101. The type IIA IgG Fc receptor gene (*fcrA76*) is present in the OF^+ strain CS110 in the same chromosomal position as are the M protein genes in OF^- strains (19). M protein and FcrA protein, however, have not yet been established to be distinct proteins in OF^+ strains, and thus the location of the M protein gene relative to the Fc receptor gene is not known. Therefore, we were interested in mapping the position of the M protein gene in OF^+ strains relative to the position of *fcrA*. We previously established that the protein expressed from the cloned M49 gene (*emm49*) did not bind an IgG1 human myeloma protein, and preliminary hybridization studies using *fcrA76*- and *emm49*-specific probes against CS101 chromosomal DNA led us to suspect that an Fc receptor or *fcrA*-like gene was situated within 3 kb upstream of *emm49* (13). To verify this hypothesis, we proceeded to amplify a fragment of CS101 DNA containing portions of both the *emm49* gene and the putative *fcrA49* gene by PCR.

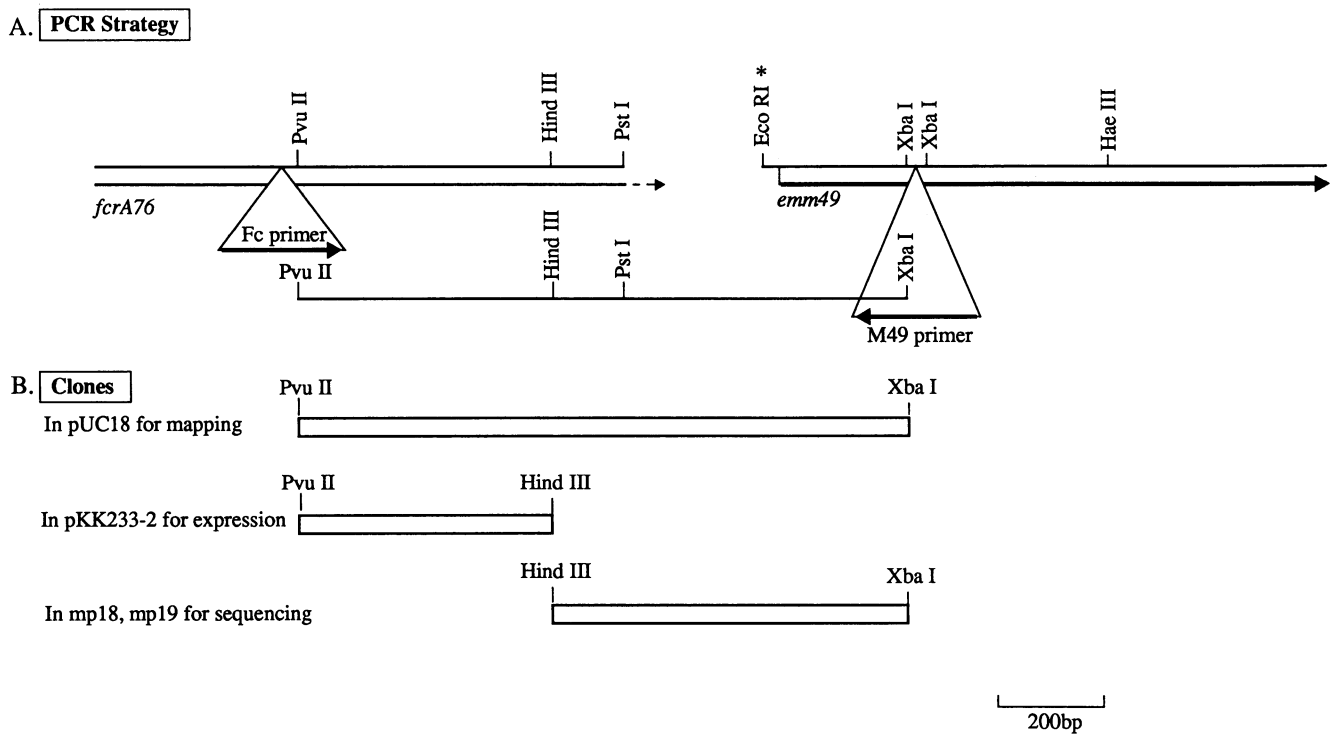

FIG. 1. (A) Strategy for amplification of the region spanning *fcrA49* and *emm49*. The top line shows maps of the regions of *fcrA76* and *emm49* used to devise the PCR; the bold arrows indicate locations of the forward and reverse primers. The line below the maps represents the 1.17-kb PCR product amplified from CS101 DNA. (B) Regions of the PCR product that were cloned into the indicated vectors.

Our strategy is outlined in Fig. 1A. PCR primers were synthesized according to the *emm49* (15) and *fcrA76* (20) sequences such that they occurred immediately flanking unique *Xba*I and *Pvu*II restriction sites, respectively. The PCR reaction amplified a 1.17-kb fragment from CS101 DNA, while reactions containing no target DNA or heterologous target DNA (herpes simplex virus type 1) produced no amplified product (data not shown).

We digested the PCR product with *Xba*I and *Pvu*II and cloned it into pUC18 for restriction and hybridization analyses (Fig. 1B). The PCR product contained a single *Hin*dIII site as predicted by the *emm49* and *fcrA76* restriction maps (15, 19). We verified the identity of the cloned PCR product by hybridization to both *fcrA76*- and *emm49*-specific probes (data not shown). Both of the probes hybridized strongly with the cloned PCR product, and they hybridized to separate fragments when the insert was digested with *Hin*dIII.

To verify that the PCR product amplified from CS101 indeed encoded a portion of a type II Fc receptor gene, we cloned the putative Fc-binding portion of the gene into the expression vector pKK233-2 by using the 12-mer *Nco*I linker exactly as described by Heath et al. (18) (Fig. 1B). We selected three clones, pKK49-3, pKK49-5, and pKK49-9, for these studies. A similar construct encoding the Fc-binding region of *fcrA76*, pKDGH-1, was used as a positive control. Sonicated extracts of IPTG-induced JM105 containing these four clones were examined in an Ouchterlony immunodiffusion against an IgG1 human myeloma protein. As shown in Fig. 2, the clones from the CS101 PCR product expressed a peptide that formed a line of identity with pKDGH-1. In addition, the expressed protein from pKK49-3 was found to bind to human IgG1, IgG2, and IgG4, but not IgG3, and bound specific antibody raised against *fcrA76* (6). We therefore concluded that this upstream gene indeed encoded the type IIa Fc receptor gene *fcrA49*, which was similar or identical to *fcrA76*.

Sequencing of the intervening region between *fcrA49* and *emm49*. According to our predicted arrangement for the layout of genes in the *vir* regulon of GAS strain CS101 (15), *emm49* maps as the central gene in trio of related genes, presumably the product of gene triplication sometime in the evolution of this strain. Using existing models for the generation of gene triplications (42), we predicted that the intervening sequence between *fcrA49* and *emm49* should be similar to the sequence between *emm49* and *ennX*. To

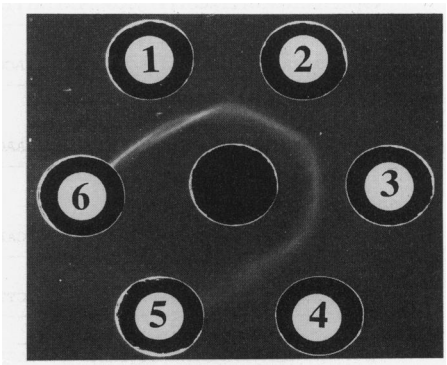

FIG. 2. Ouchterlony immunodiffusion analysis of sonicated *E. coli* carrying the Fc-binding region of *fcrA49* cloned in pKK233-2. Wells: 1, pKDGH-1 (cloned from *fcrA76*); 2, pK49-1; 3, pK49-5; 4, pK49-9; 5, pKK233-2; 6, phosphate-buffered saline. The center well contains a purified IgG1 myeloma protein.

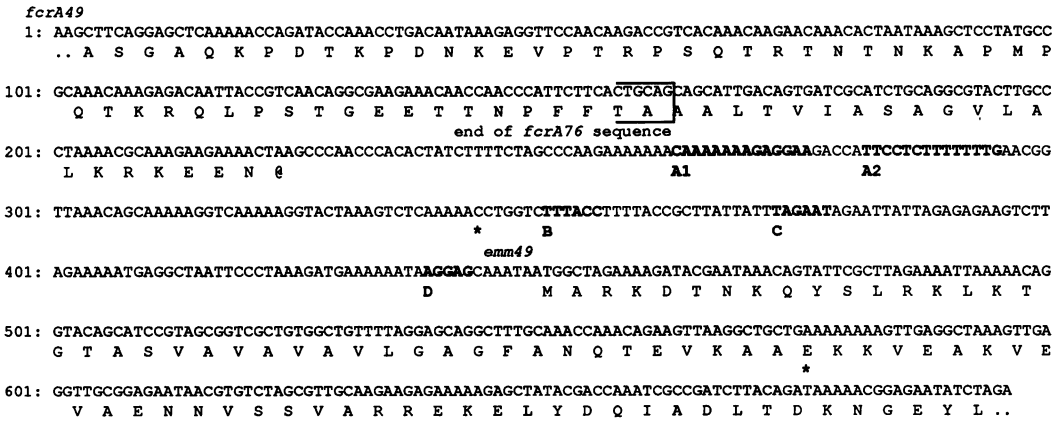

FIG. 3. Nucleotide sequence of the *Hin*dIII-*Xba*I fragment of the PCR product amplified from GAS strain CS101. The partial deduced amino acid sequences of the FcrA49 and M49 proteins are shown below the nucleotide sequences. The extent of the region that aligns with the published *fcrA76* sequence is indicated with an open box. The boundaries of the sequence that align with the published *arp4* sequence are marked with asterisks. Putative transcriptional and translational features of the sequence are shown in boldface. A1 and A2, inverted repeats representing a putative transcription termination signal for *fcrA49*; B and C, putative −35 and −10 promoter sequences for *emm49*; D, Shine-Dalgarno ribosome binding site for *emm49*.

examine this possibility, and to further verify that *fcrA49* is closely related to *fcrA76*, we determined the nucleotide sequence of the region between the *Hin*dIII site in *fcrA49* and the *Xba*I site in *emm49* (Fig. 1B). The entire sequence, shown in Fig. 3, was determined at least three times, and all of the sequence, except that previously sequenced in *emm49* (15), was determined from both strands.

The partial sequence of *fcrA49* from the *Hin*dIII site was nearly identical to the corresponding region of *fcrA76* (Fig. 4A) but included an additional 62 bp of the gene that was not sequenced in the *fcrA76* gene. This region also showed a high degree of similarity to the sequences encoding the cell wall-spanning and membrane anchor portions of M proteins, as expected from analyses of the various M protein gene sequences (comparison not shown). An inverted repeat with a free energy of 17.4 kcal (72.8 kJ)/mol, situated 46 bp downstream of the termination codon between nucleotides 263 and 295, may represent a transcription termination signal for the gene. Further downstream, between nucleotides 350 and 378, were putative −35 and −10 promoter sequences for *emm49*. The total length of the intervening region between *fcrA49* and *emm49* was 224 bp.

Although the size of this intervening region was similar to that between *emm49* and *ennX* (209 bp), the two intervening

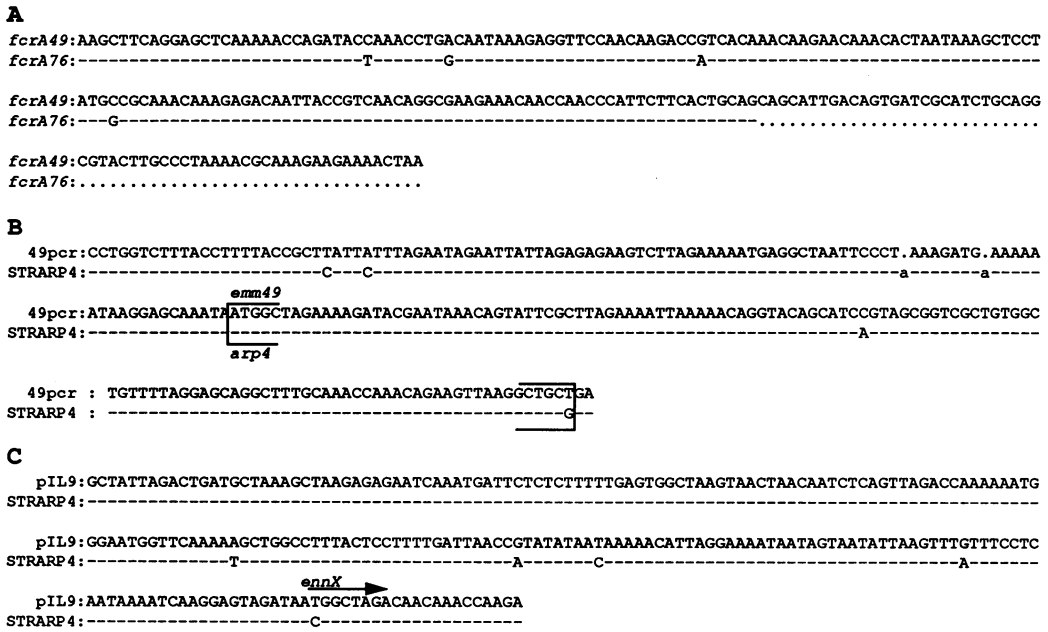

FIG. 4. Alignments of the CS101 sequences with other GAS sequences. (A) Alignment of *fcrA49* and *fcrA76* sequences from the beginning of the cloned PCR product to the termination codon of *fcrA49*. (B) Alignment of the upstream and 5′-terminal sequences of *emm49* and *arp4*. The regions encoding the signal peptides of *emm49* and *arp4* are surrounded by open boxes. (C) Alignment of sequences immediately downstream of *emm49* and *arp4*, beginning immediately after the respective termination codons. The translational start site of *ennX* is indicated.

regions shared no sequence similarity whatsoever, even when the stringency of the search was lowered to allow 40% base pair mismatch (data not shown). Given this puzzling result, we compared the 224-bp intervening sequence between *fcrA49* and *emm49* with all of the bacterial sequences in GenBank. The only similar sequence recovered was the 105-bp sequence upstream of the *arp4* gene, an IgA Fc receptor gene from an M4 GAS (10). This upstream sequence was >98% similar to the corresponding portion of our sequence, and the similarity extended into the subsequent leader sequences (Fig. 4B). Also, the regions downstream of *arp4* and *emm49* were >98% similar (Fig. 4C). This downstream similarity extended into the open reading frame that we previously designated *ennX*, but the sequence in the M4 strain containing *arp4* does not contain the ATG start codon for this open reading frame (2). In keeping with these upstream and downstream similarities, the constant regions of *emm49* and *arp4*, extending from the beginning of the C repeat regions to the end of the coding sequences, were virtually identical (data not shown).

Comparison of OF$^+$ and OF$^-$ GAS sequences in the noncoding region downstream of the *virR* gene. Recent findings show significant differences in OF$^+$ and OF$^-$ streptococcal strains based on variations in M protein C repeat regions, the presence or absence of a type IIa IgG Fc receptor, and other epidemiologic properties. We were interested in determining whether noncoding regions within the *vir* regulon also varied between the two groups. To accomplish this, we used DNA hybridization to evaluate sequence similarities upstream of the M protein and Fc receptor expression sites of a variety OF$^-$ and OF$^+$ GAS strains. These strains are listed in Table 1. Equivalent amounts of chromosomal DNAs from the various strains were cut with the restriction endonuclease *Hae*III, and the separated fragments were subjected to Southern hybridizations (43) with various DNA probes as diagrammed in Fig. 5. The same set of blots was used for all hybridizations.

The region immediately upstream of *emm1* is conserved among OF$^-$ strains. Probe 88279 is an M13 deletion subclone extending from an *Eco*RI site approximately 1 kb upstream of *emm1* to 115 bp upstream of *emm1* (Fig. 5). The nucleotide sequence of probe 88279 is 98% similar to the corresponding region in *emm12* (14, 37) and is also similar to reported sequences flanking *emm6* (22) and *emm24* (33). This upstream probe hybridized to variably sized *Hae*III fragments in all of the DNA digests tested when 20% base pair mismatch was allowed (Fig. 6). DNA fragments from all of the OF$^-$ M types, except for the M18 strain, hybridized strongly with probe 88279. Conversely, all DNAs from OF$^+$ M types hybridized weakly with probe 88279, and when stringency was adjusted to allow only 15% base pair mismatch, the less intense bands in these DNAs were not detectable (data not shown).

More distal upstream sequences are conserved among all serotypes. Probe pbf17 spanned a region that included the *Hae*III A, B, and C fragments from the M1 strain CS130 (Fig. 5). This probe includes all sequences contained in probe 88279 as well 2.1 kb of additional upstream sequence (including the entire *virR* gene) and the 5′ hypervariable end of the M1 structural gene. In contrast to the weaker hybridization signals to the 88279 portion seen among the OF$^+$ DNAs, the portion of pbf17 that spanned the *Hae*III B and C fragments (upstream of the 88279 portion) hybridized with fragments of identical sizes in nearly all digests, even when only 3% base pair mismatch was allowed (Fig. 7). At this stringency, the probe did not hybridize at all to the variably sized fragments from the OF$^+$ strains. Therefore, despite divergence in the more immediate upstream sequences as defined by probe 88279 in the OF$^+$ serotypes, sequences further upstream were highly conserved among all M types tested.

TABLE 1. GAS strains

Strain	M type	OF phenotype
CS130	1	−
CS192	3	−
CS193	5	−
CS142	6	−[a]
CS24	12	−[b]
71-687	14	−
71-690	18	−
CS126	19	−[c]
CS203	24	−[c]
74-754	55	−
PF1643	57	−
79-308	80	−
CS191	2	+
CS194	4	+
CS195	8	+
CS196	9	+
CS197	11	+
CS198	13	+
CS202	22	+
CS101	49	+
GT8237	59	+
72-466	62	+
CS110	76	+[d]

[a] Originally strain D471 from the Rockefeller University collection. This is the strain from which the M6 gene was cloned.
[b] Strain from which the M12 gene was cloned.
[c] Obtained from the V.A. Medical Center, Memphis, Tenn. CS203 was the strain from which the M24 gene was cloned.
[d] Strain from which the type II group A IgG Fc receptor gene was cloned.

Conserved OF$^+$-specific sequences upstream of the type II IgG Fc receptor. Plasmid pDH7 is a probe from the OF$^+$, M76 strain CS110 (19) and spans the same general region as does pbf17 (Fig. 5). This probe contains the 5′ end of the type II IgG Fc receptor gene (*fcrA76*) and 2.9 kb of upstream sequence. As with probe pbf17, pDH7 hybridized strongly to the *Hae*III B and C fragments in all chromosomal DNAs (Fig. 8) and at lowered stringency (30% base pair mismatch) hybridized to all of the same variably sized fragments that also hybridized to probes pbf17 and 88279 at lowered stringency (data not shown). At the higher stringency shown (10% base pair mismatch), however, pDH7 hybridized strongly to the variably sized fragments in all OF$^+$ strains and that of only one OF$^-$ strain, an M80 strain (Fig. 8). At this stringency, it did not hybridize or hybridized very weakly to the variably sized fragments in the remainder of the OF$^-$ strains. An additional *Hae*III fragment of 586 bp (*Hae*III-D) was present in all of the OF$^+$ strains (and the M80 strain). This fragment also hybridized with pbf17 at 30% base pair mismatch (data not shown). This *Hae*III D fragment maps between *Hae*III-B and the fragment analogous to *Hae*III-A in pDH7 (Fig. 5).

DISCUSSION

In this report, we show that *fcrA49* and *emm49* are separate, tandemly arrayed genes on the streptococcal chromosome. We infer, then, that IgG Fc-binding and antiphagocytic activities on the surface of the OF$^+$ GAS strain CS101 are expressed from independent proteins. We previously showed that *emm49* is flanked on its 3′ end by an M

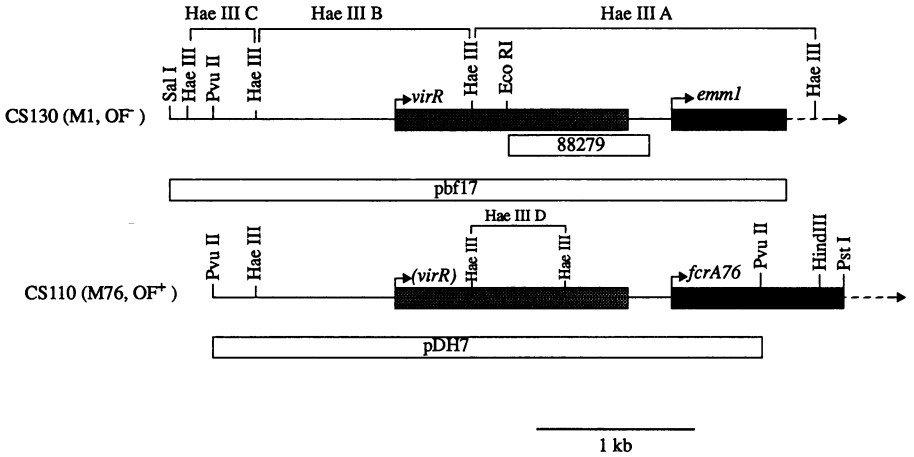

FIG. 5. Restriction maps of the regions upstream of *emm1* in GAS strain CS130 and *fcrA76* in GAS strain CS110. The hybridization probes used in this study are shown by the open boxes below the maps. The *Hae*III fragments of the CS130 clone and the additional *Hae*III fragment in the CS110 clone are marked.

protein-related gene (*ennX*), and we hypothesized that *emm49* was situated as the central gene in a triplication (15). Our data showing linkage of *emm49* downstream of an M protein-like Fc receptor gene confirms the previous gene triplication hypothesis. This family of M protein-related genes is located in the same general chromosomal location as are *emm1*, *emm6*, *emm12*, and *emm24*, as determined from upstream and downstream sequence similarities, including similarity to the sequence of the *scpA* gene (15). An updated schematic of this model, based on available data, is shown in Fig. 9.

We define *fcrA* as an M-like gene, because it is clear that the two genes arose from a common ancestor. This definition is in disagreement with that of Bessen and Fischetti (2), who

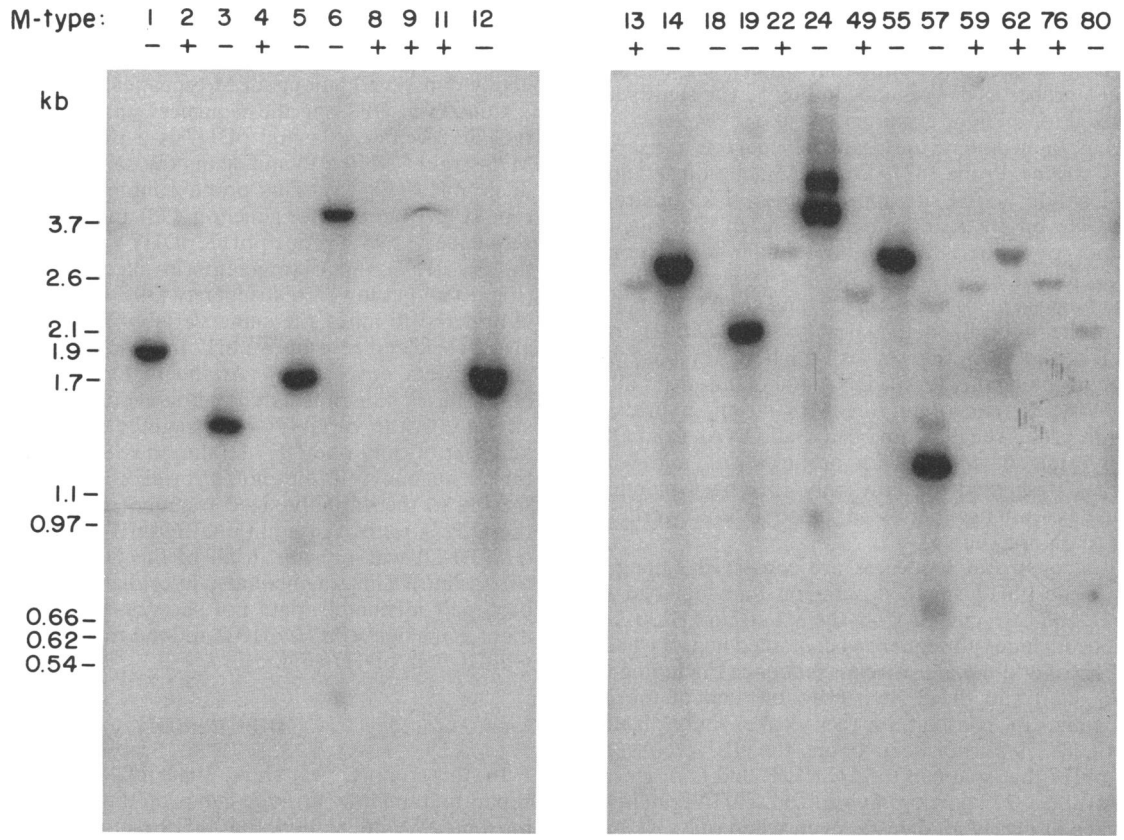

FIG. 6. *Hae*III chromosomal restriction digests of DNA from a variety of M types hybridized with probe 88279. Washes were done at 20°C below T_m. Positions of the size markers (bacteriophage lambda *Cla*I restriction fragments), shown at the left, were approximated from the stained gels.

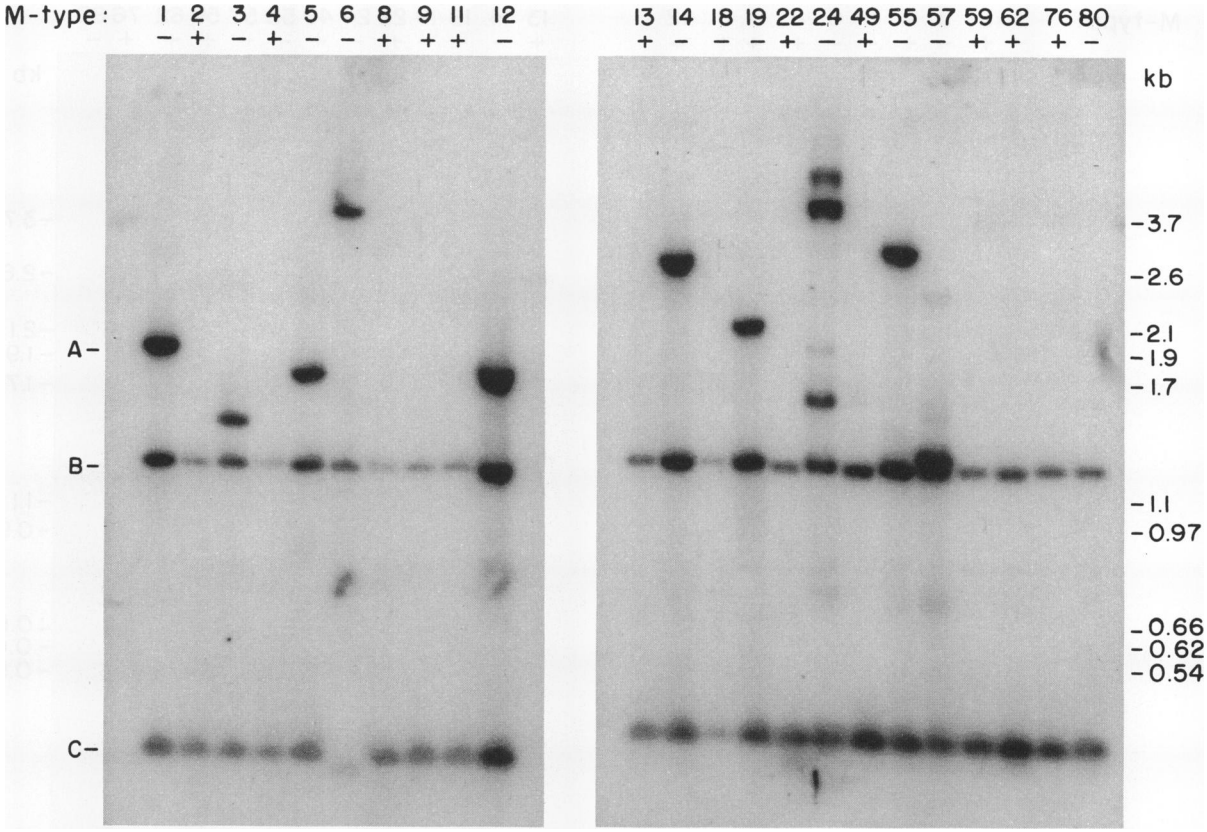

FIG. 7. *Hae*III chromosomal restriction digests of DNA from a variety of M types hybridized with probe pbf17. Washes were done at 3°C below T_m. The CS130 *Hae*III fragments that hybridize to pbf17 are labelled at the left. Positions of the size markers (bacteriophage lambda *Cla*I restriction fragments), shown at the right, were approximated from the stained gels.

define M-like genes as those having similarity to the *emm* C repeat region. The leader sequences of *fcrA76* and *emm49* are 75% similar, and portions of sequences downstream of the C repeat region exceed 80% similarity (20; data not shown). Given this extent of similarity, the genetic linkage of *fcrA49* and *emm49* further argues that the two genes are related, even though *fcrA* is clearly more divergent from the *emm* prototype than are *ennX* and its homologs.

Tandemly arrayed M protein-like genes are not a unique property of type 49 GAS. Bessen and Fischetti recently reported the sequence of two tandemly arrayed M-like genes (*emmL2.1* and *emmL2.2*) from an OF^+ M2 GAS (2), and the *arp4* IgA Fc receptor gene is also reported to be closely linked to a fibrinogen receptor gene, presumably *emm4* (28). Hybridization evidence suggests that *fcrA76* and *emm76* exist as a tandem array (17). In addition, preliminary hybridization data suggest that the sequence between *emm49* and *ennX* is conserved in the DNAs of all of the OF^+ strains that we tested (13). We hypothesize, then, that M protein-related gene families may be a common feature of OF^+ streptococcal strains.

Although no additional M-like genes have been identified in the OF^- strains from which *emm1*, *emm6*, *emm12*, and *emm24* were cloned and sequenced, we cannot rule out the possibility that multiple M-like gene families also occur in some OF^- strains. Gomi et al. (12) report the nucleotide sequence of a novel IgG Fc receptor (protein H) from an OF^-, M1 GAS strain. This M-like gene also appears to be the central gene in a tandem array of M-like genes, but the intervening sequences are not similar to those flanking *emm49* or *arp4*. Thus, this tandem array might have evolved separately from the OF^+ lineage. The M5 serotype is also reported to have multiple M-like sequences in its genome, but it is not clear whether they are tandemly arrayed with the *emm5* gene (23, 32). Bessen et al. (1) found that OF^- isolates from skin infections, as opposed to throat infections, were much more likely to express IgG Fc receptors, and this epidemiologic distinction might also correlate with the presence of multiple M-like genes. Further surveys of OF^- GAS strains will be required to determine the extent of multiple M-like gene arrays in this group.

In the model for duplications and triplications of genes (42), the initial duplication step is a very rare event caused by mismatched pairing of two daughter strands of DNA during replication, followed by unequal crossing over and nonhomologous recombination. This first step introduces a unique junction in the region between the duplicated genes. In a subsequent step, the duplicated genes on the two daughter strands can again participate in unequal crossing over, but this time by homologous recombination. The progeny of this process would be a daughter with a triplication and a daughter with a single gene. In the triplicated progeny, the two junctions between the three genes would be similar to each other.

Our sequencing results disagree with this prediction, as the intervening regions between *fcrA49* and *emm49* and between *emm49* and *ennX* are very dissimilar. Possibly, this triplication occurred by a different mechanism, or the tripli-

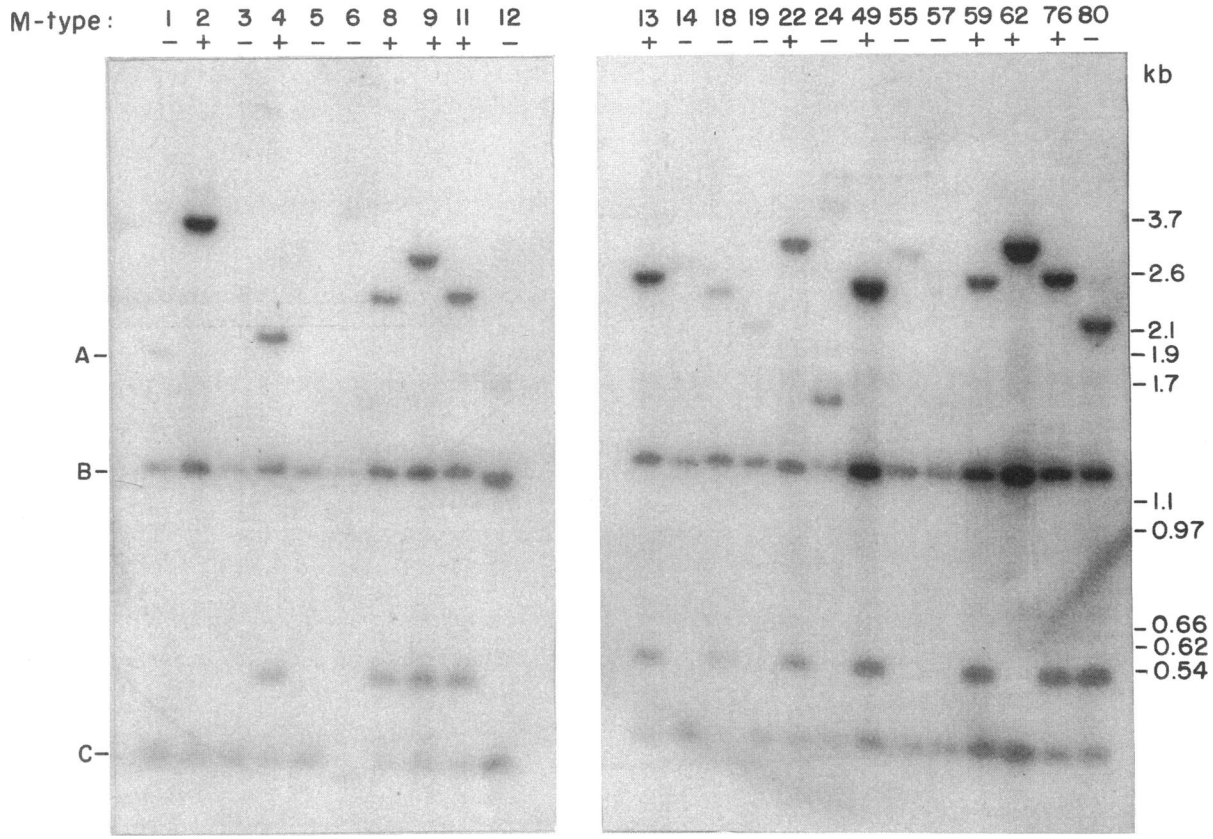

FIG. 8. *Hae*III chromosomal restriction digests of DNA from a variety of M types hybridized with probe pDH7. Washes were done at 10°C below T_m. The CS130 (M1) *Hae*III fragments that hybridize to pbf17 are labelled at the left. Positions of the size markers (bacteriophage lambda *Cla*I restriction fragments), shown at the right, were approximated from the stained gels.

cation took place very early in the evolution of streptococci, giving time for subsequent divergence by point mutations. This latter hypothesis is supported by the high degree of similarity between the regions upstream and downstream of *emm49* and *arp4*. The degree of conservation seen between different serotypes might reflect subsequent selective pressure to conserve transcriptional and regulatory control regions of the various genes. These similarities also suggest that the generation of the multiple gene families in OF^+ strains predated the generation of multiple M protein serotypes.

The arrangement of the multiple genes in different serotypes also supports the hypothesis of early establishment of the multiple M-like arrays. *arp4*, an IgA Fc receptor gene, is suggested to be distinct from the fibrinogen receptor gene, even though it appears to be located in the same chromo-

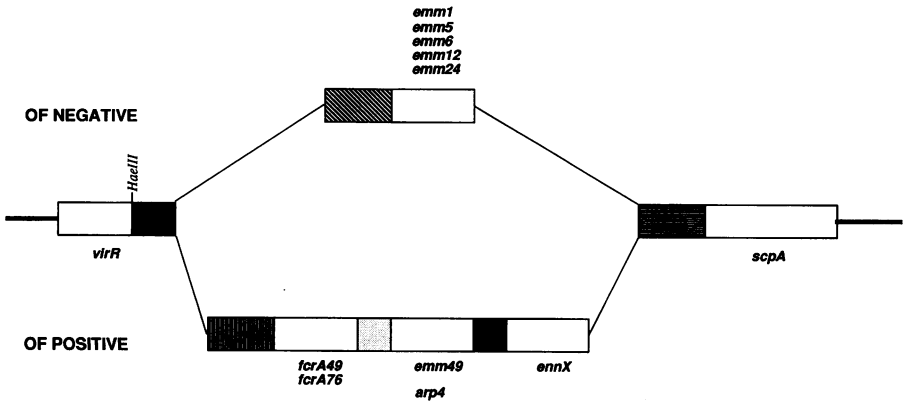

FIG. 9. Schematic comparison of the *vir* regulons of OF^- and OF^+ GAS compiled from available restriction mapping and nucleotide sequence data. The various genes are depicted as open boxes, and nonhomologous intervening sequences are depicted as variously shaded boxes.

somal location as is *emm49* and shares significant sequence similarity with *emm49* throughout the constant region. Therefore, the arrangement of M protein and other M protein-related genes in these multiple arrays could vary between different serotypes. Implicit in this hypothesis is the possibility for functional and antigenic variation via genetic exchanges between the multiple genes.

Also according to existing gene duplication theory, duplicated and triplicated genes arrays should be highly prone to expansion or contraction as a result of the opportunities for homologous recombination. In fact, one explanation for the dearth of multiple M-like genes in the OF^- serotypes that have been studied is that the multiple genes, once present, subsequently were deleted. If a triplicated array of genes became functionally advantageous to the organism, though, they could potentially become fixed in the chromosome via selective pressure. This appears to be the case with the OF^+ strains. At least two of the M protein-related genes in the M49 strain CS101 make functional proteins (*emm49* and *fcrA49*), and both could potentially benefit the cell in its effort to avoid immune surveillance and destruction by the host. Similarly, the *ennX* homolog from an M2 strain was shown to bind IgA, which could play a role in survival of the organism (2). Although we showed that *ennX* was expressed in *Escherichia coli* (15), we have not been able to demonstrate expression (or a transcript) in GAS (25). If indeed *ennX* is a pseudogene in some cases, we would expect that point mutations would accumulate, such as is seen in the ATG codon of the M4 homolog of *ennX*. Examination of *ennX* homologs in other streptococcal strains should help to determine the extent to which this gene is functional in OF^+ GAS.

Hybridizations with DNA probes upstream of the cloned *emm1* and *fcrA* genes concurred with the previous observations by showing that at least two distinct groups of GAS could be differentiated with respect to the region between *virR* and the locus that we will call the M-like expression site. Strong DNA similarity in this region was apparent within each of these groups, and these groupings correlated strongly, but not absolutely, with OF phenotype. These results extend the antibody binding and Fc receptor studies reported by Bessen et al. (1, 3, 4), who correlated differences in the M protein constant regions and the presence or absence of and IgG Fc receptor with OF phenotype.

Probe 88279, which defined the region immediately upstream of *emm1*, hybridized much more strongly to DNAs from OF^- serotypes than to DNAs from OF^+ serotypes. Conversely, probe pDH7 from the OF^+ type 76 strain CS110 showed that the sequences immediately upstream of *fcrA76* are equally conserved among the OF^+ serotypes. We suspect that the heterogeneous sizes of the hybridizing fragments were the result of hypervariable sequences present in the 5′ ends of the respective M-like structural genes. Variant restriction fragments were also detected by Scott et al. (39) when chromosomal digests of different streptococcal strains were hybridized with DNA probes containing portions of the *emm6* gene.

The pattern of the M80 strain used in this study was an exception to the hybridization patterns observed among the OF^- and OF^+ serotypes in that it resembled the pattern typical of the OF^+ serotypes. This strain, however, proved in several trials to be M protein positive but OF negative (data not shown). We must assume, then, that exceptions to the correlation between the hybridization pattern and OF phenotype exist.

Despite a divergence in sequences more immediately upstream (within 1 kb) of the OF^- M protein expression site, hybridization with probes pbf17 and pDH7 showed that sequences further upstream (the *Hae*III B and C fragments) were highly conserved among all M types tested, regardless of OF phenotype. Thus, the sequence divergence between the two streptococcal groups is localized to a region within 1 kb of the M-like expression site.

The sequence divergence seen in the region upstream of the M-like expression site is significant in light of a regulatory gene identified upstream of the M protein structural gene in several streptococcal strains. This gene, *virR*, controls (in *trans*) expression of the linked *emm* and *scpA* genes, and most likely also *fcrA* (26). Although a putative DNA binding motif of *virR* is encoded in the more conserved upstream portion of the gene (8, 36), the 3′ end of *virR* definitely lies within 1 kb of the M protein structural gene and thus includes the region of DNA dissimilarity between the OF^- and OF^+ strains. Therefore, variations in the primary sequence or conformation of *virR* could potentially result in altered regulatory activity between the two groups of strains.

We conclude, then, that two distinct divisions of streptococcal serotypes exist. OF^+ serotypes, and maybe some OF^- serotypes, feature genetically stable arrays of multiple M protein-like genes, including a type IIa IgG Fc receptor. Furthermore, variations exist between the two groups not only within the M-like genes but also in noncoding DNA sequences within the *vir* regulon.

ACKNOWLEDGMENTS

We thank Jodi Peterson for expert assistance in sequencing and Jesse Goodman for use of the PCR machine and reagents. We are grateful to Cecil Chen for helpful discussions and a critical reading of the manuscript.

This work was supported by Public Health Service grant AI16722 from the National Institute of Allergy and Infectious Diseases to P.P.C.

REFERENCES

1. **Bessen, D., and V. A. Fischetti.** 1990. A human IgG receptor of group A streptococci is associated with tissue site of infection and streptococcal class. J. Infect. Dis. **161**:747–754.
2. **Bessen, D., and V. A. Fischetti.** 1991. Nucleotide sequences of two adjacent M or M-like protein genes of group A streptococci: different RNA transcript levels and identification of a unique immunoglobulin A-binding protein. Infect. Immun. **60**:124–135.
3. **Bessen, D., K. F. Jones, and V. A. Fischetti.** 1989. Evidence for two distinct classes of streptococcal M protein and their relationship to rheumatic fever. J. Exp. Med. **169**:269–283.
4. **Bessen, D. E., and V. A. Fischetti.** 1990. Differentiation between two biologically distinct classes of group A streptococci by limited substitutions of amino acids within the shared region of M protein-like molecules. J. Exp. Med. **172**:1757–1764.
5. **Boyle, M. D. P. (Medical College of Ohio).** Personal communication.
6. **Boyle, M. D. P., E. L. Faulmann, R. A. Otten, and D. G. Heath.** 1990. Streptococcal immunoglobulin binding proteins, p. 19–44. *In* E. M. Ayoub (ed.), Microbial determinants of virulence and host response. American Society for Microbiology, Washington, D.C.
7. **Caparon, M. G., and J. R. Scott.** 1987. Identification of a gene that regulates expression of M protein, the major virulence determinant of group A streptococci. Proc. Natl. Acad. Sci. USA **84**:8677–8681.
8. **Chen, C. C.** 1990. Ph.D thesis. University of Minnesota, Minneapolis.
9. **Chen, C. C., and P. P. Cleary.** 1989. Cloning and expression of the streptococcal C5a peptidase gene in *Escherichia coli*: linkage to the type 12 M protein gene. Infect. Immun. **57**:1740–1745.

10. **Collins, M. L., and W. R. Hunsaker.** 1985. Improved hybridization assays employing tailed oligonucleotide probes: a direct comparison with 5'-end labelled oligonucleotide probes and nick translated plasmid probes. Anal. Biochem. **151:**211–224.
11. **Frithz, E., L.-O. Heden, and G. Lindahl.** 1989. Extensive sequence homology between IgA receptor and M proteins in *Streptococcus pyogenes*. Mol. Microbiol. **3:**1111–1119.
12. **Gomi, H., T. Hozumi, S. Hattori, C. Tagawa, F. Kishimoto, and L. Bjorck.** 1990. The gene sequence and some properties of protein H. A novel IgG-binding protein. J. Immunol. **144:**4046–4052.
13. **Haanes-Fritz, E.** 1989. Ph.D. thesis. University of Minnesota, Minneapolis.
14. **Haanes-Fritz, E., W. Kraus, V. Burdett, J. B. Dale, E. H. Beachey, and P. P. Cleary.** 1988. Comparison of the leader sequences of group A streptococcal M protein genes. Nucleic Acids Res. **16:**4667–4677.
15. **Haanes, E. J., and P. P. Cleary.** 1989. Identification of a divergent M protein gene and an M protein-related gene family in *Streptococcus pyogenes* serotype 49. J. Bacteriol. **171:**6397–6408.
16. **Hallas, G., and J. P. Widdowson.** 1982. The opacity factor of group A streptococci. J. Med. Microbiol. **15:**451–464.
17. **Heath, D. G.** 1988. Ph.D. thesis. University of Minnesota, Minneapolis.
18. **Heath, D. G., M. D. P. Boyle, and P. P. Cleary.** 1990. Isolated DNA repeat region from *fcrA76*, the Fc-binding protein gene from an M-type 76 strain of group A streptococci, encodes a protein with Fc-binding activity. Mol. Microbiol. **4:**2071–2079.
19. **Heath, D. G., and P. P. Cleary.** 1987. Cloning and expression of the gene for an immunoglobulin G Fc-receptor protein from a group A streptococcus. Infect. Immun. **55:**1233–1238.
20. **Heath, D. G., and P. P. Cleary.** 1989. Fc-receptor and M-protein genes of group A streptococci are the products of gene duplication. Proc. Natl. Acad. Sci. USA **86:**4741–4745.
21. **Hill, M. J., and L. W. Wannamaker.** 1968. The serum opacity reaction of *Streptococcus pyogenes*: general properties of the streptococcal factor and of the reaction in aged serum. J. Hyg. (Cambridge) **66:**37–48.
22. **Hollingshead, S. K., V. A. Fischetti, and J. R. Scott.** 1986. Complete nucleotide sequence of type 6 M protein of the group A streptococcus—repetitive structure and membrane anchor. J. Biol. Chem. **261:**1677–1686.
23. **Kehoe, M. A., T. P. Poirier, E. H. Beachey, and K. N. Timmis.** 1985. Cloning and genetic analysis of serotype 5 M protein determinant of group A streptococci: evidence for multiple copies of the M5 determinant in the *Streptococcus pyogenes* genome. Infect. Immun. **48:**190–197.
24. **Lancefield, R. C.** 1962. Current knowledge of type-specific M antigens of group A streptococci. J. Immunol. **89:**301–313.
25. **La Penta, D., and P. P. Cleary.** Unpublished results.
26. **La Penta, D., E. J. Haanes, X. P. Zhang, D. G. Heath, and P. P. Cleary.** 1990. Coordinate control of the IgG fc-receptor with other virulence factors in group A streptococci, abstr. B-249, p. 68. Abstr. 90th Annu. Meet. Am. Soc. Microbiol. 1990.
27. **Lethbak, A. C., C. Christiansen, and A. Stenderup.** 1970. Bacterial genome sizes determined by DNA renaturation studies. J. Gen. Microbiol. **64:**377–380.
28. **Lindahl, G.** 1989. Cell surface proteins of a group A streptococcus type M4: the IgA receptor and a receptor related to M proteins are coded for by closely linked genes. Mol. Gen. Genet. **216:**372–379.
29. **Maxted, W. R.** 1978. Group A streptococci: pathogenesis and immunity, p. 107–125. *In* F. A. Skinner and L. B. Quesnel (ed.), Streptococci. Academic Press, New York.
30. **Meinkoth, J., and G. Wahl.** 1984. Hybridization of nucleic acids immobilized on solid supports. Anal. Biochem. **138:**267–284.
31. **Messing, J.** 1983. New vectors for cloning. Methods Enzymol. **101:**20–78.
32. **Miller, L., L. Gray, E. H. Beachey, and M. Kehoe.** 1988. Antigenic variation among group A streptococcal M proteins. Nucleotide sequence of the serotype 5 M protein gene and its relationship with genes encoding types 6 and 24 M proteins. J. Biol. Chem. **263:**5668–5673.
33. **Mouw, A., E. H. Beachey, and V. Burdett.** 1988. Molecular evolution of streptococcal M-protein: cloning and nucleotide sequence of the type 24 M protein gene and relation to other genes of *Streptococcus pyogenes*. J. Bacteriol. **170:**676–684.
34. **Mullis, K. B., and F. A. Faloona.** 1978. Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. Methods Enzymol. **155:**335–350.
35. **Musser, J. M., A. R. Hauser, M. H. Kim, P. M. Schlievert, K. Nelson, and R. K. Selander.** 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.
36. **Perez-Casal, J., M. G. Caparon, and J. R. Scott.** 1991. *mry*, a *trans*-acting positive regulator of the M protein gene of *Streptococcus pyogenes* with similarity to the receptor proteins of two-component regulatory systems. J. Bacteriol. **173:**2617–2624.
37. **Robbins, J. C., J. G. Spanier, S. J. Jones, W. J. Simpson, and P. P. Cleary.** 1987. *Streptococcus pyogenes* type 12 M protein gene regulation by upstream sequences. J. Bacteriol. **169:**5633–5640.
38. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
39. **Scott, J. R., W. M. Pulliam, S. K. Hollingshead, and V. A. Fischetti.** 1985. Relationship of M protein genes in group A streptococci. Proc. Natl. Acad. Sci. USA **82:**1822–1826.
40. **Simpson, W. J., D. LaPenta, C. C. Chen, and P. P. Cleary.** 1990. Coregulation of type 12 M protein and streptococcal C5a peptidase genes in group A streptococci: Evidence for a virulence regulon controlled by the *virR* locus. J. Bacteriol. **172:**696–700.
41. **Singh, L., and K. W. Jones.** 1984. The use of heparin as a simple, cost-effective means of controlling background in nucleic acid hybridization procedures. Nucleic Acids Res. **12:**5627–5638.
42. **Smith, G.** 1976. Evolution of repeated DNA sequences by unequal crossover. Science **191:**528–535.
43. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98:**503–517.
44. **Top, F. H., and L. W. Wannamaker.** 1968. The serum opacity reaction of *Streptococcus pyogenes*: demonstration of multiple, strain-specific lipoproteinase antigens. J. Exp. Med. **127:**1013–1034.
45. **Top, F. H., and L. W. Wannamaker.** 1968. The serum opacity factor of *Streptococcus pyogenes*: frequency of production of streptococcal lipoproteinase by strains of different serological types and the relationship to M protein production. J. Hyg. (Cambridge) **66:**49–58.
46. **Wexler, D. E., R. D. Nelson, and P. P. Cleary.** 1983. Human neutrophil chemotactic response to group A streptococci: bacteria-mediated interference with complement-derived chemotactic factors. Infect. Immun. **39:**239–246.
47. **Yanisch-Perron, C., J. Vieira, and J. Messing.** 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13 mp and pUC vectors. Gene **33:**103–119.