Mosaicism in the alpha-like protein genes of group B streptococci

C. S. Lachenauer*†‡, R. Creti*§, J. L. Michel*¶, and L. C. Madoff*

*Channing Laboratory and Division of Infectious Diseases, Department of Medicine, Brigham and Women's Hospital, †Division of Infectious Diseases, Children's Hospital, and ¶Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, MA; and §Laboratory of Bacteriology and Medical Mycology, Istituto Superiore di Sanità, Rome, Italy

Communicated by Frederick M. Ausubel, Harvard Medical School, Boston, MA, June 16, 2000 (received for review March 28, 2000)

Members of a family of repeat-containing surface proteins of group B streptococci (GBS) defined by the alpha C and Rib proteins exhibit size variability and cross-reactivity and have been studied as potential vaccine components. We report evidence of horizontal DNA transfer with subsequent recombination as a mechanism generating diversity within this antigen family. Alp2 and Alp3 are additional members of the alpha C protein family identified in strains of the emerging GBS serotypes V and VIII. Each contains an overall genetic organization highly similar to that of the alpha C and Rib proteins, including a tandem repeat region and conserved N- and C-terminal regions. Among different strains, protein size varies according to the number of tandem repeats within the corresponding gene. Unlike the alpha C and Rib proteins, however, the newly described alpha-like proteins contain other regions, including one similar to the IgA-binding region of the GBS beta C protein, a nontandem repeat region, and an isolated repeat highly homologous to the alpha C repeat. Sequence analysis of the regions flanking the alpha C protein gene on a 13.7-kb insert reveals several ORFs that are likely to be involved in basic metabolic pathways. Analysis of corresponding flanking regions in other GBS strains, including the parent strains of the newly described alpha-like proteins, shows striking conservation among all strains studied. These findings indicate that the alpha-like proteins are encoded by mosaic variants at a single genomic locus and suggest that recombination after horizontal DNA transfer is a means of generating diversity within this protein family.

Group B streptococcus (GBS; *Streptococcus agalactiae*) is an important cause of invasive infection in newborns, pregnant women, and immunocompromised adults. Present on most GBS strains is a surface-associated protein antigen belonging to a family of GBS proteins, of which the alpha C protein is the prototype (1–5). Although the precise role of the alpha C and related proteins in the pathogenesis of GBS disease is not known, deletion of the alpha C protein gene results in attenuated virulence of GBS in animals (6). The ability of the alpha C protein and other related proteins to elicit antibody-mediated protection in experimental animals has led to investigation into their potential role in maternal GBS vaccines.

A striking feature of the gene encoding the alpha C protein is the series of identical, tandem repeating subunits that make up the majority of the gene (7). After a typical Gram-positive signal sequence and a unique N-terminal encoding region, the prototype alpha C protein from type Ia/C strain A909 contains a series of nine identical, 246-bp tandem repeating subunits, a partial identical repeat 33 bp in length, and a region encoding a C-terminal cell wall-associating motif common to many Gram-positive surface proteins (8). Other GBS strains possess alpha C proteins of varying sizes, corresponding to varying numbers of genetic repeats (9). The existence of a protein family was suggested by the observation that the Rib protein, purified from a type III GBS strain, is a homolog of the alpha C protein gene. It has a similar N-terminal encoding region, a series of 12 237-bp identical tandem repeats, and a highly conserved C-terminal encoding region (10). Phenotypic characteristics of alpha-like proteins from type V and VIII GBS suggest that these proteins also are related (4, 5, 11) and likely possess a similar genetic organization. Some of the proteins, such as the alpha C and Rib proteins, are immunologically distinct from each other, but others, such as the alpha C and an alpha-like protein from type V GBS, cross-react on Western blots (4, 5).

A common feature among several Gram-positive surface protein families is an ability to display variation within antigenic epitopes, presumably as a means of evading host immunity. For example, in the M protein family of group A streptococci, a variety of mechanisms including point mutations (12), homologous recombination between repeat regions (13), and horizontal transfer of DNA (12, 14) have been reported to contribute to sequence diversity. Pneumococcal surface protein A (PspA) is highly variable in the surface-exposed N-terminal region; mechanisms accounting for this variability have not been clearly defined, although mutations and horizontal genetic exchange have been suggested (15). In GBS, we showed previously that deletions in the tandem repeat region are an important means by which antigenic diversity is generated within the alpha C protein (16, 17). Isolates with reduced numbers of repeats are able to escape recognition by host antibodies elicited by the full-length protein.

Here, we report the sequences of the genes encoding two members of the alpha C protein family from the emerging GBS serotypes V and VIII (4, 5). Like the alpha C and Rib protein genes, these genes each contain a tandem repeat region and conserved N- and C-terminal encoding regions, but they also contain an additional nontandem repeat region. Comparison of the genes encoding alpha-like proteins from types V and VIII GBS with other GBS protein genes reveals that the newly described protein genes are mosaics, that is, they appear to be composed of conserved structures rearranged from genes encoding different GBS proteins. Comparison of the regions flanking the alpha C and related protein genes indicates that these genes are alleles at a specific genomic locus. These results indicate that recombination of horizontally acquired DNA is a means by which antigenic diversity is generated in the alpha C protein family of GBS.

Materials and Methods

Bacterial Strains, Plasmids, Proteins, and Antisera. GBS type V strains G-107 and HIA-00086 (blood isolates), H4B-0052 (rectal isolate), and H4A-0158 (vaginal isolate) were studied previously as part of a panel of type V strains (4). Type V strain CJB-110 and type VIII strains JM9–130013, SMU 071, and SMU 093 have been described (5). Type Ia/C strain A909 (alpha C protein prototype strain) (7),

Abbreviations: GBS, group B streptococci; Esp, enterococcal surface protein.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF291065–AF291072).

‡To whom reprint requests should be addressed at: Channing Laboratory, 181 Longwood Avenue, Boston, MA 02115. E-mail: catherine.lachenauer@channing.harvard.edu.

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.

type II strain 18RS21 (nonexpresser of alpha C or an alpha-like protein) (4), and type III strain BM-110 (Rib protein prototype strain) (10) have been described. pJMS23 is a 13.7-kb restriction fragment isolated from GBS strain A909 that contains the complete alpha C protein structural gene (1). pCRII and *Escherichia coli* strain INVα were purchased (Invitrogen), as was *E. coli* strain XLI-Blue (Stratagene). Alpha-like proteins purified from strains CJB-110 and JM9–130013 and rabbit antisera raised to these proteins have been described (4, 5).

Cloning and Sequencing of the Alpha-Like Protein Genes from Types V and VIII GBS. Chromosomal DNA was prepared from GBS by phenol/chloroform extraction (16) or by a colony chromosomal DNA method (18). Amplification of chromosomal DNA was carried out by using the following primers synthesized according to sequence flanking *bca*, the gene encoding the alpha C protein: 5′-TGGTGGACAAGAAAAAGTTCT-3′ (forward primer) and 5′-TGTTCACACCAATAAATGGTGA-3′ (reverse primer). PCR products amplified from strains CJB-110 and JM9–130013 were cloned into *E. coli* by using pCRII according to the manufacturer's instructions (Invitrogen). Expression of the alpha-like protein genes was confirmed by Western blotting of SDS extracts of clones with homologous antiserum (4).

Amplified product pooled from at least 10 PCRs was sequenced on an automated sequencer (ABI PRISM 377 DNA Sequencer; Perkin–Elmer/Cetus) at the Brigham and Women's Hospital DNA Sequencing Facility. Initial sequencing reactions were performed with the primers used to amplify the genes; additional primers were synthesized according to new sequence (Midland Certified Reagent, Midland, TX). Sequence assembly and editing were performed with SEQUENCHER 3.0 (Gene Codes, Ann Arbor, MI). GCG8 was used for sequence analysis (Genetics Computer Group, Madison, WI). GenBank searches were performed with Basic Local Alignment Search Tool (BLAST). For homology searches with the enterococcal genome, preliminary sequence data were obtained from The Institute for Genomic Research web site at http://www.tigr.org. For homology searches with the GAS genome, preliminary sequence data were obtained from the University of Oklahoma *Streptococcus pyogenes* Genome Blast Server at http://www.genome.ou.edu/strep_blast.html.

Restriction Enzyme Analysis of Alpha-Like Protein Gene Repeats. The tandem repeat region of the alpha-like protein genes from strains JM9–130013, H4A-0158, and HIA-00086 were amplified as described above by using primers flanking these regions: 5′-GATGATCTTAAAGCTAAGTAT-3′ (forward primer) and 5′-GCTGGTAGTTTATTTCCTTTACC-3′ (reverse primer). PCR amplicons were digested overnight with *Bgl*II or *Hpa*II at 37°. Digested DNA was visualized by electrophoresis on ethidium bromide-stained 2.0% agarose gel.

Analysis of Regions Flanking the Alpha C and Alpha-Like Protein Genes. The 13.7-kb insert of pJMS23, exclusive of the previously sequenced *bca* gene (7), was sequenced at the Harvard Medical School Microbiology DNA Core Facility and at the Beth Israel Deaconess Medical Center Molecular Medicine Unit with automated sequencers (ABI Prism 377 DNA Sequencer; Perkin–Elmer/Cetus). For the flanking region 5′ to *bca* and for approximately 2.1 kb immediately 3′ to *bca*, overlapping sequences derived from enzyme digestion of pJMS23 were subcloned into *E. coli* XLI-Blue by using the phagesmid pBluescript (Stratagene). The remaining 3.9-kb downstream flanking region was amplified and sequenced directly from pooled PCRs. Initial sequencing primers were based on known sequence of the vector or on primers used to generate PCR product; additional primers were synthesized according to new sequence (Great American Gene Company, Ramona, CA). GCG8 was used for sequence analysis. Putative functions were assigned to ORFs according to homology searches of GenBank by using BLAST.

PCR and selected sequencing were used to generate maps of the flanking regions of the alpha-like protein genes in the following strains: CJB-110 (Alp2), JM9–130013 (Alp3), BM110 (Rib), and 18RS21 (alpha-like protein nonexpresser). Chromosomal DNA was prepared from these strains as described above. For the PCR analysis, the following primer pairs were synthesized according to the sequence of pJMS23 (Fig. 3): N-1, 5′-ACGATACCACCACAAGTCTGAA-3′ (forward primer) and 5′-GGAGAACTTTTTCTTGTCCACC-3′ (reverse primer); N-2, 5′-GGGTCAATAACATTGCTACGCCAG-3′ (forward primer) and 5′-CCGTTGTTTCGTCTGTGAAGTATC-3′ (reverse primer); C-1 5′-CCAGCAACAGGTGAGAATGCAACTC-3′ (forward primer) and 5′-ATGGTAATGCTCAGTTCCGAGTTC-3′ (reverse primer); and C-2 5′-GGTCATTTGCTTCTGTAAGTGGAAG-3′ (forward primer) and 5′-GACACTGCCAAGCTGATCGATCAAC-3′ (reverse primer).

Where indicated, PCR amplicons were sequenced by using the primers used to generate the PCR fragments as the initial sequencing primers.

Results

Cloning and Sequencing of Alpha-Like Proteins from Types V and VIII GBS. Amplification of chromosomal DNA prepared from GBS strains CJB-110 (type V) and JM9–130013 (type VIII) with primers directed to regions flanking the alpha C protein gene yielded products of approximately 2.8 and 3.1 kb, respectively. The PCR products each were cloned into *E. coli* strain INVα. Western blotting of SDS extracts of the clones with antiserum raised to the homologous protein demonstrated protein expression (data not shown).

Analysis of the PCR amplicon from type V strain CJB-110 demonstrated a single ORF of 2,358 bp (*alp2*) (GenBank accession no. AF208158) encoding a protein of 786 aa with a predicted molecular size of 78,855 Da (Alp2) (Fig. 1*A*). Analysis of the PCR amplicon from type VIII strain JM9–130013 demonstrated a single ORF of 2,595 bp (*alp3*) (GenBank accession no. AF245663), encoding a protein of 865 aa with a predicted molecular size of 87,680 Da (Alp3) (Fig. 1*B*). Because the repeat region of the *alp3* is too large to be spanned by a single sequencing reaction, the complete sequence was derived from two contiguous sequences, one each originating from the 5′ and the 3′ ends of the amplicon. The number of repeats (five) within the repeat region of *alp3* was calculated from the PCR amplicon size.

For both *alp2* and *alp3*, the putative prokaryotic promoter consensus (TATAAT) begins 45 bases upstream from the initiation codon and is identical to the corresponding region of the alpha C protein gene. For each, the length of the 56-aa signal peptide was determined by comparison of the deduced sequence with the actual amino acid sequence determined by Edman degradation of the purified proteins (4, 5). Homology studies with the alpha C and Rib protein sequences show that the N-terminal region common to Alp2 and Alp3 is 172 aa in length, with a calculated molecular size of 18,932 Da, and is 70% identical to the N-terminal sequence of the alpha C protein and 60% identical to that found in Rib (7, 10) (Fig. 1). Immediately adjacent to the N-terminal region in both Alp2 and Alp3 is an "A" region containing 51 aa with a predicted molecular size of 5.8 kDa. A "U" region of 137 aa with a predicted molecular mass of 14.9 kDa follows. The U region is 34% identical to a region within the beta C protein, a nontandem repeat-containing GBS surface protein that binds to the Fc portion of human IgA (19, 20). Western blot analysis showed no binding of IgA or beta-specific antibody to Alp2 or Alp3 (data not shown).

After the U region, the sequences of Alp2 and Alp3 abruptly diverge. In Alp2, a second (i.e., nontandem) A repeat follows, identical at the amino acid level to the first and differing at the

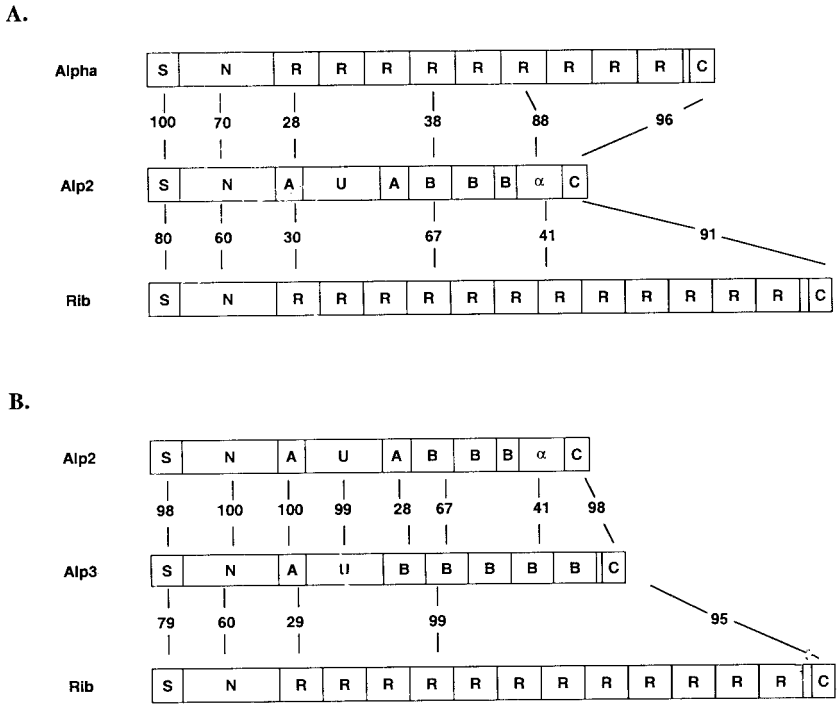

Fig. 1. Structure of alpha-like proteins from type V strain CJB-110 (Alp2) and type VIII strain JM9–130013 (Alp3). (*A*) Structure of Alp2 compared with the alpha C protein (*Top*) and protein Rib (*Bottom*). (*B*) Structure of Alp3 compared with Alp2 (*Top*) and protein Rib (*Bottom*). Percentages refer to amino acid identity between corresponding regions. Alpha C protein sequence is from ref. 7. Rib sequence is from ref. 10.

nucleotide level by only 1 bp. After the second A repeat is a "B" repeat region containing two complete repeats of 76 aa (differing at the nucleotide level by 1 bp) and one partial repeat of 41 aa (identical at the nucleotide level to the first B repeat). The deduced amino acid sequence of the B repeat is 67% identical to the Rib repeat and 38% identical to the alpha C protein repeat. Immediately after the B repeat region is an 82-aa sequence that is 88% identical to the alpha C protein repeat. The 45-aa C-terminal region is 96% identical to the corresponding region of the alpha C protein and 91% identical to the C-terminal region of the Rib protein. This region incorporates a cell wall-associating motif (LPXTGX) common to many Gram-positive surface proteins.

In Alp3, the U region is followed by a B repeat region containing five tandem repeats, each 79-aa long and 99% identical to the repeat region of the Rib protein and a partial repeat 13 aa long. The first 39 bp of the first B repeat are highly similar, but not identical, to the first 39 bp in subsequent B repeats and may represent a junctional area resulting from a recombination event. Otherwise, the tandem repeats are identical to each other at the nucleotide level except for a single base-pair substitution present in alternating repeats. After the B repeat region is a C-terminal region that is nearly identical to that of Alp2. Except for the number of repeats, the sequence of this protein is 98% identical to the recently published sequence of R28 from *S. pyogenes* (21).

Size of the Alpha-Like Proteins from Types V and VIII GBS Is Determined by the Size of the Tandem Repeat Region. PCR analysis of the alpha C protein from different GBS isolates demonstrated that variation in the number of repeats on a Western blot corresponded to the number of tandem nucleotide repeats (16). It is hypothesized that the same phenomenon is responsible for the size variability of protein Rib (10). Here, we show that the number of repeats within the tandem repeat regions of the Alp3 gene determines the size of the corresponding protein (Fig. 2). Western blot analysis of six additional type V and VIII strains using antiserum to Alp3 demonstrated alpha-like proteins of various sizes. The genes encoding these proteins were amplified from the parent strains and sequenced, revealing that each of the genes is identical to *alp3* except for the number of tandem B repeats. For four strains with three or fewer repeats, the entire gene was sequenced. For two additional strains (H4A-0158 and HIA-00086) in which the tandem repeat region was too large to be spanned by a single sequencing reaction, the nonrepeat regions of the gene and at least two repeats at either end of the repeat region were sequenced. To confirm that the remainder of the repeat region was composed entirely of tandem repeats, the region was amplified from each strain and subjected to endonuclease digestion with *Bgl*II or with *Hpa*II, each of which recognizes only one site within each tandem repeat. Agarose gel electrophoresis of the digested product demonstrated a prominent band of 237 bp, corresponding to the size of the individual repeat, and two additional bands, corresponding to the predicted size of the fragments at either end of the repeat region (data not shown).

Analysis of the Flanking Regions of Alpha-Like Protein Genes in a Panel of GBS Strains. Similarities and divergences in the chromosomal regions flanking the alpha-like protein genes from selected GBS strains were analyzed to determine whether these genes were present in a conserved locus on the GBS chromosome and to identify features that might suggest a virulence cassette or other mobile genetic element. First, the regions flanking the *bca* gene contained on the 13.7-kb plasmid pJMS23 were sequenced [GenBank accession no. AF248037 (upstream region) and AF248038 (downstream region)]. Analysis of these regions revealed 14 ORFs (Fig. 3), most of which appear to be housekeeping genes, based on BLAST analysis (Table 1). ORF 7 appears to be a truncated integrase on the basis of sequence homology with integrases from Gram-positive bacteriophages. No other genes involved with integration were apparent. No significant homology with an alpha-like protein gene was found by BLAST search in the GAS or enterococcal genome databases.

Maps of the chromosomal regions flanking the alpha-like protein genes from strains CJB-110 (Alp2), JM9–130013 (Alp3),

Fig. 2. Correlation of the number of tandem repeats with size of alpha-like proteins from type V and VIII GBS strains. (*A*) Western blot of SDS extracts of clinical type V and VIII GBS strains with antiserum raised to purified alpha-like protein from strain JM9–130013 demonstrating protein size variation. (*B*) Structure of corresponding alpha-like protein genes. Except for the number of repeats in the tandem repeat region, all sequences were identical (>99%) to Alp3. The size of the tandem repeat region correlated with protein size for each strain. Generation of the laddering patterns on the Western blot is presumed to be a result of proteolysis at specific sites within each tandem repeat (10).

BM-110 (Rib), and 18RS21 (alpha-like protein nonexpresser) were constructed by using data obtained by PCR and partial sequencing and show that flanking regions in all strains studied are highly conserved (Fig. 3). PCR using the N-1 primer pair generated a product of approximately 3,850 bp for all strains except 18RS21, for which the PCR product was 800 bp longer. Sequencing of approximately 700 bp from either end of the N-1 PCR products revealed near-identity (>95%) among pJMS23, CJB-110, and 18RS21. Additional sequencing of the N-1 amplicon from 18RS21 demonstrated a region with near-identity to an IS1381 variant from strain A909 (22). The sizes of the N-2 amplicons were indistinguishable among all strains studied.

Analysis of the 3′ flanking regions also revealed a high degree of conservation. PCR amplicons were generated with the C-1 and C-2 primer pairs for all strains studied. Sequencing of the 2.8-kb C-1 amplicon from 18RS21 revealed near-identity with the corresponding region of pJMS23 except for the presence of a truncated IS1191 element (23). The C-1 amplicons from JM9–130013 and BM110 both were approximately 2.5 kb, and single sequence reactions from the ends revealed near-identity (>95%) of each with that of pJMS23. The C-1 amplicon from CJB-110 was approximately 6.5 kb; sequencing of approximately 200 bp from the 5′ end of the amplicon showed >95% identity with the corresponding region of pJMS23. The sequence of the C-2 amplicon from CJB-110 essentially is identical (>99%) to that from pJMS23. The C-2 amplicons from JM9–130013, BM110, and 18RS21 are each approximately 650 bp in length and >99% identical to each other and to pJMS23, except for a large deletion.

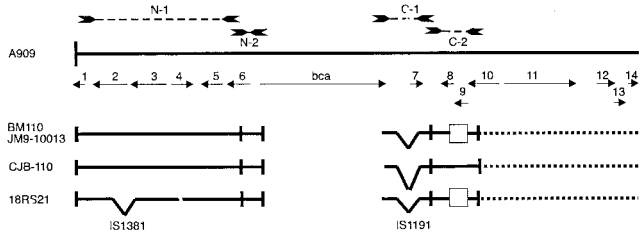

Fig. 3. Homology of the flanking regions of the alpha C and alpha-like protein genes. The 13.7-kb insert of pJMS23, exclusive of the previously sequenced *bca* gene, was sequenced. pJMS23 originally was constructed from strain A909 (7). ORFs are designated by solid arrows. Numbers above the solid arrows correspond to the ORF numbers listed in Table 1. *bca* is the gene encoding the alpha C protein. PCR amplicons using primer pairs (N-1, N-2, C-1, C-2) constructed according to sequence from pJMS23 are designated by dashed lines with arrow ends at the top of the figure. Maps of the regions flanking the alpha-like protein genes in strains BM-110 (protein Rib), CJB-110 (Alp2), JM9–130013 (Alp3), and 18RS21 (protein nonexpresser) were constructed by using PCR and partial sequencing (see text) and are designated by solid lines below the pJMS23 sequence. Inverted hatches in the maps refer to insertions. Open rectangles indicate deletions. Dotted lines within the maps indicate regions that have not been analyzed. IS1381 and IS1191, insertion sequences (see text).

Highly Conserved Regions within Members of the Alpha C Protein Family. Multiple sequence alignments were performed to identify common motifs among the Alp2, Alp3, alpha C, and Rib proteins from GBS, R28 from GAS, and the repeat-containing protein Esp (enterococcal surface protein) from *Enterococcus faecalis*. The sequences of the N-terminal encoding regions of Alp2 and Alp3 were nearly identical to those of the alpha C and Rib proteins through position 333, beyond which variability is apparent (not shown); no significant homology with Esp was seen in the N-terminal encoding region.

Alignment of the repeat regions indicated several motifs that were highly conserved among all of the GBS and GAS repeat elements and the B and C repeats of the enterococcal Esp protein (Fig. 4). Except for the B repeat of Alp2, the 5′ ends of all of the repeat elements were located within a highly conserved 5′ region (motif A); the location of this motif at the junction of different mosaic blocks suggests that it facilitates recombination. The 5′ end of the first B repeat of Alp2 is located in another conserved region (motif B) and, thus, also may represent a recombination site. A third conserved motif (motif C), V-E/V-V-T-Y-P-D-G-T/S-K-D-T-V, that has been described previously in Esp, Rib, and the alpha C protein appears not to be a junctional motif, which suggests that it instead plays a functional role at the amino acid level (24).

Discussion

In this report, we provide evidence that diversity in the alpha-like protein gene family of GBS arises from a combination of recombination events operating at a specific genomic locus. We describe the sequences of Alp2 and Alp3, two new members of a GBS protein family that has been defined previously by the alpha C protein and protein Rib. Alp2 and Alp3 are included in this protein family by virtue of immunologic cross-reactivity and

Table 1. ORFs in the flanking regions of *bca* and their homologies to other bacterial proteins

			Homologous protein					
ORF	Position	Length, aa	Function	Species	Length, aa	Statistical significance, *P*	% Identity/ % similarity	Accession no.
1	1–289	96	Dehydrogenase	*Haloferax volcanii*	389	5e-20	50/60	AAB71801
2	409–1,242	278	Putative reductase protein	*Bacillus subtilis*	276	8e-53	42/63	CAB15345
3	1,327–2,190	288	*czcD*, zinc responsible operon*	*Staphylococcus aureus*	325	3e-22	22/45	BAA36686
4	2,322–2,846	175	Putative transcriptional regulator	*Streptomyces coelicolor*	216	1e-08	24/44	CAB55667
5	3,041–3,559	173	*msmR*, multiple sugar metabolism regulator	*Streptococcus pyogenes*	209	7e-16	31/54	AAC97150
6	3,666–4,235	190	—	—	—	—	—	—
7	8,118–8,444	108	Integrase	Bacteriophage phi-LC3	374	2e-06	30/46	A47085
8	8,879–9,179	100	Predicted coding region	*Helicobacter pylori*	99	2e-04	39/56	AAD07951
9	9,181–9,519	112	Unknown	*Methanobacterium thermoautotrophicum*	99	4e-15	31/68	AE000906
10	9,539–10,303	254	Methyltransferase	*Haemophilus influenzae*	251	2e-69	52/66	057060
11	10,421–12,167	578	Putative transcriptional regulator	*Bacillus stearothermophilus*	697	2e-19	21/42	AAC44464
12	12,666–13,115	149	Phosphotransferase system fructose-specific enzyme IIBC component	*Bacillus halodurans*	160	2e-12	30/49	BAA75339
13	13,119–13,406	95	Phosphotransferase system, galactitol-specific IIB component	*Escherichia coli*	94	.004	25/49	P37188
14	13,425–13,685	334	Phosphotransferase system, galactilol-specific IIC component	*Escherichia coli*	334	2e-11	37/62	BAA15959

*Also homologous to cation transport proteins in multiple species.

characteristic features that are shared with the alpha C and Rib proteins, including a large, internal, tandem repeat region and extremely similar N- and C-terminal encoding regions. In addition, Alp2 and Alp3 exhibit features not seen in the alpha C or Rib proteins, notably a modular structure, suggesting incorporation of genetic regions originating from diverse sources. Sequence analysis of the flanking regions of the alpha C protein gene present on the 13.7-kb insert of pJMS23 revealed several ORFs that are likely to be involved in basic metabolic pathways. Analysis of additional GBS strains that express other alpha-like proteins showed striking conservation of the flanking regions. These findings suggest that the alpha-like proteins are encoded by variant genes at a single location on the GBS genome.

We previously showed that deletions in the alpha C protein repeat region occur *in vivo* under selective antibody pressure and lead to alterations in antigenicity, probably as a result of conformational changes (17). Isolates with reduced numbers of repeats are able to escape recognition by host antibodies generated to the full-length protein (16); a likely mechanism is homologous recombination between identical repeats. As in the alpha C and Rib proteins, variation in the number of nucleotide repeats in the tandem repeat regions accounts for the size heterogeneity of Alp3, suggesting that intragenic recombination contributes to allelic diversity. Although a selective advantage for a decreased number of repeats has been demonstrated, at least in the alpha C protein, the value to the organism of an increased number of repeats is not known but may relate to the ability of the molecule to adhere to host tissues. Tandem repeat regions of other Gram-positive surface proteins are involved in binding to polysaccharides or other proteins (15, 25), including epithelial cell receptors (26). Recent data show that the GAS R28 antigen (nearly identical to Alp3) is involved in binding to epithelial cells (21).

The modular structures of Alp2 and Alp3 suggest that that they are products of mosaic formation, another well described process contributing to antigenic diversity in a number of pathogens, including *Neisseria gonnorheae* (27), GAS (14, 28), *Streptococcus pneumoniae* (29), and *Hemophilus influenzae* (30). Mosaics are thought to result from horizontal gene transfer of partial or whole

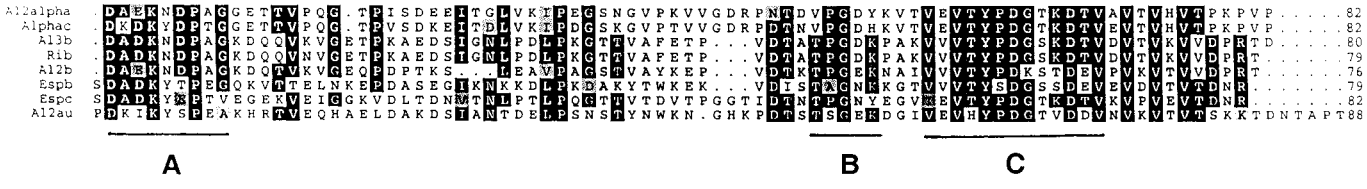

Fig. 4. Multiple sequence alignments of the deduced amino acid sequences of the repeat regions of the alpha-like proteins. Alignments of the repeat regions illustrated conserved motifs near the beginnings and ends of the repeats. Black boxes indicate identity, and gray boxes indicate similarity. Al2alpha, alpha-like repeat element from Alp2; Alphac, alpha C protein repeat element; Al3b, B repeat element from Alp3; Rib, repeat element from protein Rib; Al2b, B repeat element from Al2; Espb, B repeat element from Esp; Espc, C repeat element from Esp; Al2au, A repeat element and beginning of U region from Alp2A. Because the sequence of R28 from GAS essentially is identical to that of Alp3 (98% overall, 96% within the region shown above), it was not included as a separate protein in this figure. Alpha C protein sequence is from ref. 7. Rib sequence is from ref. 10. Esp sequence is from ref. 24.

genes and their subsequent incorporation by recombination. In the present study, the N-terminal half of Alp3 contains modules that are highly similar to those of the alpha C and beta C proteins, whereas the C-terminal half essentially is identical to that of protein Rib. Alp2 contains the same N-terminal modules, a tandem repeat region apparently derived from protein Rib, and an additional repeat apparently derived from the alpha C protein. Because the alpha C and Rib proteins are thought not to coexist on individual GBS strains, we hypothesize that the individual blocks of DNA that make up the mosaic proteins of the newer serotypes were transmitted horizontally before recombination, although we cannot formally exclude the possibility of intragenomic recombination events. Multiple sequence analysis of Alp2, Alp3, alpha C, Rib, and the related Esp reveals a conserved motif at the junction of several repeat sequences that is present even in repeat elements that otherwise do not share significant homology. Substitutions within this motif are largely nonsynonymous, an indication that it is at the nucleotide level that conservation occurs. A putative role of this motif as a recombination hotspot is consistent with these data and supports the recombination model.

We showed previously that some of the repeat-containing GBS proteins are serologically cross-reactive (4, 5). This cross-reactivity probably arises from conserved modules present within different proteins. Immunologic methods of identifying GBS surface proteins with currently available antisera therefore may be misleading and should be used with caution, especially for epidemiologic classification of strains.

Except for the number of repeats, the sequence of Alp3 is nearly identical to the sequence of R28 from *S. pyogenes*, which became available during the preparation of this manuscript (21). This determination at the genetic level of a surface protein common to both group A and group B streptococci supports the hypothesis that the protein plays an important role in virulence of these organisms, perhaps by mediating attachment to mucosal surfaces (21). This gene transfer among streptococcal species is not without precedent; highly similar genes encoding a C5a peptidase have been sequenced from GAS, GBS, and group G streptococci (GGS) (31). *Emm12*, encoding the M12 protein gene, has been demonstrated in divergent GAS and GGS strains, thus suggesting horizontal acquisition (32). In GGS, a mosaic sequence containing elements of the GAS *vir* regulon has been identified, the structure of which suggests *en bloc* transfer of the *vir* regulon from GAS to GGS, followed by rearrangement events (33). These data and the findings we presented here imply that there is a global gene pool for these streptococcal species, analogous to those described for neisserial species (27) and for other streptococci (34). By this model, related species occasionally exchange DNA from a pool of genetic material available to related strains (27). The high degree of cross-species identity in R28 and Alp3 is strong evidence for a recent transfer event, which is striking given the natural incompetence of both of these species. The mechanism for gene transfer between GBS and GAS is unknown; sequence analysis of the flanking regions of the R28 gene might reveal evidence of a particular mechanism, such as a recombination event at a highly homologous chromosomal locus or a transposon insertion. Database searches of the GAS and enterococcal genome sequencing projects have not revealed a related protein in the strains being sequenced.

With few exceptions, GBS strains of a given polysaccharide serotype contain the same alpha-like protein (disregarding tandem repeat number) (2, 4, 5, 9). Thus, the alpha C protein is largely exclusive to type Ia, Ib, and II strains, protein Rib appears to be exclusive to type III strains, and Alp3 is present on most type V and VIII strains, but not on strains of other serotypes. This polysaccharide/protein concordance is in contrast to the distribution of PspA, for example, which is independent of capsular serotype (15). This concordance may reflect clonal lineages within a species that is naturally incompetent and is consistent with several studies showing a predominance of individual clones among clinical GBS isolates (35, 36). Alternatively, it is conceivable that there is a functional linkage between the alpha-like proteins and polysaccharide serotypes that would require this specificity.

In conclusion, our data suggest that multiple factors promote a high degree of local variability at the GBS alpha-like protein locus. We hypothesize that these genes contain hotspots for recombination that have resulted in mosaic variants after horizontal gene acquisition, perhaps from an extensive gene pool. The finite number of variants recognized so far indicates that these events are uncommon, or at least that only certain mosaic variants subsequently are maintained. This selection could occur via fulfillment of functional criteria, e.g., for epithelial cell binding or for a specific capsular polysaccharide interaction. Selection by the host immune system plays a role in determining the number of tandem repeats in the tandem repeat region and also may contribute to the selection of stable mosaic variants, although no data specifically address this. The degree to which random mutations and other mechanisms contribute additionally to sequence diversity remains to be studied.

We are grateful to Dennis Kasper, Fred Ausubel, Susan Hollingshead, Andrew Berry, and Graziella Orefici for insightful discussions. We thank Laura Stulgis for expert technical assistance. This research was supported in part by Public Health Service Grants AI01388, AI38424, and AI33963, by Public Health Service contracts NOI-AI-25152 and NOI-AI-75326, and by a Child Health Research Grant from the Charles H. Hood Foundation.

1. Michel, J. L., Madoff, L. C., Kling, D. E., Kasper, D. L. & Ausubel, F. M. (1991) *Infect. Immun.* **59,** 2023–2028.
2. Stalhammar, C. M., Stenberg, L. & Lindahl, G. (1993) *J. Exp. Med.* **177,** 1593–1603.
3. Flores, A. E. & Ferrieri, P. (1989) *J. Clin. Microbiol.* **27,** 1050–1054.
4. Lachenauer, C. S. & Madoff, L. C. (1996) *Infect. Immun.* **64,** 4255–4260.
5. Lachenauer, C. S., Kasper, D. L., Shimada, J., Ichiman, Y., Ohtsuka, H., Kaku, M., Paoletti, L. C., Ferrieri, P. & Madoff, L. C. (1999) *J. Infect. Dis.* **179,** 1030–1033.
6. Li, J., Kasper, D. L., Ausubel, F. M., Rosner, B. & Michel, J. L. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 13251–13256.
7. Michel, J. L., Madoff, L. C., Olson, K., Kling, D. E., Kasper, D. L. & Ausubel, F. M. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10060–10064.
8. Fischetti, V. A., Pancholi, V. & Schneewind, O. (1990) *Mol. Microbiol.* **4,** 1603–1605.
9. Madoff, L. C., Hori, S., Michel, J. L., Baker, C. J. & Kasper, D. L. (1991) *Infect. Immun.* **59,** 2638–2644.
10. Wastfelt, M., Stalhammar-Carlemalm, M., Delisse, A., Cabezon, T. & Lindahl, G. (1996) *J. Biol. Chem.* **271,** 18892–18897.
11. Ferrieri, P. & Flores, A. E. (1997) in *Streptococci and the Host*, ed. Horaud, T. (Plenum, New York), Vol. 418, pp. 635–637.
12. Whatmore, A. M., Kapur, V., Musser, J. M. & Kehoe, M. A. (1995) *Mol. Microbiology* **15,** 1039–1048.
13. Hollingshead, S. K., Fischetti, V. A. & Scott, J. R. (1987) *Mol. Gen. Genet.* **207,** 196–203.
14. Bessen, D. E. & Hollingshead, S. K. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 3280–3284.
15. Briles, D. E., Tart, R. C., Swialto, E., Dillard, J. P., Smith, P., Benton, K. A., Ralpha, B. A., Brooks-Walter, A., Crain, M. J., Hollingshead, S. K., *et al.* (1998) *Clin. Microb. Rev.* **11,** 645–657.
16. Madoff, L., Michel, J., Gong, E., Kling, D. & Kasper, D. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 4131–4136.
17. Gravekamp, C., Rosner, B. & Madoff, L. (1998) *Infect. Immun.* **66,** 4347–4354.
18. Chaffin, D. O. & Rubens, C. E. (1998) *Gene* **219,** 91–99.
19. Jerlstrom, P. G., Chhatwal, G. S. & Timmis, K. N. (1991) *Mol. Microbiol.* **5,** 843–849.
20. Heden, L.-O., Frithz, E. & Lindahl, G. (1991) *Eur. J. Immun.* **21,** 1481–1490.
21. Stalhammar-Carlemalm, M., Areschoug, T., Larsson, C. & Lindahl, G. (1999) *Mol. Microbiol.* **33,** 208–219.
22. Tamura, G. S., Herndon, M., Przekwas, J., Rubens, C. E., Ferrieri, P. & Hillier, S. L. (1999) *J. Infect. Dis.* **181,** 364–368.
23. Guedon, G., Bourgoin, F., Pebay, M., Roussel, Y., Colmin, C., Simonet, J. & Descaris, B. (1995) *Mol. Microbiol.* **16,** 69–78.
24. Shankar, V., Baghdayan, A. S., Huycke, M. M., Lindahl, G. & Gilmore, M. S. (1999) *Infect. Immun.* **67,** 193–200.
25. Wren, B. W. (1991) *Mol. Microbiol.* **5,** 797–803.
26. Perez-Casal, J., Okada, N., Caparon, M. G. & Scott, J. R. (1995) *Mol. Microbiol.* **15,** 907–916.
27. Maiden, M. C. J., Malorny, B. & Achtman, M. (1996) *Mol. Microbiol.* **21,** 1297–1298.
28. Whatmore, A. & Kehoe, M. (1994) *Mol. Microbiol.* **11,** 363–374.
29. Brooks-Walter, A., Briles, D. & Hollingshead, S. K. (1999) *Infect. Immun.* **67,** 6533–6542.
30. Gilsdorf, J. R. (1998) *Infect. Immun.* **66,** 5053–5059.
31. Chmouryguina, I., Suvorov, A., Ferrieri, P. & Cleary, P. P. (1996) *Infect. Immun.* **64,** 2387–2390.
32. Simpson, W. J., Musser, J. M. & Cleary, P. P. (1992) *Infect. Immun.* **60,** 1890–1893.
33. Sriprakash, K. S. & Hartas, J. (1996) *Microb. Pathog.* **20,** 275–285.
34. Reichmann, P., Konig, A., Linares, J., Alcaide, F., Tenover, F. C., McDougal, L., Swidsindki, S. & Hakenbeck, R. (1997) *J. Infect. Dis.* **176,** 1001–1012.
35. Musser, J. M., Mattingly, S. J., Quentin, R., Goudeau, A. & Selander, R. K. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 4731–4735.
36. Blumberg, H. M., Stephens, D. S., Modansky, M., Erwin, M., Elliot, J., Facklam, R. R., Schuchat, A., Baughman, W. & Farley, M. M. (1996) *J. Infect. Dis.* **173,** 365–373.