CHAPTER 4

GBS PROTEIN GENE PROFILES

Fanrong Kong[1], Sonia Gowan[2], Diana Martin[2], Gregory James[1], Gwendolyn L. Gilbert [1]

1. Centre for Infectious Diseases and Microbiology, Insititute of Clinical Pathology and Medical Research, Westmead Hospital, Westmead, NSW 2145, Australia
2. Environmental Science and Research, Porirua, Wellington, New Zealand

This chapter has been published in *Journal of Clinical Microbiology* 2002; 40:620-626. (see Appendices section)

Statement of Joint Authorship

Kong, F. (candidate)

Did all the molecular work, interpreted the data and wrote the manuscript.

Gowan, S. and Martin, D.

Provided and serotyped the New Zealand GBS strains.

James, G.

Provided all the needed molecular experiment equipment, reagents and software.

Gilbert G. L. (supervisor)

Supervised the overall project. Assisted in research design, in the analysis and interpretation of data, and made a significant contribution to the manuscript.

4.1. SUMMARY

The study of surface protein antigens of group B streptococcus (GBS) is important for an understanding of the pathogenesis and epidemiology of infection, and several have been proposed as components of GBS conjugate vaccines. In a previous study, we developed a novel PCR and sequencing system for identification of GBS serotypes and serosubtypes based on the capsular polysaccharide synthesis (*cps*) gene cluster. In this study, we used published sequences to develop PCR assays for identification of genes encoding GBS surface proteins including C alpha (*bca*), C alpha-like 2 and 3 (*alp2* and *alp3*), Rib (*rib*) and C beta (*bac*). We showed that the prototype R reference strain, Prague 25/60, contained a novel alpha-like protein antigen 4 gene (*alp4*). Initial evaluation of these gene-specific assays showed excellent specificity. By combining *cps* serotypes, serosubtypes and surface protein gene profiles, we were able to divide 224 GBS isolates into 31 serovariants. GBS *bac*-positive strains could be further subtyped into 11 groups and 20 subgroups. Our results confirmed and extended reported associations between some *cps* serotypes and serosubtypes with surface protein genes namely: serosubtypes III-1 and III-2 were associated with *rib*; serosubtype III-3 with *alp2*; serotype Ib with *bca* and *bac*, serotype V with *alp3*. The associations between serotype Ia and *bca*, *bca* repetitive unit or *bca* repetitive unit-like sequence-containing genes need to be studied further. These PCR-based methods will provide an alternative and objective tool for subtyping of GBS based on surface protein antigen genes.

4.2. INTRODUCTION

Group B streptococcus (GBS; *Streptococcus agalactiae*) is the commonest cause of neonatal and obstetric sepsis and an increasingly important cause of septicaemia in the elderly and immunocompromised patients (Schuchat, 1999). There are nine capsular polysaccharide GBS serotypes (based on the capsular polysaccharide

synthesis [*cps*] gene cluster), which vary in their distribution between geographic areas, disease types and patient age-groups (Harrison *et al*., 1998; Hickman *et al*., 1999). Capsular polysaccharides are important virulence factors and epidemiological markers and the main components of conjugate vaccines. For studies of epidemiology and pathogenesis, it is important to identify as many phenotypic or molecular markers as possible to increase the discriminatory power of typing systems (Hauge *et al*., 1996). In addition to capsular polysaccharide antigens, GBS surface protein antigens, which also contribute to the pathogenesis of GBS disease and induce protective immunity, are potentially useful markers (Lachenauer *et al*., 2000). Their use in polysaccharide conjugate vaccines is under investigation (Gravekamp *et al*., 1997; Larsson *et al*., 1996). Identification of surface protein antigens, combined with *cps* serotyping, allows subdivision of GBS strains into a large number of serovariants, which can facilitate epidemiological, pathogenesis, and other related studies of GBS infection (Kvam *et al*., 1995).

The genes encoding the proteins C alpha (*bca*), C alpha-like 2 and 3 (*alp2* and *alp3*), and Rib (*rib*) have been well studied and their gene sequences published in GenBank (Lachenauer *et al*., 2000; Michel *et al*., 1992; Wastfelt *et al*., 1996). They are members of a family of surface proteins containing repetitive elements, which produce variations in protein size and antigenicity (Lachenauer *et al*., 2000; Wastfelt *et al*., 1996). The gene encoding the C beta or IgA binding protein (*bac*) also has been well described (Heden *et al*., 1991; Jerlstrom *et al*., 1991). C alpha, C beta, and Rib proteins all have been proposed as potential vaccine components (Larsson *et al*., 1996; Madoff *et al*., 1996).

Numerous methods have been used to identify GBS surface antigens – using monoclonal (Moyo *et al*., 2001) or polyclonal antibodies (Bevanger *et al*., 1995) or genes, by hybridization with probes (Suvorov *et al*., 1997), PCR (Brakstad *et al*., 1997; Maeland *et al*., 1999, 2000; Mawn *et al*., 1993) and/or sequencing (Brakstad *et al*., 1997; Lachenauer *et al*., 2000; Maeland *et al*., 1999, 2000). PCR-based methods are attractive because of their high discriminatory power and

reproducibility (Brakstad *et al*., 1997; Maeland *et al*., 1999, 2000). PCR methods to detect C alpha and C beta protein genes have been published (Brakstad *et al*., 1997; Maeland *et al*., 1999, 2000; Mawn *et al*., 1993), but the specificity, clinical application and interpretation of these methods require further study. Specific PCR methods to identify genes encoding Rib and C alpha-like proteins 2 and 3, which are present in the more virulent serotypes, III and V, have not yet been described (Lachenauer *et al*., 2000; Wastfelt *et al*., 1996). Associations between *cps* serotypes and some protein antigens have been described (Kvam *et al*., 1995; Suvorov *et al*., 1997). They are likely to vary over time and in different populations and geographic locations (Harrison *et al*., 1998; Hickman *et al*., 1999) and should be useful for studies of the epidemiology and pathogenesis of GBS infection.

In this study, we used published sequences of surface protein antigen genes, including *bca*, *alp2*, *alp3*, *rib*, and *bac* (Heden *et al*., 1991; Jerlstrom *et al*., 1991; Lachenauer *et al*., 2000; Michel *et al*., 1992; Wastfelt *et al*., 1996), to improve and/or develop protein gene-specific PCR assays. We used these assays to examine the distribution and variation of surface protein genes and their associations with *cps* genes, in a large collection of GBS isolates collected over the past decade in Australia and New Zealand.

4.3. MATERIALS AND METHODS

4.3.1. GBS isolates, serotyping and serosubtyping.

Isolates used in this study and the serotype and serosubtype identification methods have been described in chapter 3 (Kong *et al*., 2002a). Isolates included well-characterised reference panels kindly provided by Dr Lawrence Paoletti, Channing Laboratory, Boston USA (serotypes Ia to VIII; reference panel 1) and Dr Diana Martin, Streptococcus Reference Laboratory, at ESR, Wellington, New Zealand (serotypes Ia to VI; reference panel 2) and 206 clinical isolates. All isolates were

serotyped by the conventional and molecular methods and some were serosubtyped by PCR and sequencing. Antisera used for serotyping were prepared against serotypes Ia, Ib, Ic, II-VIII and the R protein antigen. The prototype R reference strain Prague 25/60 was used to raise R antiserum.

4.3.2. Oligonucleotide primers.

Oligonucleotide primers used in this study, their target sites in the gene sequences and their melting temperatures are shown in Table 4.1. The primers were synthesised according to our specifications by Sigma-Aldrich (Castle Hill, NSW, Australia). Six previously published oligonucleotide primers (Brakstad *et al*., 1997; Maeland *et al*., 1999, 2000; Mawn *et al*., 1993), and a series of new primers designed by us were used to sequence parts of and/or to specifically amplify genes encoding GBS surface proteins. All new primers, except two used only for sequencing *rib* and six previously published (unmodified), were designed with high melting temperatures (>70 ^{o}C) for use in rapid cycle PCR (Table 4.1.).

4.3.3. DNA preparations and PCR.

DNA was prepared from GBS cultures (Mawn *et al*., 1993) and PCR performed as previously described (Kong *et al*., 2002a) and the same as in chapter 3. The denaturation, annealing and elongation temperatures and times used were 96^{o}C for 1 second, 45-72^{o}C (according to the primer *T*m values or as previously described) for 1 second and 74^{o}C for 1 to 30 seconds (according to the length of amplicons), respectively, for 30-35 cycles, using a Perkin Elmer Thermal Cycler 9600.

10 µl PCR products were analysed by electrophoresis on 1.5% agarose gels, which were stained with 0.5 µg ethidium bromide ml^{-1}. For detection and/or subtyping, the presence of expected length PCR amplicons, shown by ultraviolet transillumination, were accepted as positive. For sequencing, 40 µl volumes of PCR products were

Table 4.1. Primers used in GBS protein gene profiling study.

Primer[reference]	**Target genes**	**Tm °C[a]**	**GenBank numbers**	**Sequence[b, c]**
IgAagGBS[e]	*bac*	73.8	X59771	**2663**GCGATTAAACAA CAA ACT ATT TTT GAT A TTG ACA ATG CAA**2702**
IgAS1[d]	*bac*	72.8	X59771	**2765**GCT AAA TTT CAA AAA GGT CTA GAG ACA AAT ACG CCA G**2801**
IgAA1[d]	*bac*	78.9	X59771	**3157**CCC ATC TGG TAA CTT CGG TGC ATC TGG AAG C**3127**
RIgAagGBS[e]	*bac*	76.3	X59771	**3284**CAGCCAACTCTTTC GTC GTT ACT TCC TTG AGA TGT AAC**3247**
GBS1360S[f]	*bac*	72.3	X59771	**1325**GTGAAATTGTAT AAG GCT ATG AGT GAG AGC TTG GAG**1360**
GBS1717S[d]	*bac*	75.0	X59771	**1685**ACA GTC ACA GCT AAA AGT GAT TCG AAG ACG ACG**1717**
GBS1937A[f]	*bac*	75.9	X59771	**1976**CCGTTTTAGAATCTTT CTG CTC TGG TGT TTT AGG AAC TTG**1937**
bcaRUS[g]	*bca* repetitive unit	73.5	M97256	**769**GATAAATATGATCCAA CAG GAG GGG AAA CAA CAG TAC**805**
bcaRUA[g]	*bca* repetitive unit	77.2	M97256	**1003**CTGGTTTTGGTGTCACAT GAA CCG TTA CTT CTA CTG TAT CC**963**
bcaS1[d]	*bca*/*alp2*/*alp3*	71.7	M97256/AF291065	**208/533**GGT AAT CTT AAT ATT TTT GAA GAG TCA ATA GTT GCT GCA TCT AC**251/576**
bcaS2[d]	*bca*/*alp2*/*alp3*	78.0	M97256/AF291065	**256/581**CCAGGGA GTG CAG CGA CCT TAA ATA CAA GCA TC**288/613**
balS[d]	*alp2/alp3*	73.8	AF291065	**677**GAT CCT CAA AAC CTC ATT GTA TTA AAT CCA TCA AGC TAT TC**717**

bcaA[d]	*bca*	74.2	M97256	**597**CGTTCTAACTT CTT CAA TCT TAT CCC TCA AGG TTG TTG**560**
balA[d]	*alp2/alp3*	73.6	AF291065	**978**CCA GTT AAG ACT TCA TCA CGA CTC CCA TCA C**948**
bal23S1[d]	*alp2/alp3*	70.9	AF208158/AF291065	**1093/1373**CAG ACT GTT AAA GTG GAT GAA GAT ATT ACC TTT ACG G**1129**/**1409**
bal23S2[d]	*alp2/alp3*	72.9	AF208158/AF291065	**1174/1454**CTT AAA GCT AAG TAT GAA AAT GAT ATC ATT GGA GCT CGT G**1213**/**1493**
bal2A1[d]	*alp2*	78.3	AF208158	**1426**CGT GTT GTT CAA CAG TCC TAT GCT TAG CCT CTG GTG**1391**
bal2A2[d]	*alp2*	70.8	AF208158	**1518**GGT ATC TGG TTT ATG ACC ATT TTT CCA GTT ATA CG**1484**
bal3A[d]	*alp3*	79.2	AF291065	**1693**GAC CGT TTG GTC CTT ACC TTT TGG TTC GTT GCT ATC C**1657**
#ribS1[d]	*rib*	65.2	U58333	**216**TAC AGA TAC TGT GTT TGC AGC TGA AG**241**
ribS2[d]	*rib*	73.0	U58333	**238**GAAGTAATTTCAG GAA GTG CTG TTA CGT TAA ACA CAA ATA TG**279**
ribA1[d]	*rib*	78.8	U58333	**431**GAA GGT TGT GTG AAA TAA TTG CCG CCT TGC CTA ATG**396**
ribA2[d]	*rib*	72.6	U58333	**462**AAT ACT AGC TGC ACC AAC AGT AGT CAA TTC AGA AGG**427**
#ribA3[d]	*rib*	61.3	U58333	**570**CAT CTA TTT TAT CTC TCA AAG CTG AAG**554**

Notes.

\# For sequencing use only, not entirely specific for ***rib***.

a. The primer *T*m values were provided by the primer synthesiser (Sigma-Aldrich, Castle Hill, NSW, Australia).

b. Numbers represent the numbered base positions at which primer sequences start and finish (numbering start point "1" refer to the start point "1" of corresponding GenBank accession number, of which there are two for some sequences).

c. Underlined sequences show bases added to modify previously published primers.

d. Primers designed by us for this study.

e. From Mawn ***et al.***, 1993.

f. From Brakstad ***et al.***, 1997.

g. From Maeland ***et al.***, 2000.

further purified by polyethylene glycol precipitation method (Ahmet *et al*., 1999).

4.3.4. Sequencing.

To confirm the specificity of newly designed or modified primer pairs, we sequenced 10, 13 and 10 selected amplicons produced by bcaS1-bcaA (targeting 5'-end of *bca*), ribS1-ribA3 (targeting *rib*) and GBS1360S-GBS1937A (targeting *bac*), respectively, from the two panels of reference strains and 31 randomly selected clinical isolates. All amplicons of primer pairs bcaS1-balA (targeting *alp2* and *alp3*), bal23S1-bal2A2 (targeting *alp2*) and IgAagGBS-RIgAagGBS (targeting *bac*), from any of the 224 isolates were sequenced.

The PCR products were sequenced using Applied Biosystems (ABI) *Taq* DyeDeoxy terminator cycle-sequencing kits according to standard protocols. The corresponding amplification primers or inner primers were used as the sequencing primers.

4.3.5. Database similarity searching and sequence comparison.

Databases were searched for sequence similarity using *FastA* program in SeqSearch program group. Sequences were compared using *Bestfit* and *Gap* programs in Comparison program group. Translate program in Translation program group was used to translate from DNA sequence to amino acid sequence. All programs are provided in WebANGIS, ANGIS (Australian National Genomic Information Service).

4.3.6. Surface protein gene profile codes.

Each isolate was given a protein gene profile code according to positive PCR results using various primer pairs, as shown in Table 4.2.

Table 4.2. Specificity and expected lengths of amplicons of GBS protein gene profiling primer pairs.

Primer pairs*	Specificity	Length of amplicons (base pairs)	Protein gene profile codes
IgAagGBS-RIgAagGBS	*bac*	532-838	B
IgAS1-IgAA1	*bac*	303-591	B
GBS1360S-GBS1937A	*bac*	652	B
GBS1717S-GBS1937A	*bac*	292	B
bcaS1-bcaA	5'-end of *bca*	390	A
bcaS2-bcaA	5'-end of *bca*	342	A
BcaRUS-BcaRUA	*bca* repetitive unit or *bca* repetitive unit-like region	235	a or as
bcaS1-balA	*alp2* or *alp3*	446	alp2 or alp3
bcaS2-balA	*alp2* or *alp3*	398	alp2 or alp3
BalS-balA	*alp2* or *alp3*	302	alp2 or alp3
bal23S1-bal2A1	*alp2*	334	alp2
bal23S2-bal2A1	*alp2*	253	alp2
bal23S1-bal2A2	*alp2*	426	alp2
bal23S2-bal2A2	*alp2*	345	alp2
bal23S1-bal3A	*alp3*	321	alp3
bal23S2-bal3A	*alp3*	240	alp3

#ribS1-ribA3	*rib*	355	R or alp4
ribS2-ribA1	*rib*	194	R
ribS2-ribA2	*rib*	225	R
ribS2-ribA3	*rib*	333	R

Notes.

* See Table 4.1. for primer sequences. For primer pair ribS1-ribA3 the annealing temperature used was 55-60^{o}C. For three previously described primer pairs (unmodified) annealing temperatures were as described (Maeland *et al*., 1997, 2000; Mawn *et al*., 1993); and for all other new primer pairs, they were 68-72^{o}C (according to the *T*m values).

\# For sequencing use only, not entirely specific for rib gene (see text for more detail).

4.3.7. Nucleotide sequence accession numbers.

The sequences generated during this study were depositied in GenBank with the following accession numbers: AF367974 (partial *bac* sequence, with an insertion sequence IS*1381* from one isolate); AF362685-AF362704 (partial *bac* sequences for all *bac*-positive isolates); AF373214 (partial *alp4* for reference strain Prague 25/60, an R protein standard strain).

Previously published gene sequences used in this study and their GenBank accession numbers are as follows: M97256 (*bca*); X58470, X59771 (*bac*); U58333 (*rib*); AF208158 (*alp2*), AF291065-AF291072 (*alp3*); AJ488912 (partial *alp4*); AF064785 (IS*1381*). AY179867 (R4 protein gene); AE014211 (containing Rib protein gene [SAG0433] in the 2603V/R genome).

4.4. RESULTS

4.4.1. PCR results.

With few exceptions, all primer pairs produced amplicons of predicted length from isolates giving positive results (Table 4.2.). The exceptions included one isolate that was positive by PCR using primer pairs GBS1360S-GBS1937A and GBS1717S-GBS1937A (both targeting *bac*) but produced amplicons significantly longer than those of other *bac*-positive isolates. Sequencing showed that the amplicon contained the insertion sequence IS*1381* with minor variations compared with the published sequences (Tamura *et al.*, 2001). The amplicons produced using primers IgAagGBS-RIgAagGBS and IgAS1-IgAA1 (also targeting *bac*) varied in length (Berner *et al.*, 2001) and were sequenced for further subtyping (see below and Table 4.3.).

4.4.2. Evaluation of the protein gene-specific primer pairs by direct sequencing of PCR amplicons.

All 10 amplicons of primer pair bcaS1-bcaA and 12 of 13 (except strain Prague 25/60, see below) of primer pair ribS1-ribA3 were identical with the corresponding portion of the gene sequences in GenBank (M97256, *bca* and U58333, *rib*, respectively). Four of 10 amplicons of primer pair GBS1360S-GBS1937A (targeting *bac*) were identical with the corresponding gene sequences in GenBank (X58470, X59771). A single point mutation (A to G, 1441 of X59771) was found in the remaining six *bac*-positive amplicons, including the one, which contained the insertion sequence IS*1381* (see above and AF367974).

Fifty isolates produced amplicons using primer pair bcaS1-balA. The sequences of nine were identical with the corresponding portions of the published sequence of *alp2* (AF208158) and 41 with that of *alp3* (AF291065). There are two consistent heterogeneity sites between *alp2* and *alp3* in the sequences of bcaS1-balA amplicons, which can be used to distinguish them, in addition to *alp2* and *alp3*-specific PCR. All nine amplicons of primer pair bal23S1-bal2A2 were identical with the corresponding portion of the *alp2* sequence in GenBank (AF208158). The primer pair IgAagGBS-RIgAagGBS identified *bac* in 52 isolates. There was considerable sequence variation, which allowed separation of *bac*-positive isolates into 11 groups and 20 subgroups based on amplicon length and sequence heterogeneity, respectively (Table 4.3.). The groups contained small numbers (one to five) of isolates except for B1 (20 isolates, two subgroups) and B4 (11 isolates, three subgroups). In general, the presence or absence of short repetitive sequences was responsible for differences in amplicon length (Berner *et al.*, 2001; Jerlstrom *et al.*, 1991).

4.4.3. Further confirmation of specificity of surface protein gene-specific primer pairs.

Table 4.3. Genetic groups and subgroups of *bac* (C beta protein gene) based on amplicon length (using primers IgAagGBS-RIgAagGBS) and sequence heterogeneity.

Group or Subgroup	N=	Amplicon length	GenBank numbers	No. of different sites compared with (c.f.) main group	Molecular sero/subtypes[a]
B1	19	532	AF362685[b]		17 = Ib; 2 = II
B1a	1	532	AF362686	1 (c.f. B1)	Ib
B2	3	550	AF362687		Ib, II, III-4
B3	2	586	AF362688		2=Ib
B3a	1	586	AF362689	4 (c.f. B3)	V
B3b	1	586	AF362690	21 (c.f. B3)	VI
B3c	1	586	AF362691	24 (c.f. B3)	Ib
B4	8	604	AF362692		4 = Ib; 4 = II
B4a	1	604	AF362693	1 (c.f. B4)	II
B4b	2	604	AF362694	2 (c.f. B4)	2 = Ib
B5	2	622	AF362695[c]		Ia, VI
B5a	1	622	AF362696	2 (c.f. B5)	Ia
B6	1	640	AF362697		Ib
B7	1	658	AF362698		Ib
B7a	1	658	AF362699	34 (c.f. B7)	VI
B8	1	712	AF362700		Ib
B9	2	748	AF362701		2 = II
B9a	1	748	AF362702	13 (c.f. B9)	Ib
B10	2	820	AF362703		2 = Ib
B11	1	838	AF362704		Ib

Notes.

a. See Table 4.4. for further details of *cps* sero/subtype relationships with protein antigen genes.

b. AF362685 sequence is identical with the corresponding portion of X58470.

c. AF362695 sequence is identical with the corresponding portion of X59771.

To confirm primer pair specificity, we compared the results of PCR using the primer pairs we had designed or modified for *bac* PCR, with those of PCR using previously published primer pairs (Brakstad *et al*., 1997; Maeland *et al*., 1999, 2000; Mawn *et al*., 1993) and found 100% correlation.

The previously reported non-specificity of the published primer pair bcaRUS-bcaRUA (targeting the *bca* repetitive unit) was confirmed (Brakstad *et al*., 1997; Maeland *et al*., 1999, 2000). Using these primers, all nine *alp2* positive (bcaS1/bcaA negative) isolates and 53 isolates (including the reference strain Prague 25/60**,** also see below), which were PCR negative using the primer pairs bcaS1-bcaA, bcaS2-bcaA (targeting the 5'-end of *bca*), bal23S1-bal2A2 and bal23S2-bal2A1 (targeting the 5'-end of *alp2*) produced amplicons. Our sequencing showed that *bca*, *alp2* and *alp4* (see below**)** have significant homology in the regions targeted by bcaRUS-bcaRUA allowing amplicon formation from *alp2*- and *alp4*-positive strains (Lachenauer *et al*., 2000; Brakstad *et al*., 1997; Maeland *et al*., 1999, 2000). These false positive results could be due to the presence of other C alpha-like protein genes, containing regions homologous with the *bca* repetitive unit (*bca* repetitive unit-like sequence).

We also showed that the results of PCR using two or more primer pairs that we had designed for individual genes (*rib*, *alp2*, and *alp3*) correlated well, supporting the specificity of each set. The only exception, as mentioned above, was ribS1-ribA3, which produced a non-specific amplicon from one of 224 isolates tested.

4.4.4. Prague 25/60 contains another new alpha-like surface protein antigen gene – *alp4*.

The strain Prague 25/60 (which is used to raise R antiserum), in reference panel 2, produced an amplicon with primer pair ribS1-ribA3 but not with ribS2-ribA1, ribS2-ribA2, and ribS2-ribA3. It was therefore assumed not to contain *rib*, although the amplicon sequence showed considerable homology with *rib* and other members of

the family of surface proteins (see below). This isolate was the only one, of 224 tested, for which PCRs were negative using ribS2-ribA1 and ribS2-ribA2 but positive using ribS1-ribA3. The latter primer pair is then not entirely specific for *rib* and was therefore used only for sequencing.

Sequencing of the Prague 25/60 ribS1-ribA3 amplicon showed considerable homology with other members of surface protein gene family defined by *bca-rib*, and it is finally proved to be identical with a GenBank *alp4* sequence later published (AJ488912, *alpha*-like protein 4 gene [*alp4*] for reference strain NCTC9828, a NT-R3R4 standard strain). The *alp4* also contained the *bca* repetitive unit-like sequence, and probably with multiple copies, which led to the amplicon of "a". The *alp4* (AJ488912) similarity ratios compared with DNA sequences of *bca*, *rib*, *alp2* and *alp3* were 70.5%, 71.9%, 72.1%, and 78.6%, respectively. Since this amplicon sequence is most similar to that of *bca,* which encodes C alpha, the prototype of the surface protein family, the gene was named *alp4* (C alpha-like protein antigen 4 gene). The region of *bca* repetitive unit-like sequence for *alp2* and *alp4* share very high homology (94.5%), suggested their possible common source (Lachenauer *et al*., 2000).

4.4.5. Surface protein gene profiles.

For each GBS surface protein gene (except *bca* repetitive unit and *bca* repetitive unit-like region), we selected two primer pairs to identify and characterise them by PCR. Four common profiles accounted for 203 of 224 (90.6%) isolates: "R" (62 isolates), "AaB" (51 isolates), "a" (49 isolates) and "alp3" (41 isolates) (see Table 4.4.). Only two isolates contained no surface protein gene markers. All but one isolate with the *bac* ("B") also had *bca* with its repetitive unit ("Aa"); one had *rib* ("R"). All "alp2" isolates contained single *bca* repetitive unit-like sequences ("as"); a "alp4" isolate contained multiple *bca* repetitive unit-like sequences ("a"). "A", "R", "alp2", "alp3" and "alp4" were all mutually exclusive. 62 of 63 isolates with *rib* ("R") and 41 of 41 isolates with *alp3* had no other protein antigen gene markers.

4.4.6. Relationship between surface protein antigen gene profiles and *cps* serotypes and serosubtypes.

Development of the molecular serotype (MS) identification method and comparison with conventional serotyping (CS) have been described elsewhere (Kong *et al*., 2002a). A *cps* MS was assigned to all isolates and the results correlated with CS results except for 19 of 224 isolates that were nontypable using antisera. The relationship between surface protein gene profiles and *cps* molecular serotypes are summarised in Table 4.4.

The following strong associations were confirmed or demonstrated between: MS Ia with *bca* repetitive unit or *bca* repetitive unit-like sequence (most with profile "a"), MS III-1 and III-2 with *rib*, MS III-3 with *alp2*, MS Ib with *bca* and *bac* and MS V with *alp3*. MS II showed the most varied surface protein gene profiles. However, the relationships were not absolute and different combinations of polysaccharide *cps* serotypes and protein gene profiles produced 31 serovariants or 51 when *bac* (B) subgroups were considered.

4.4.7. Relationship between surface protein antigens and protein gene profiles.

Based on conventional serotyping, 33 isolates (belonging to CS Ia/c, Ib/c, Iic, Iib, IIIc or IIIb) reacted with the C antiserum. The surface protein gene profiles of all of these isolates contained *bca* (A) and/or *bca* repetitive unit-related markers (a or as) as follows: Aa (three isolates), AaB (18 isolates), a (11 isolates), alp2as (one isolate). Twenty-nine isolates reacted with the R antiserum and, of these, 22 contained *rib* and six contained *alp3*. The remaining isolate was Prague 25/60 (the reference strain Used to raise the R protein antiserum), which contained the new alpha-like protein 4 gene – *alp4* (see above).

Table 4.4. The relationship between GBS protein gene profiles[a] and capsular polysaccharide synthesis (*cps*) gene molecular sero/subtypes.

Sero/subtype[b]	N=	None	Aa	AaB	R	alp3	a	as	alp2as	RB	alp4[c]a
Ia	43	-	-	2	-	-	35	3	3	-	-
Ib	37	-	1	35	-	1	-	-	-	-	-
II	29	-	3	10	8	2	5	-	-	-	1[3]
III-1	30	-	-	-	30	-	-	-	-	-	-
III-2	22	-	-	-	22	-	-	-	-	-	-
III-3	5	-	-	-	-	-	-	-	5	-	-
III-4	3	-	-	1	-	1	-	-	1	-	-
IV	9	-	-	-	1	-	8	-	-	-	-
V	38	1	-	-	1	35	-	-	-	1	-
VI	5	-	1	3	-	-	1	-	-	-	-
VII	1	-	-	-	-	1	-	-	-	-	-
VIII	2	1	-	-	-	1	-	-	-	-	-
Total	224	2	5	51	62	41	49	3	9	1	1

Notes.

a. Protein antigen gene profile codes are:
   "A": 5'end of ***bca*** positive; "a" or "as": ***bca*** repetitive unit or ***bca*** repetitive unit-like region positive, with multiple or single band amplicons, respectively; "B": ***bac*** positive; "R": ***rib*** positive; "alp2": ***alp2*** positive; "alp3": ***alp3*** positive; "None": isolate contains none of the above protein genes.
b. See text for explanation of ***cps*** serosubtypes (Kong ***et al***., 2002a).
c. This isolate was the atypical reference strain Prague 25/60, which was nonserotypable using capsular polysaccharide antisera but was assigned to molecular serotype II. It expresses a protein that is antigenically similar to R (and is used to generate R antiserum), encoded by the gene "***alp4"***.

4.5. DISCUSSION

In our previous study, all the isolates used in the present study were serotyped by conventional and molecular methods, that identified their *cps* serotypes and, in some cases, serosubtypes (Kong *et al*., 2002a). In this study, we developed PCR-based methods to identify GBS surface protein genes and further characterise these isolates. Using the published *bac* sequence, we modified *bac*-specific primers (Brakstad *et al*., 1997; Maeland *et al*., 1999, 2000; Mawn *et al*., 1993) and designed new primers, with high melting temperatures (>70 ^{o}C) suitable for rapid cycle PCR (Ke *et al*., 2000; Kong *et al*., 2002a) targeting all major surface protein genes.

As previously reported, a published PCR primer pair targeting the *bca* repetitive unit (at the 3'-end of *bca*), was not entirely specific for *bca* (Brakstad *et al*., 1997; Maeland *et al*., 1999, 2000). We designed two new primer pairs targeting the 5'-end of *bca*, to improve the specificity. However, very few serotype Ia strains gave positive results using these two primer pairs whereas all were PCR positive using primer pair targeting the *bca* repetitive unit (Brakstad *et al*., 1997; Maeland *et al*., 1999, 2000). These results were consistent with a previous report (Hauge *et al*., 1996) that a probe targeting the 5'-end of *bca* hybridized with only one of nine serotype Ia strains whereas a large *bca* probe, including the tandem repeat region, hybridized with all nine. Further study is required to define the sequences and specificities of different portions of *bca* and their effects, if any, on the structure and functions of C alpha and related proteins.

PCR specific for *rib*, *alp2* and *alp3* have not been described previously. The primer pairs we designed mainly targeted the 5'-ends of the genes and were chosen after comparing their heterogeneity with related gene sequences (Lachenauer *et al*., 2000; Wastfelt *et al*., 1996). We designed two or more primer pairs for each gene to check primer specificity by comparison of results of different PCR targeting the same genes. Protein gene profiles "alp2" and "alp3" were distinguished on the basis of the *alp2* and *alp3*-specific PCR and/or two sequence heterogeneity sites in the

amplicons of bcaS1/balA, or bcaS2/ balA.

To confirm the specificity of our primers, we used them to examine two reference panels and selected GBS isolates. The longest amplicons produced by PCR for each gene were sequenced, to provide maximal sequence information and ensure that the inner primers were not located at strain heterogeneity sites. Our sequencing results confirmed the specificity of the primers. Two pairs of primers for each gene were compared, with similar results. Finally, six gene/region specific primer pairs (including the one targeting the *bca* repetitive unit) were used to define protein antigen gene profiles for all 224 isolates.

The study showed that only one member of the surface protein gene family containing repetitive sequences – *rib*, *bca*, *alp2*, *alp3* and *alp4* – was present in any single isolate (Kvam *et al.*, 1995; Lachenauer *et al.*, 2000). However, all isolates containing *bac*, which is not member of the surface protein gene family containing repetitive sequences, also contained either *bca* (51/52) or *rib* (1/52) (Kvam *et al.*, 1995).

The C beta protein gene, *bac*, was present in 23% of isolates, a similar proportion to those (19-22%) previously reported (Berner *et al.*, 1999). In common with others, we found variations in the *bac* (Berner *et al.*, 2002) amplicons due to variable small internal repetitive sequences (Berner *et al.*, 2001; Jerlstrom *et al.*, 1991) that, unlike those of the *bca-rib* family, were irregular. Their role is not clear, but they are potentially useful molecular markers for epidemiological studies (Berner *et al.*, 2001; Heden *et al.*, 1991).

Our study confirmed previously reported relationships between *cps* serotypes and surface protein gene profiles (Lachenauer *et al.*, 2000). For example: some serotype III isolates (our MS III-1 and III-2) were closely associated with *rib* (Stalhammar-Carlemalm *et al.*, 1993), and others (our MS III-3) with *alp2* (Lachenauer *et al.*, 2000). Serotype Ib was associated with *bca* and *bac* (Kvam *et al.*, 1995) and

serotype V with *alp3* (Lachenauer *et al*., 2000). However, as the relationship was not absolute, different combinations of *cps* serotypes/protein gene profiles identified many serovariants, which will be useful in epidemiological studies and in formulation of conjugate vaccines (Kvam *et al*., 1995; Lachenauer *et al*., 2000). Based on PCR only, we were able to divide our 224 isolates into 31 serovariants based on *bac* ("B") groups or 51, based on subgroups. Theoretically, there are likely to be additional serovariants.

Comparison of protein antigen (C and R proteins) serotyping results with the protein gene profiles, showed that the presence of the gene does not necessarily indicate the expression of the corresponding protein. This is one reason for discrepancies between genetic and serotyping results; another is that C and R protein antisera are not entirely specific (Lachenauer *et al*., 2000). Our analysis showed that reaction with C antiserum generally correlated with the presence of genes encoding C alpha (*bca*) or alpha-like protein 2 (*alp2*). Reaction with R antiserum correlated with presence of genes encoding R protein (*rib*), the alpha-like protein 3 (*alp3*) or the new, rare alpha-like protein 4 (*alp4*) (found in the study). Apparently antigenic cross-reactivity does not necessarily reflect genetic similarity, since the *alp4* sequence studied was more similar to that of *bca* than *rib* or *alp3*. More extensive analysis of these genes and the relationships between the proteins they encode is required. These methods will be useful in further studies of the effects of various antigen profiles on virulence and to further define the genealogy of GBS serotypes and various subtypes.

New study showed that R4 protein gene (GenBank number: AY179867) is nearly identical to Rib protein gene (GenBank number: U58333), suggest R4 is actually Rib (Bevanger *et al*., 1995; Fasola *et al*., 1996).