CHAPTER 5

GBS MOBILE GENETIC ELEMENTS

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

Statement of Joint Authorship

Kong, F. (candidate)

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

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.

5.1. SUMMARY

Our aim is to develop an integrated genotyping system for group B streptococcus (GBS, *Streptococcus agalactiae*) that can be used to study the population genetics of the organism and the pathogenesis and epidemiology of GBS disease. In our previous studies, we used two sets of markers – the capsular polysaccharide synthesis (*cps*) gene cluster and surface protein antigen genes – to assign molecular serotypes (MS) and protein gene profiles (pgp) to more than 200 isolates. In the present study, we have used five mobile genetic elements (mge) as a third set of markers, to further characterise 194 invasive isolates, recovered from blood or cerebrospinal fluid (CSF). Of these, 97% contained one or more mge, the distribution of which was related to MS and pgp, as illustrated by MS III, which is divisible into four serosubtypes (sst) with different combinations of mge (or none). We identified 56 different genotypes and eight genetic clusters, each with different combinations of the three sets of molecular markers. Five predominant genotypes (Ia-1, Ib-1, III-1, III-2 and V-1) contained 62% of isolates and five of the eight genetic clusters contained 92% of isolates. The 17 cerebrospinal fluid (CSF) isolates were relatively widely distributed between 10 genotypes and across seven of the eight clusters. Further study is needed to determine whether these genotypes or clusters share common markers of increased virulence. In future, comparison of invasive with colonising strains of GBS may elucidate the significance of these findings.

5.2. INTRODUCTION

Group B streptococcus (GBS, *Streptococcus agalactiae*) is the commonest cause of neonatal sepsis and an increasingly important cause of septicaemia in elderly and immunocompromised patients (Schuchat, 1998). However, it is also frequently carried in the normal faecal and/or vaginal flora. Integrated studies of strain

virulence, and the epidemiology and pathogenesis of infection, are needed to identify strains that are most likely to cause disease (van Belkum *et al*., 2001). Our aim is to develop a GBS genotyping system with enough discriminatory ability to be used for such studies.

Nine GBS serotypes are defined by differences in capsular polysaccharide antigens (Chaffin *et al*., 2000), which are important virulence factors, encoded by the *cps* gene clusters. A family of variable surface protein antigens, including Rib, C alpha, C alpha-like and the IgA binding protein, C beta, also contribute to virulence (Lachenauer *et al*., 2000; Mawn *et al*., 1993; Wastfelt *et al*., 1996). Numerous mobile genetic elements (mge), including the four insertion sequences (IS) and a group II intron selected for use in the study, have been identified in GBS (Franken *et al*., 2001; Granlund *et al*., 1998, 2001; Rubens *et al*., 1989; Spellerberg *et al*., 2000; Tamura *et al*., 2000). Their presence and the number of copies can affect gene expression and virulence (Mahillon & Chandler, 1998; Mahillon *et al*., 1999; Martinez-Abarca & Toro, 2000) and they are potentially useful epidemiological markers (Tamura *et al*., 2000).

In our previous studies, we developed methods, based on PCR and sequencing, to identify serotypes and to define surface protein gene profiles, using a panel of 224 GBS isolates and compared the results with those of conventional serotyping (Kong *et al*., 2002a, b). Molecular and conventional serotyping results correlated well and, in addition, a molecular serotype was assigned to isolates that were nontypable using antisera (Kong *et al*., 2002a, b). We confirmed previously reported associations between capsular polysaccharide and surface protein antigens (Lachenauer *et al*., 2000; Wastfelt *et al*., 1996) and divided our 224 GBS isolates into more than 50 molecular types corresponding with different antigen combinations (and antigenic variants) (Kong *et al*., 2002b).

In this study, we used five selected mge as a third set of molecular markers, to further increase the discriminatory ability of our genotyping system (Kong *et al*.,

2002a, b). We then used this typing system to examine the population genetic structure and age-related disease distribution of genotypes among 194 invasive GBS isolates.

5.3. MATERIALS AND METHODS

5.3.1. Invasive GBS clinical isolates.

All 194 isolates used in this study were recovered from the blood (177) or CSF of 191 patients (107 female, 80 male, four sex unrecorded; three cultures each contained mixed growth of two GBS serotypes). 108 isolates were from specimens submitted for culture to the Centre for Infectious Diseases and Microbiology (CIDM), ICPMR, Sydney, Australia during 1996-2001 and 83 were referred to Institute of Environmental Science and Research (ESR), Porirua, Wellington, New Zealand for serotyping, from various diagnostic laboratories in New Zealand, during 1994-2000.

Isolates and methods used to identify molecular serotypes, serosubtypes and surface protein gene profiles have been previously described in detail (Kong *et al*., 2002a, b). Patients were classified into age-groups for analysis of genotype distribution as follows: neonatal, early onset (0-6 days); neonatal, late onset (7 days to 3 months); infant and child (4 months-14 years); young adult (15-45 years); middle-aged (46-60 years); elderly (>60 years).

5.3.2. Oligonucleotide primers, DNA preparations and PCR.

The target sites, sequences, melting temperatures and numbered base positions of the oligonucleotide primers used in this study are shown in Table 5.1. Expected amplicon lengths of different primer pairs can be calculated from the 5'-end positions of the corresponding primers. Primers were designed, specifically for this

Table 5.1. Primers used in GBS mobile genetic element study.

Primer	Target	Tm ^{o}C^a	GenBank numbers	Sequence[b]
IS861S	IS*861*	77.4	M22449	**445**GAG AAA ACA AGA GGG AGA CCG AGT AAA ATG GGA CG**479**
IS861A1	IS*861*	77.3	M22449	**831**CAC GAT TTC GCA GTT CTA AAT AAA TCC GAC GAT AGC C**795**
IS861A2	IS*861*	76.1	M22449	**1020**CAA ACT CCG TCA CAT CGG TAT AGC ACT TCT CAT AGG**985**
IS1548S	IS*1548*	76.5	Y14270	**143**CTA TTG ATG ATT GCG CAG TTG AAT TGG ATA GTC GTC**178**
IS1548S1	IS*1548*	77.0	Y14270	**539**GTT TGG GAC AGG TAG CGG TTG AGG AGA AAA GTA ATG**574**
IS1548A1	IS*1548*	77.0	Y14270	**574**CAT TAC TTT TCT CCT CAA CCG CTA CCT GTC CCA AAC**539**
IS1548A2	IS*1548*	70.3	Y14270	**915**CCC AAT ACC ACG TAA CTT ATG CCA TTT G**888**
IS1548A3	IS*1548*	78.0	Y14270	**930**CGT GTT ACG AGT CAT CCC AAT ACC ACG TAA CTT ATG CC**893**
IS1381S1	IS*1381*	80.1	AF064785/ AF367974	**272/818**CTT ATG AAC AAA TTG CGG CTG ATT TTG GCA TTC ACG**307/853**
IS1381S2	IS*1381*	81.7	AF064785/ AF367974	**497/1040**GGC TCA GGC GAT TGT CAC AAG CCA AGG GAG**526/1069**
IS1381A	IS*1381*	73.1	AF064785/ AF367974	**881/1424**CTA AAA TCC TAG TTC ACG GTT GAT CAT TCC AGC**849/1392**
ISSa4S	ISS*a4*	78.5	AF165983	**326**CGT ATC TGT CAC TTA TTT CCC TGC GGG TGT CTC C**359**
ISSa4A1	ISS*a4*	75.2	AF165983	**639**GCC GAT GTC ACA ACA TAG TTC AGG ATA TAG CCA G**606**
ISSa4A2	ISS*a4*	74.5	AF165983	**780**CGT AAA GGA GTC CAA AGA TGA TAG CCT TTT GA ACC**745**
GBSi1S1	GBSi1	78.6	AJ292930	**721**CAT CTC GGA ACA ATA TGC TCG AAG CTT ACA AGC AAG TG**758**

GBSi1S2	GBSi1	77.3	AJ292930	**789**GGG GTC ACT ATC GAG CAG ATG GAT GAC TAT CTT CAC**824**
GBS i1A1	GBSi1	83.9	AJ292930	**1058**AAT GGC TGT TTC GCA GGA GCG ATT GGG TCT GAA CC**1024**
GBS i1A2	GBSi1	80.5	AJ292930	**1161**CCA GGG ACA TCA ATC TGT CTT GCG GAA CAG TAT CG**1127**

Notes.

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" refers to the start point "1" of corresponding gene GenBank accession number).

study, using published sequences of the relevant mge. DNA extraction from GBS cultures, and PCR were performed as previously described (Kong *et al.*, 1999, 2000a, b, c, 2001, 2002a).

5.3.3. Sequencing.

To evaluate the specificity of our primer pairs, we sequenced selected amplicons produced by primer pairs IS1548S-IS1548A3 (21 isolates), ISSa4S-ISSa4A2 (9 isolates), IS861S-IS861A2 (10 isolates), IS1381S1-IS1381A (2 isolates) and GBSi1S1-GBSi1A2 (3 isolates). For sequencing, 40 µl volumes of PCR products were further purified by polyethylene glycol precipitation method (Ahmet *et al.*, 1999). 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.

5.3.4. Multiple sequence alignments and sequence comparison.

Multiple sequence alignments were performed with *Pileup* and *Pretty* programs in Multiple Sequence Analysis program group. Sequences were compared using *Bestfit* program in Comparison program group. All programs are provided in WebANGIS, ANGIS (Australian National Genomic Information Service), 3rd version.

5.3.5. Nucleotide sequence accession numbers.

The insertion sequence IS*1381,* sequenced in our previous study (Kong *et al.*, 2002b), has appeared in GenBank with accession number AF367974 (IS*1381* in partial C beta antigen gene). The five mge previously sequenced by others have appeared in GenBank with the following accession numbers: AF064785 (IS*1381*); M22449 (IS*861*); Y14270 (IS*1548*); AF165983 (ISSa*4*); and AJ292930 (GBSi1). Two recently released genome sequences (Glaser *et al.*, 2002; Tettelin *et al.*, 2002)

have appeared in GenBank with the following accession numbers: NC_004116 (*S. agalactiae* 2603V/R, complete genome), NC_004368 (*S. agalactiae* NEM316, complete genome).

5.3.6. Definitions of molecular serotype (MS), serosubtype (sst) and protein gene profiles (pgp).

Our previous study (Kong *et al*., 2002a) defined molecular serotypes (MS) Ia, Ib, II-VIII (as for conventional serotypes) and four serosubtypes (sst) of MS III, based on heterogeneity within several regions of the *cps* gene cluster. A second study (Kong *et al*., 2002b) defined pgp based on the presence of genes encoding Rib, C alpha, C alpha-like, C beta proteins (Kong *et al*., 2002b).

5.3.7. Statistical analysis and dendrogram.

SSPS version 11 software was used for statistic analysis. A dendrogram was formed using Average Linkage (between groups) and Hierarchical Cluster Analysis in SSPS version 11 software. The presence or absence of each marker - MS Ia, Ib, II, IV-VI, sst III-1-4; pgp "A", "R", "a", "as", "alp2", alp3"; *bac* subgroups 1, 1a, 2, 3, 3a, 3b, 3c, 4, 4b, 5a, 7, 7a, 8, 9, 9a, 10, N1, N2; and mge IS*1381*, IS*861*, IS*1548*, ISSa*4*, GBSil - were included in the analysis. The genotypes were each characterized by a distinct combination of the MS or sst, pgp and mge.

5.4. RESULTS

5.4.1. Specificity of primer pairs.

The sequencing results showed that 10 IS*861*, 21 IS*1548*, nine ISSa*4* and three GBSi1 amplicon sequences were either identical with their corresponding sequences in GenBank (M22449, Y14270, AF165983, AJ292930 and NC_004116) or had very

minor mutations (one to three sites, respectively).

In our previous study, we found a full-length IS*1381* (AF367974) within the C beta antigen gene (*bac)* of one clinical isolate. Our sequence showed several differences compared with the original published sequence (AF064785): the terminal inverted repeats contained 17 rather than 20 base pairs (bp); between positions 419 and 429, there was a three bp deletion and four individual bp differences as follows: **GGG ATC CGA TT** (AF064785) vs **CAG A-- -GG TA** (AF367974; our sequence). The sequence of the amplicons of primer pair IS1381S1-IS1381A, from two selected clinical isolates, were identical to that of our sequence (AF367974) and to the IS*1381* sequences in the GBS (2603V/R) genome (NC_004116), but different, as above, from the original reported IS*1381* sequence (AF064785).

In addition to sequencing, we evaluated the specificity of our primer pairs by comparing PCR results for two or more primer pairs for each target (Table 5.1.). In all cases, the same sets of isolates produced amplicons when tested with different primers targeting the same mge, thus confirming the specificity of the primer pairs.

5.4.2. Distribution of mge.

The numbers of isolates containing different mge combinations (from none to four per isolate) are shown in Table 5.2. IS*1381,* IS*861,* IS*1548,* IS*Sa4* and GBSi1 were identified in 87%, 52%, 17%, 6% and 18% of isolates, respectively. Six (3%) isolates contained none of the five selected mge.

5.4.3. Predominant relationships between MS/sst, pgp and mge.

Figure 5.1. shows the relationships between the various genetic markers.
IS*1381* was present in nearly all isolates of MS Ia, Ib, IV, V and VI, but in none ofsst III-2 or III-3. IS*1548* and GBSil were found most commonly in serotypes II orIII; three isolates (all MS II) contained both GBSi1 and IS*1548*. IS*861* was found

inall sst III-1 and III-2 and most MS II and Ib isolates but only in 14% of other MS isolates. ISSa*4* was present in only 6% of isolates, more than half of which were MS II; it was present in one invasive isolate obtained before 1996 (1994). IS*1381* was found in most isolates except those in cluster 8, pgp "alp2", which had none of the four insertion sequences (Glaser *et al*., 2002). IS*861* was found in most genotypes with pgp "AaB" (clusters 3 and 4) and all genotypes with pgp "R" (clusters 6 and 7).

5.4.4. Genotypes based on MS/sst, pgp, *bac* subtypes and mge.

MS/sst, pgp, *bac* subtype (for isolates with pgp "B") and the presence of various combinations of mge provide a PCR/sequencing-based genotyping system. The 194 invasive isolates in this study represented seven serotypes, ten MS/sst, 41 subtypes based on the distributions of pgp and mge or 56 genotypes when *bac* subtypes (mainly in MS Ib) were included (Figure 5.1.).

5.4.5. Theoretical GBS clonal population structure.

Theoretically there are 13 possible GBS MS/sst (eight MS - Ia, Ib, II, IV-VIII, four sst III 1-4 and *cps* gene cluster absent) and at least ten pgp (none, "Aa", "AaB", "a", "as", "R", "RB", "alp2as", "alp3" or "alp4a"). If the 22 *bac* subgroups identified so far are included, there are up to 31 pgp. If the five mge were independently, randomly distributed and present or absent, there would be $13 \times 31 \times 2^5 = 12,896$ different possible combinations of molecular markers. The fact that only 56 different combinations were found (Figure 5.1.), demonstrates that markers are not randomly distributed or, in other words, these invasive Australasian GBS isolates have a clonal population structure. It is possible, but unlikely, that these isolates represent a very limited number of GBS genotypes.

Table 5.2. Distribution of mobile genetic elements among 194 invasive GBS isolates.

	Mobile genetic elements present					
Total N =	IS*1381*	IS*861*	IS*1548*	IS*Sa4*	GBSi1	None
6	-	-	-	-	-	6
78	78	-	-	-	-	-
2	-	-	-	-	2	-
37	37	37	-	-	-	-
1	1	-	1	-	-	-
3	3	-	-	3	-	-
29	29	29	29	-	-	-
6	6	6	-	6	-	-
8	8	8	-	-	8	-
18	-	18	-	-	18	-
1	1	-	-	-	1	-
1	1	-	1	-	1	-
2	2	2	2	-	2	-
2	2	-	-	2	2	-
Total (n=194)	**168 (87%)**	**100 (52%)**	**33 (17%)**	**11 (6%)**	**34 (18%)**	**6 (3%)**

Note.

Data are numbers of isolates containing various combinations of mge.

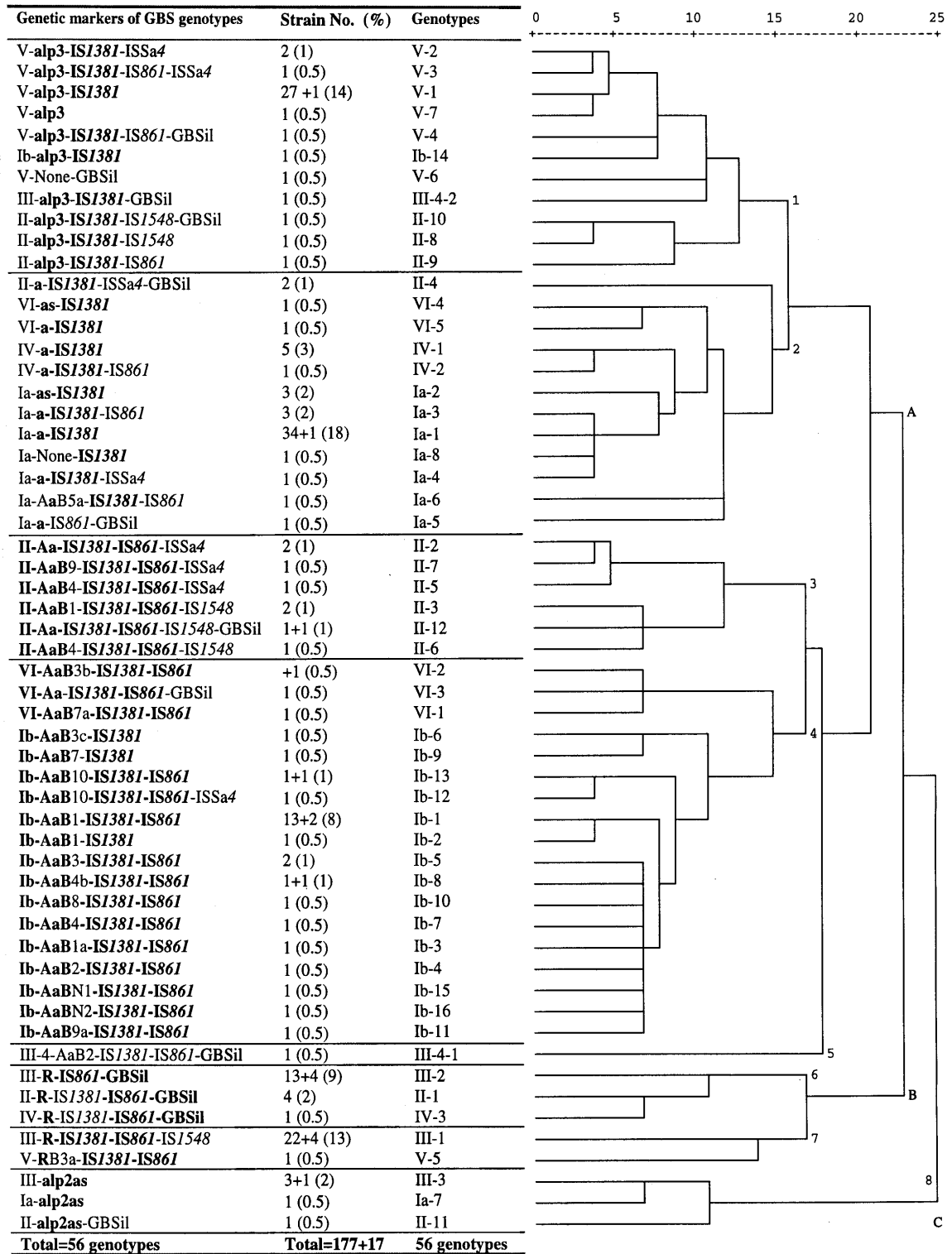| Genetic markers of GBS genotypes	Strain No. (%)	Genotypes
V-alp3-IS*1381*-ISSa*4*	2 (1)	V-2
V-alp3-IS*1381*-IS*861*-ISSa*4*	1 (0.5)	V-3
V-alp3-IS*1381*	27 +1 (14)	V-1
V-alp3	1 (0.5)	V-7
V-alp3-IS*1381*-IS*861*-GBSil	1 (0.5)	V-4
Ib-alp3-IS*1381*	1 (0.5)	Ib-14
V-None-GBSil	1 (0.5)	V-6
III-alp3-IS*1381*-GBSil	1 (0.5)	III-4-2
II-alp3-IS*1381*-IS*1548*-GBSil	1 (0.5)	II-10
II-alp3-IS*1381*-IS*1548*	1 (0.5)	II-8
II-alp3-IS*1381*-IS*861*	1 (0.5)	II-9
II-a-IS*1381*-ISSa*4*-GBSil	2 (1)	II-4
VI-as-IS*1381*	1 (0.5)	VI-4
VI-a-IS*1381*	1 (0.5)	VI-5
IV-a-IS*1381*	5 (3)	IV-1
IV-a-IS*1381*-IS*861*	1 (0.5)	IV-2
Ia-as-IS*1381*	3 (2)	Ia-2
Ia-a-IS*1381*-IS*861*	3 (2)	Ia-3
Ia-a-IS*1381*	34+1 (18)	Ia-1
Ia-None-IS*1381*	1 (0.5)	Ia-8
Ia-a-IS*1381*-ISSa*4*	1 (0.5)	Ia-4
Ia-AaB5a-IS*1381*-IS*861*	1 (0.5)	Ia-6
Ia-a-IS*861*-GBSil	1 (0.5)	Ia-5
II-Aa-IS*1381*-IS*861*-ISSa*4*	2 (1)	II-2
II-AaB9-IS*1381*-IS*861*-ISSa*4*	1 (0.5)	II-7
II-AaB4-IS*1381*-IS*861*-ISSa*4*	1 (0.5)	II-5
II-AaB1-IS*1381*-IS*861*-IS*1548*	2 (1)	II-3
II-Aa-IS*1381*-IS*861*-IS*1548*-GBSil	1+1 (1)	II-12
II-AaB4-IS*1381*-IS*861*-IS*1548*	1 (0.5)	II-6
VI-AaB3b-IS*1381*-IS*861*	+1 (0.5)	VI-2
VI-Aa-IS*1381*-IS*861*-GBSil	1 (0.5)	VI-3
VI-AaB7a-IS*1381*-IS*861*	1 (0.5)	VI-1
Ib-AaB3c-IS*1381*	1 (0.5)	Ib-6
Ib-AaB7-IS*1381*	1 (0.5)	Ib-9
Ib-AaB10-IS*1381*-IS*861*	1+1 (1)	Ib-13
Ib-AaB10-IS*1381*-IS*861*-ISSa*4*	1 (0.5)	Ib-12
Ib-AaB1-IS*1381*-IS*861*	13+2 (8)	Ib-1
Ib-AaB1-IS*1381*	1 (0.5)	Ib-2
Ib-AaB3-IS*1381*-IS*861*	2 (1)	Ib-5
Ib-AaB4b-IS*1381*-IS*861*	1+1 (1)	Ib-8
Ib-AaB8-IS*1381*-IS*861*	1 (0.5)	Ib-10
Ib-AaB4-IS*1381*-IS*861*	1 (0.5)	Ib-7
Ib-AaB1a-IS*1381*-IS*861*	1 (0.5)	Ib-3
Ib-AaB2-IS*1381*-IS*861*	1 (0.5)	Ib-4
Ib-AaBN1-IS*1381*-IS*861*	1 (0.5)	Ib-15
Ib-AaBN2-IS*1381*-IS*861*	1 (0.5)	Ib-16
Ib-AaB9a-IS*1381*-IS*861*	1 (0.5)	Ib-11
III-4-AaB2-IS*1381*-IS*861*-GBSil	1 (0.5)	III-4-1
III-R-IS*861*-GBSil	13+4 (9)	III-2
II-R-IS*1381*-IS*861*-GBSil	4 (2)	II-1
IV-R-IS*1381*-IS*861*-GBSil	1 (0.5)	IV-3
III-R-IS*1381*-IS*861*-IS*1548*	22+4 (13)	III-1
V-RB3a-IS*1381*-IS*861*	1 (0.5)	V-5
III-alp2as	3+1 (2)	III-3
Ia-alp2as	1 (0.5)	Ia-7
II-alp2as-GBSil	1 (0.5)	II-11
Total=56 genotypes	**Total=177+17**	**56 genotypes**

Figure 5.1. Genetic relationship of 194 invasive Australasian GBS strains (or 56 genotypes).

Notes for column headed "Genetic markers of GBS genotypes":

See text for explanation of *cps* serosubtypes (sst).

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.

See text for explanation of five mobile genetic elements.

The molecular markers in bold type show the common features in each cluster.

Notes for column headed "Strains No. (%)":

After "+" are the numbers of CSF isolates, the others are blood isolates.

Notes for column headed "Genotypes":

Each genotype was characterized by a distinct combination of the *cps* genes, protein gene profiles and mobile genetic elements. The predominant genotype in each serotype were named as the number "1" genotype of that serotype.

Notes for the dendrogram:

At about distance 16, the 56 genotypes could be separated into eight clusters (1-8); at about distance 22.5, the 56 genotypes could be separated into three cluster groups (A, B, C).

5.4.6. The phylogenetic relationship of Australasian invasive GBS.

The 56 genotypes formed eight clusters, if they were separated at a genetic distance of about ~16 (or three cluster groups separated at a distance of ~22.5). The pgp was the main determinant of cluster separation (Figure 5.1.). 94% of isolates belonged to five MS (Ia, Ib, II, III and V), 62% belonged to five (9%) genotypes (Ia-1, Ib-1, III-1, III-2 and V-1) and 92% belonged to the five largest clusters (1, 2, 4, 6 and 7). Cluster group A, the largest, contained 139 (72%) isolates and 48 (86%) genotypes, 45 of which contained fewer than five isolates, whereas cluster group B contained 49 (25%) isolates and five (9%) genotypes. The main characteristics of each cluster were as follows:

Cluster 1. "alp3", IS*1381* (39 isolates, four MS, 11 genotypes; predominant genotype V-1).

Cluster 2: "a" or "as", IS*1381* (55 isolates, four MS, 12 genotypes, predominant genotype Ia-1).

Cluster 3: "Aa" or "AaB", MS II, IS*1381,* IS *861* (ten isolates, six genotypes).

Cluster 4: "AaB", IS*1381,* IS*861* (35 isolates, two MS: VI or Ib; 18 genotypes; predominant genotype Ib-1).

Cluster 5. "AaB", IS*861,* GBSi1, genotype III-4-1 (one isolate).

Cluster 6: "R", IS*861* and GBSil (22 isolates, three MS/genotypes; predominant genotype III-2).

Cluster 7: "R", IS*1381* and IS*861* (27 isolates; two MS/genotypes; predominant genotype III-1).

Cluster 8: "alp2as", none of the four selected IS (six isolates; three MS/genotypes; one contained GBSi1).

Table 5.3. Relationship between genotypes of invasive GBS isolates and patients' age-groups.

Genotype				Age-group			
	0-6 d	7 d-3 m	4 m-14 yr[b]	15-45 yr	46-60 yr	>60 yr	Total
Ia-1	14	4+1[a]	1	7	3	6	35+1
Ia-(2-8)	4	2	-	1	-	3	10
Ia total	18	6+1	1	8	3	9	45+1
Ib-1	2	1+1	-	3	2	5+1	13+2
Ib-(2-16)	3	4+2	-	3	1	5	16+2
Ib total	5	5+3	-	6	3	10+1	29+4
II	8	1	-	4+1	1	4	18+1
III-1	6+1	4	1+1	1+1	6+1	4	22+4
III-2	5	5+4[c]	1	2	-	-	13+4
III-(3-4)	1+1	1	-	1	1	1	5+1
III total	12+2	10+4	2+1	4+1	7+1	5	40+9
IV total	3	-	-	-	-	4	7
V-1	3	3	2	4	2	13+1	27+1
V-(2-7)	1	1	-	1	-	4	7
V total	4	4	2	5	2	17+1[d]	34+1
VI total	1	-	-	-	+1	3	4+1
Total	**51+2=53**	**26+8=34**	**5+1=6**	**27+2=29**	**16+2=18**	**52+2=54**	**177+17=194**

Notes:

a. Numbers after "+" refer to CSF isolates; all others are from blood.

b. Five aged 4 m-1 yr and one case was aged 3 yr.

c. Sst III-2 in late onset infection (9/34=26%) compared with all other patient age groups (8/160=5%): p=0.0005, odds ratio (OR) 6.8; 95% confidence interval (CI) 2.4-19.4.

d. MS-V in elderly (18/54=33%) compared with all other age-groups 17/140 (12%): p=0.001, OR 3.6; 95% CI 1.7-7.7.

5.4.7. The relationship between genotypes and GBS disease patterns.

The distribution of MS and genotypes in different age groups of patients with invasive GBS disease is shown in Table 5.3. All common MS were represented in more than one patient group. However, there were highly significant associations (when compared with all other age-groups) between sst III-2 and late onset neonatal infection (p=0.0005) and MS V and infection in the elderly (p=0.001).

There were 17 isolates from cerebrospinal fluid specimens, nine (53%) of which were MS III (from three different sst/genotypes, each in a different cluster). The other eight isolates were distributed among five MS, seven genotypes and four clusters. Meningitis occurred in all age-groups but comprised 23% of cases in the late onset neonatal group compared with 5% in all other groups.

5.5. DISCUSSION

Our aim is to develop a comprehensive genotyping system to study the epidemiology and pathogenesis of GBS infection. It should be reproducible, objective and transportable between laboratories, comparable with and complementary to other typing methods and able to incorporate known virulence markers (Hauge *et al*., 1996). Based on these criteria, we first developed a molecular serotyping (MS) method based on the *cps* gene cluster. It compared favourably with, but was more sensitive than, conventional serotyping (CS) (Kong *et al*., 2002a) and allowed us to identify several subtypes of serotype (sst) III, as described by others (Bohnsack *et al*., 2001). A second subtyping method based on the family of genes encoding variable surface protein antigens (*bca/rib/alp2/alp3/alp4*) and the IgA binding protein C beta (*bac*), is more sensitive and objective than conventional protein serotyping, which cannot type all isolates and is sometimes misleading (Lachenauer *et al*., 2000). Our methods also can identify more members of the

family of variable antigen genes and distinguish numerous *bac* subgroups (Kong *et al*., 2002b).

To extend our typing system we have now used five mobile genetic elements (mge) (Mahillon & Chandler, 1998; Mahillon *et al*., 1999; Martinez-Abarca & Toro, 2000), including four different insertion sequences (IS) and a group II intron, which have been identified in GBS (Granlund *et al*., 1998, 2001; Rubens *et al*., 1989; Spellerberg *et al*., 2000; Tamura *et al*., 2000). Two additional insertion sequences (IS*Sag1* and IS*Sag2*) found in GBS, were reported to be present in all human GBS isolates (Franken *et al*., 2001) and are therefore unlikely to increase the discriminatory ability of our typing system.

IS*861* was the first IS to be identified in GBS. The presence of multiple copies in some serotype III isolates was associated with increased *cps* gene expression (and hence greater virulence) (Rubens *et al*., 1989). Multiple copies of IS*1381* are found in most GBS and some other *Streptococcus* spp., including *S. pneumoniae* (Sanchez-Beato *et al*., 1997). They have been used as probes for restriction fragment length polymorphism (RFLP) analysis of GBS for epidemiological studies (Tamura *et al*., 2000). ISSa*4* was first identified in a nonhemolytic GBS isolate, in which it caused insertional inactivation of the gene *cylB,* which is part of an ABC transporter involved in production of hemolysin. Only a small proportion of (mainly hemolytic) GBS isolates (4%) contained ISSa*4,* all of which had been isolated since 1996 and it was postulated that ISSa*4* had been newly acquired by GBS (Spellerberg *et al*., 2000). IS*1548* was first discovered in some hyaluronidase-negative GBS serotype III isolates, in which it caused insertional inactivation of the gene *hyl*B (one of a gene cluster responsible for production of hyaluronidase, an important GBS virulence factor) (Granlund *et al*., 1998). Another copy of IS*1548* is also found downstream of the C5a peptidase gene (also associated with virulence), in isolates that contain it.

Group II introns - large self-splicing mge that transpose via an RNA intermediate - are found in many bacteria, including some clinically significant species, such as *S. pneumoniae* (Martinez-Abarca & Toro, 2000). Recently a previous unknown group II intron, GBSi1, was identified in more than one third of type III GBS isolates (Granlund *et al*., 2001), none of which contained IS*1548,* suggesting that these two mge were mutually exclusive markers of different genetic lineages. The location of some mge in proximity to genetic markers associated with virulence suggests that they may be involved in pathogenicity island-like structures in GBS, which may also contribute to its clonal population structure (Bohnsack *et al*., 2002; Franken *et al*., 2001; Hauge *et al*., 1996; Quentin *et al*., 1995).

In this study, we developed PCR methods to detect the presence of each of these five mge in GBS isolates, but did not determine the number of copies or attempt to correlate their presence with virulence. We have significantly increased the level of discrimination of our GBS genotyping system for epidemiological studies. It could be increased further by combination with other methods, such as multilocus enzyme electrophoresis (MLEE) (Musser *et al*., 1989), multilocus sequence typing (MLST) (Enright & Spratt, 1999), random amplified polymorphic DNA (RAPD) (Chatellier *et al*., 1997; Limansky *et al*., 1995; Martinez-Abarca & Toro, 2000) or restriction length fragment polymorphisms (RFLP) (Tamura *et al*., 2000). Combining typing methods with identification of virulence markers and clinical data, would allow better definition of potentially virulent genotypes.

We used only invasive GBS isolates to demonstrate the practical value of our genotyping system, confirm their clonal population structure and determine the distribution of genotypes in different patient groups. The isolates originated from patients of all ages with GBS sepsis. About half were consecutive GBS isolates from blood or CSF, at a large diagnostic laboratory in a general adult hospital, with an obstetric unit (i.e there were no isolates from children other than neonates). The rest were consecutive isolates referred for serotyping from all over New Zealand. Thus the overall age distribution is representative of that in the population affected

by GBS disease, except that children beyond the early neonatal period are probably under-represented. However, the distribution of genotypes within each age-group should be representative.

Among our 194 Australasian invasive GBS isolates we identified 56 genotypes, of which five (Ia-1, Ib-1, III-1, III-2 and V-1) accounted for 62% of isolates. Whether this indicates that these genotypes are more virulent or just more prevalent among GBS isolates in general is not clear. Previous studies, mostly confined to isolates from neonates and pregnant carriers, have suggested that there is less diversity among invasive than colonising strains (Musser *et al*., 1989; Quentin *et al*., 1995). Another study confirmed associations between genotypes and virulence factors (hyaluronidase, IgA binding protein, C alpha and variants of C beta proteins) but found no differences between invasive neonatal and colonising genital isolates (Hauge *et al*., 1996). We are collecting vaginal isolates from unselected pregnant women, to compare the distribution of genotypes with that in our invasive isolates from neonates and pregnant women. Considering the genotypic heterogeneity of invasive isolates, even among those from neonatal infections, it is likely that a large number of vaginal isolates will be required to demonstrate significant differences reliably. It will be more difficult to collect large numbers of appropriate noninvasive strains (faecal isolates) for comparison with invasive isolate from immunocompromised and elderly patients.

The phylogenetic tree derived from our results showed both similarities to, and differences from, that derived by Hauge *et al* (1996). Between them, the two studies examined most known GBS genotypic and phenotypic markers (more than 30 together). Both showed relationships between *cps* serotype and pgp and demonstrated a similar level of discrimination - 56 (this study) and 58 (Hauge *et al*., 1996) genotypes among 194 and 85 isolates, respectively. Both showed that certain known virulence markers - C beta, C alpha variants and hyaluronidase production (indirectly in our study - see below) - were associated with distinct clonal lineages. The main difference between the two studies was in the source of isolates. Besides

the isolates came from two different geographic areas, ours included a high proportion from adults with invasive disease, in which serotype V was common, whereas the other study examined invasive isolates from infants and genital isolates, with a greater proportion of serotype III.

Our 17 CSF isolates were genetically heterogeneous, although more than half of them belonged to MS III. They were distributed among all age-groups but the highest proportion of cases of meningitis was among cases of late onset neonatal infection. Others have found that CSF isolates from neonates are much more likely than vaginal isolates to be serotype III (77% vs 27%) and were confined to fewer genotypes, based on RAPD and MLEE typing, than genital tract isolates (Chatellier *et al*., 1997). The relative heterogeneity of our isolates probably is related to the wide age-range of patients from whom they were isolated.

GBS serotype III has been subdivided previously into two or more phylogenetic subgroups or subtypes on the basis of differences among *cps* gene clusters (Sellin *et al*., 1995), *hylB* gene plus RFLP (Bohnsack *et al*., 2001) or rRNA genes (Chatellier *et al*., 1997). Our subdivision of serotype III into four serosubtypes (sst), based on differences within the *cps* gene cluster (Kong *et al*., 2002a) was supported by corresponding differences in surface pgp (Kong *et al*., 2002b) and distribution of mge. We did not test our isolates for hyaluronidase activity, but it is likely that our sst III-1, which contains *rib,* IS*1548*, IS*861* and IS*1381*, corresponds with the hyaluronidase negative subtype III-2, described by Bohnsack *et al.* (2001), which also contains IS*1548*. Our sst III-2, with *rib*, IS*861* and GBSi1, probably corresponds with subtype III-3 of Bohnsack *et al.* (2002), which also contains GBSi1 and *rib* . Sst III-1 and III-2 are the predominant genotypes of MS III.

Sst III-3 and sst III-4 genotypes are apparently related to MS Ia and MS II, respectively. They share sequence homology in sections of the corresponding *cps* clusters (Kong *et al*., 2002a), and genotype Ia-7, MS II genotypes in cluster 3 and genotype II-10 are closely related to genotypes III-3, III-4-1 and III-4-2,

respectively (Figure 5.1.) (Kong *et al*., 2002b), suggesting that they have common ancestors. These results suggest that sst III-3 and III-4 have arisen by recombination between serotypes, as occurs quite commonly in *S. pneumoniae* (Jiang *et al*., 2001).

Our genotyping system, based on three sets of genetic markers, is highly discriminatory. Because it provides useful phenotypic data, including antigenic composition, it will be useful for epidemiological surveillance of GBS, especially in relation to potential GBS vaccine use. Further investigation of the distribution and multiplicity of mge and their associations with virulence markers is needed to determine their role, if any, in the pathogenesis of GBS disease. The relationships between putative high-virulence genotypes (Musser *et al*., 1989) and patient characteristics (age and/or underlying risk factors), and whether there are significant differences between CSF isolates (or genotypes) and other invasive or colonising strains, requires further study, which will be facilitated by our genotyping system. Using this system, we have demonstrated a clonal population structure among invasive Australasian GBS isolates and plan to compare it with that of colonising GBS isolates, in order to identify markers of virulence.