CHAPTER 7

IN SILICO INTRASPECIES COMPARATIVE GENOMICS OF *STREPTOCOCCUS AGALACTIAE* – IN THE CONTEXT OF GENETIC POPULATION STRUCTURE

Fanrong Kong, Gwendolyn L. Gilbert

Centre for Infectious Diseases and Microbiology, Insititute of Clinical Pathology and Medical Research, Westmead Hospital, Westmead, NSW 2145, Australia

Statement of Joint Authorship

Portions of this chapter will be included in a paper on **Comparative Genomics of *Streptococcus agalactiae*** to be written by Dr Glaser (leading author) ***et al.***, and will be submitted to *Journal of Bacteriology*.

Kong, F. (candidate)
Did all of the comparative and molecular analysis in the chapter, interpreted the data and wrote the manuscript.
Gilbert G. L. (supervisor)
Supervised the overall project. Assisted in the analysis and interpretation of data, and made a significant contribution to the manuscript.

7.1. SUMMARY

Streptococcus agalactiae is a commensal bacterium, which colonizes a significant proportion of the human population. However, it is also a cause of serious illness in newborns, pregnant women and adults with underlying chronic medical conditions. Two *S. agalactiae* genomes have been published, one for a serotype III (serosubtype III-3 in our genotyping system) strain NEM316 (ATCC 12403), and another for a serotype V strain 2603 V/R (ATCC BAA-611). The third *S. agalactiae* genome, for a serotype Ia strain A909 (ATCC 27591), will be completed and available soon. In order to better understand *S. agalactiae* heterogeneity and possible disease pathogenesis, we compared the general features of the two published genomes and analysed three sets of selected gene sequences, namely: *cps* gene clusters, surface protein antigen genes and mobile genetic elements. These *in silico* analyses revealed significant genetic heterogeneity between the two *S. agalactiae* genomes. In particular, most of the heterogeneity sites were clustered within about 19 genomic islands, which contribute most of the genetic diversity between the two genomes. Some of these genomic islands may be pathogenicity islands [PIs], with potentially important roles in virulence acquisition. Finally, based on the results of genotyping of 1,066 *S. agalactiae* isolates, we have shown that the three sequenced *S. agalactiae* strains are atypical human isolates and suggest that genome sequence data analysis should be interpreted in the context of *S. agalactiae* genetic population structure.

7.2. INTRODUCTION

Streptococcus agalactiae (group B streptococcus, GBS) was first recognized as a pathogen in bovine mastitis (Keefe, 1997). Although *S. agalactiae* usually behaves as a commensal organism that colonizes the gastrointestinal or genitourinary tract of 25-50% (Hansen *et al*., 2004) of healthy women, it can cause life-threatening invasive infection in susceptible hosts: newborn infants, pregnant women, and

nonpregnant adults with underlying chronic illnesses (Schuchat, 1998). Since guidelines, recommending intrapartum antibiotic prophylaxis (IAP) for high-risk or colonized women, were issued in 1996 the incidence of neonatal infections has decreased (Schrag *et al*., 2002). However, invasive *S. agalactiae* infections in immunocompromised adults, elderly persons and those with underlying chronical diseases have become relatively more common and a serious cause of morbidity and mortality (Schuchat, 1998; Farley, 2001).

Capsular polysaccharide is an important *S. agalactiae* virulence determinant. *S. agalactiae* is divided into at least nine known serotypes according to the antigenic reactivity of the capsular polysaccharide (Chaffin *et al*., 2000). Of the nine serotypes described so far, Ia, Ib, II, III and V are responsible for the majority of invasive human *S. agalactiae* diseases. Serotype III is particularly important because it causes a significant percentage of cases of early onset neonatal disease (EOD) and most late-onset disease (LOD) (Schuchat, 1998; Kong *et al*., 2003). Overall, serotype III is responsible for 80% of cases of neonatal *S. agalactiae* meningitis (Schuchat, 1998; Glaser *et al*., 2002). Serotype V is the commonest serotype associated with invasive infection in nonpregnant adults (Schuchat, 1998; Tettelin *et al*., 2002; Amaya *et al*., 2004). The genomes of the two serotypes – serotypes III and V – have been sequenced (Glaser *et al*., 2002; Tettelin *et al*., 2002). While a single genome analysis provides tremendous biological insights into GBS, intraspecies comparative genomics of multiple serotypes or strains provides substantially more information (Fraser *et al*., 2000; Whittam & Bumbaugh, 2002). To elucidate the heterogeneity and possible disease pathogenesis of *S. agalactiae*, we compared the genomes of the serotype III strain NEM316 (http://genolist.pasteur.fr/SagaList) (Glaser *et al*., 2002) and serotype V strain 2603 V/R (http://www.tigr.org) (Tettelin *et al*., 2002). In addition, we used our previously described genotyping system (Kong *et al*., 2003) to analyse the genetic population structure of 1,066 GBS isolates. This provides a context and guide for comparative analyses of *S. agalactiae* genome sequences (Joyce *et al*., 2002; Spratt & Maiden, 1999).

7.3. MATERIALS AND METHODS

7.3.1. Comparison of methods for prediction of open reading frames (ORFs).

The methods and software used for predicting ORFs, gene identification and annotation differed significantly for the two published genomes. In particular, more "stringent" parameters were used for analysis of NEM316 than for 2603 V/R (Glaser *et al.*, 2002; Tettelin *et al.*, 2002). For example, the predicted minimum protein/peptide length was 30 aa. for genome 2603 V/R, compared with 40 aa. for genome NEM316, which led to annotation of more short "genes" in genome 2603 V/R (see Results and Discussion section).

7.3.2. *In silico* genome comparison.

The two *S. agalactiae* genome sequences and gene lists were obtained through the Website of the National Center of Biological Information :
NEM316 (ATCC 12403) III-3
(http://www.ncbi.nlm.nih.gov:80/cgi-bin/Entrez/framik?db=genome&gi=264)
2603 V/R (ATCC BAA-611)
(http://www.ncbi.nlm.nih.gov:80/cgi-bin/Entrez/framik?db=genome&gi=252)

In order to examine the heterogenity of gene content, lengths and orders of genes in NEM316 and 2603 V/R genomes, we downloaded the "Feature tables" (protein coding genes and structural RNAs, including their start and stop locations) into our Microsoft Excel and Access files in Microsoft Windows 2000. This allowed us to demonstrate corresponding gene locations within the two genomes, after manual alignment of genes.

7.3.3. Sequence management, search, comparison, and multiple sequence alignments.

The Australian National Genomic Information Service (ANGIS) provided all programs used in the study (http://www1.angis.org.au/WebANGIS/): in particular, sequence file management (WebFM), sequence search (*BLAST* and *FastA* programs in Database Similarity Searches program group), two sequence comparison (*Bestfit* in Comparison program group), multiple sequence alignments (*Pileup* and *Pretty* in Multiple Sequence Analysis program group).

7.3.4. *In silico* restriction map and pulsed-field gel electrophoresis (PFGE).

The website (http://www.in-silico.com/) provided the online service of *in silico* restriction digest of complete genomes of NEM316 and 2603 V/R. The *Sma*I (Recognition sequence: CCC'GGG) was selected for theoretical (*in silico*) PFGE.

7.3.5. Genetic population study.

Our previously described *S. agalactiae* genotyping system (Kong *et al*., 2002a, b, 2003) was used to characterize 27 *S. agalactiae* reference strains and 1,039 clinical isolates from Australia, New Zealand, Canada, Korea, Japan, and Germany. Of the 1,039 clinical isolates, 900 were human invasive (about 320) or colonization (about 580) isolates and 139 were bovine milk isolates (Martinez *et al*., 2000). 115 out of 140 *bac* (encoding protein C•) positive reference strains and clinical isolates from Australia, New Zealand, and Germany were further subtyped based on *bac* gene sequencing heterogeneity (Kong *et al*., 2002b; Berner *et al*., 2002)

7.4. RESULTS AND DISCUSSION

7.4.1. Comparison of two genomes general features.

Table 7.1. General features of NEM316 and 2603 V/R genomes.

General features	NEM316	2603 V/R
Length (base pairs)	2, 211, 485	2, 160, 267
Gene number	2, 118	2, 175
Gene density (gene/kbp)	0.958	1.006
Biological roles assigned – gene no. (%)	1, 313 (62)	1, 333 (61)
Matched unknown function – gene no. (%)	529 (25)	623 (29)
No database match – gene no. (%)	276 (13)	219 (10)
G+C content (%)	35.6	35.7
Transcribed genes in one direction (%)	81	78
rRNA gene order (copy no.)	16S-23S-5S (7-7-7)	16S-23S-5S (7-7-2*)
Distribution range of 7 set rRNA (kbp)	455	406
tRNA gene number	80	80

Notes.

*See text for more explanation. Briefly, NEM316 and 2603 V/R have nearly identical sequences in the corresponding regions of the seven sets of rRNA genes. Unlike NEM316, in which seven copies of 5S rRNA were annotated, only two copies were annotated in 2603 V/R.

Though there are some differences (see below), the general features of NEM316 and 2603 V/R genomes are quite similar and are shown in Table 7.1.

7.4.1.1. The backbone of *S. agalactiae* genomes.

Orthologous genes typically have the same function (Tatusov *et al.*, 1997). In the study, we tried to align orthologous genes that were located at the same region. These genes can be seen as the "backbone" of the two *S. agalactiae* genome (Nakagawa *et al.*, 2003); gene insertions and deletions (indels) within the backbone were designated as "gene indels" (Gupta & Griffiths, 2002). Based on these considerations and our calculations, the total number of ORFs was 2,417 for both NEM316 and 2603 V/R. Our analysis showed that, apart from significant heterogeneity in the regions of several genomic islands (GIs) (see below and Table 7.5.) and some other islets or minor indels (Gupta & Griffiths, 2002; Britten *et al.*, 2003), the "backbones" of the two genomes are highly conserved (78.2% shared orthologous genes). This supports results of previous studies (Dmitriev *et al.*, 1998; Nakagawa *et al.*, 2003), and was itself supported by theoretical NEM316 and 2603 V/R genome *Sma*I restriction maps (see below).

7.4.1.2. *In silico* restriction map and PFGE.

Genome analysis showed that there were 24 *Sma*I cleavage sites in NEM316 and 21 in 2603 V/R (http://www.in-silico.com/). Based on our protein coding genes and structural RNA start and stop location files, we found (Table 7.2.) that 21 cleavage sites were located in corresponding gene regions within the backbones of the two genomes. No cleavage sites were located within rRNA operons, as previously described (Dmitriev *et al.*, 1998). Two of the three extra cleavage sites in NEM316 were located in the genomic islands, GIs X and XII (see below), and the third in the gbs660 and gbs661 intergenic spacer region. This suggested that most differences in PFGE patterns, between closely related strains are due to indels (Gupta & Griffiths, 2002; Britten *et al.*, 2003). In another words, heterogeneity of *S. agalactiae* PFGE

patterns – due to either fragment length differences or the presence of additional bands – is largely due to indels (Dmitriev *et al*., 1998). If indels (without cleavage sites within indels) were introduced within the regions of two cleavage sites, they will mainly cause the fragment length differences; if indels contained cleavage sites, they may cause significant differences in both fragment lengths and band numbers.

7.4.1.3. Gene density.

The gene density in genome 2603 V/R is apparently higher than that in genome NEM316, which is mainly due to differences in annotation methods resulting in a great number of shorter genes in 2603 V/R. For example, in NEM316, there were 211 coding sequences (CDS) of fewer than 100 codons and 20 of fewer than 50 codons, compared with 288 and 82 respectively, in 2603 V/R. Mira *et al.* (2002) have suggested that the number of shorter genes in some microbial genomes have been overestimated, resulting in annotated gene lists containing a number of ORFs that are not true genes.

7.4.1.4. Annotation errors or real heterogeneity?

Comparison of the annotations of the two genomes showed that there were at least 101 genes (~5%) for which apparent differences in length were probably due to different annotation start points. However, there were at least 56 genes (~2.5%), in which the in-frame stop codon or small deletions or insertions (minor indels) that led to frame shifts. In addition, heterogeneity in at least 32 genes (~1.6%) was related to their being annotated as transcribed or nontranscribed non-functional pseudogenes (caused by premature stop codons and frame shifting mutations) (Mounsey *et al*., 2002).

Table 7.2. ***Sma*I restriction map of NEM316 and 2603 V/R genomes.**

NEM316 cleavage position	NEM316 length of sequence	Length difference of sequence[a]	2603V/R cleavage position	2603V/R length of sequence
954	958	0	954	958
5826	4872	1	5825	4871
6786	960	0	6785	960
10179	3393	0	10178	3393
31170	20991	-178	31347	21169
74945	43775	320	74802	43455
75905	960	0	75762	960
148661	72756	-313	148831	73069
149621	960	0	149791	960
235298	85677	1510	233958	84167
236258	960	0	234918	960
295087	58829	528	293219	58301
334032	38945	-1	332165	38946
334992	960	0	333125	960
450253	**115261**	**47077**	**401309**	**68184**
451213	960	0	402269	960
677272	**226059**	**226059**	-	-
988118	**310846**	**-252297**	**965412**	**563143**
1258572	**270454**	**270454**	-	-
1410444	**151872**	**151872**	-	-
1744390	**333946**	**-341988**	**1641346**	**675934**
2071569	**327179**	**-56430**	**2024955**	**383609**
2153859	82290	4605	2102640	77685
2211481	57622	-1	2160263	57623

Notes.

a. The bold numbers show significant differences in *Sma*I cleavage sites and fragment lengths between the two genome. The underlined sites are three extra *Sma*I sites in NEM316.

b. Length difference of sequence was calculated by NEM316 length of sequence minus 2603 V/R length of sequence.

Because of the very accurate genome sequencing, the minor indels could represent genuine differences (Gupta & Griffiths, 2002; Britten *et al.*, 2003), but may also include some pseudo-heterogeneity or false predictive results (Mira *et al.*, 2002).

7.4.1.5. rRNA gene complex and tRNA genes.

Apart from some minor heterogeneities (see below), the RNA genes generally were highly conserved between NEM316 and 2603 V/R. It is of note that both genomes contained seven rRNA operons rather than six as previously reported (Dmitriev *et al.*, 1998).

101 RNA genes were annotated in NEM316 compared with 96 in 2603 V/R. This was due to five extra copies of 5S rRNA genes that were annotated in NEM316 only. Further analysis showed that the sequences of seven copies of the rRNA gene complex in NEM316 and the corresponding regions in 2603 V/R were otherwise nearly identical. Therefore we assume that the apparent difference in the number of copies of the 5S rRNA was an annotation artefact. The two copies of 5S rRNA genes that were annotated in both genomes were given different lengths, 143 bp in NEM316 and 162 bp in 2603 V/R. Further, the two copies of 5S rRNA genes showed significant sequence heterogeneity compared with the other five copies, which may contribute to differences in annotation between the two strains.

All seven copies of 23S rRNA genes in both genomes were with the same length (2,903 bp), but the seven copies of 16S rRNA gene in NEM316 (1,409 bp) were apparently 98 bp shorter than those in 2603 V/R (1,507 bp) because of a 39 bp (upstream) and 59 bp (downstream) shorter annotation.

There was one heterogeneity site in the 16S rRNA gene at bp 193 (according to the 2603 V/R 16S rRNA gene start point), which was G in NEM316, and A in 2603 V/R. This is consistent with previously reported sequences in GenBank (AB023574, AF459432, AF015927, X59032, AB002479 and AB002480). There was another

heterogeneity site in the 5S rRNA gene at bp 12 (according to the 2603 V/R 5S rRNA gene start point), which was C in NEM316 and T in 2603 V/R. Besides these two strain level heterogeneities, NEM316 16S rRNA gene operon 6 (the copy located at 350560-351968) had a T at bp 807 (according to 2603 V/R 16S rRNA gene start point) compared with C at this site in all other six copies. This was the only inter-copy heterogeneity site between 14 copies of 16S rRNA genes in NEM316 and 2603 V/R.

Both NEM316 and 2603 V/R have 80 tRNA genes and no tRNA gene length heterogeneity was found. However, we found genome fragment rearrangements at two sites related to differences in the order of the tRNA genes. They were tRNA-Asp-tRNA-Lys-tRNA-Leu-tRNA-Thr within 2603 V/R region SAG0085-SAG0086 and tRNA-Gly within the corresponding NEM316 region, gbs0085-gbs0086; versus tRNA-Asp-tRNA-Lys-tRNA-Leu-tRNA-Thr within NEM316 region gbs0445-gbs0446 and tRNA-Gly within the corresponding 2603 V/R region, SAG0410-SAG0411. Rolland *et al.* (2002) also identified a rearrangement corresponding to that in NEM316, but only in highly virulent strains. This is consistent with the fact that strain NEM316 was isolated from a fatal case of *S. agalactiae* sepsis (Glaser *et al*., 2002).

7.4.2. Three sets of virulence-related molecular markers.

S. agalactiae expresses a variety of products, which are implicated in virulence. Among these are the products of two sets of molecular markers included in our previously described genotyping system (Kong *et al*., 2003), namely the capsular polysaccharide synthesis (*cps*) gene clusters and surface proteins genes. In addition, a number of mobile genetic elements (mge) are associated with various other specific virulence factors. We used these three sets of virulence-related molecular markers to further compare the two genomes.

7.4.2.1. Capsular polysaccharides synthesis (*cps*) gene clusters.

S. agalactiae possesses two cell wall-associated surface polysaccharides: group B specific carbohydrate common to all *S. agalactiae* serotypes and the serotype-specific capsular polysaccharide – one of the most important *S. agalactiae* virulence factors. Serotype-specific capsular polysaccharide prevents deposition of the host complement factor C3b and inhibits complement-mediated opsonophagocytosis (Chaffin *et al*., 2000; Glaser *et al*., 2002).

The 2603 V/R and NEM316 *cps* gene clusters consist of 19 and 17 genes, respectively, including the transcriptional regulatory gene *cpsY* (Koskiniemi *et al*., 1998). Regions encoding glycosyltransferases and related proteins (SAG1162-SAG1170/gbs1237-gbs1243), direct the synthesis of the respective polysaccharide repeat units. They comprise nine genes in 2603 V/R but only seven in NEM316, from which genes corresponding to SAG1166 and SAG1167 are missing. Of the seven shared genes, three are the same length and four have length and sequence heterogeneities. This serotype-specific region is flanked on either side by genes conserved in all *S. agalactiae* serotypes (Chaffin *et al*., 2000; Glaser *et al*., 2002; Tettelin *et al*., 2002). Downstream are genes that encode enzymes for biosynthesis and activation of sialic acid (SAG1158-SAG1161/gbs1233-gbs1236). Upstream are genes (SAG1171-SAG1175/gbs1244-gbs1248) found not only in all nine *S. agalactiae* serotypes but also in a variety of other polysaccharide-producing streptococci (Chaffin *et al*., 2000).

The sequences of all 19 genes of the 2603 V/R *cps* gene cluster are largely consistent (99.995%) with those of a serotype V strain, CNCTC 1/82 (ATCC 49446) previously deposited in GenBank (AF349539: 18,239bp). There is only one base heterogeneity or mutation at the 5'-end of *cpsK*; it was C in CNCTC 1/82 but T in 2603 V/R. In addition, at the 3'-end of *cpsD*, 2603 V/R does not contain a 9 bp repetitive sequence (TTACGGCGA), which is present in CNCTC 1/82 and all serotype V isolates that we have studied (see below). This finding, our genetic population analysis based on more than 1,000 isolates (see below), comparative genome hybridization (CGH) and phylogeny studies (Tettelin *et al*., 2002) and

multilocus sequence typing (MLST) (Jones *et al*., 2003) all indicate that 2603 V/R is an atypical serotype V strain, which is closely related to serotype II (serotype II may or may not have the 9 bp repetitive sequence).

The sequences of all 17 genes of the NEM316 *cps* gene cluster are largely consistent (99.519%, with five gaps) with those of a serotype III strain, COH1 (a serosubtype III-2 according to our genotyping system) previously deposited in GenBank (AF163833: 17,276 bp) (Chaffin *et al*., 2000). However, one region (4,411 bp) of the NEM316 sequence was nearly identical (99.995%=4409/4411) to the corresponding region (AF332897) in a serosubtype III-3 reference strain (NZRM 912 [NCDC SS 620]), previously sequenced by us. Our analysis showed that among the 81 heterogeneity sites between COH1 and NEM316, 59 sites were identical with corresponding sites in serotype Ia strain OI1 (AB028896: 25,021 bp), 22 sites were different from either AB028896 (Ia) or AF163833 (III-2) and so were assumed to be serosubtype III-3 specific. Four of the five gaps caused by 1 bp insertion in COH1 were assumed to be due to sequencing errors or mutations of AF163833 after careful comparison with the other known *cps* gene cluster sequences (GenBank accession numbers: AB028896, AB050723, AF355776, AF349539 and AF337958, respectively; and the two genome *cps* gene clusters). The fifth gap was due to the fact that NEM316 contains the 9 bp repetitive sequence (TTACGGCGA) at the 3'-end of *cpsD*, but COH1 (III-2) does not – a difference that distinguishes serosubtypes III-2 and III-3 in our genotyping system (Kong *et al*., 2002a). This finding, as well as the presence of Alp2 (see below) and MLST results (Jones *et al*., 2003), indicate that NEM316 belongs to our genotype III-3, which is closely related to serotype Ia (most of which have the 9 bp repetitive sequence), probably as a result of recombination.

7.4.2.2. Surface and secreted proteins.

Some *S. agalactiae* surface and secreted proteins are potential virulence factors or targets of protective immunity. Selected "virulence"-related proteins in NEM316

and 2603 V/R genomes are shown in Table 7.3. The amino acid heterogeneities of the cell wall anchored proteins of the two genomes are shown in Table 7.4. 19 of 34 (56%) cell wall related proteins differed in length, between the two genomes. Although both genomes shared many proteins in both categories, there was considerable heterogeneity (length or binary diversity; Tables 7.3. and 7.4.) and their significance in pathogenesis needs to be studied further.

The function of the protein gbs1087 (410 aa long), which does not have any streptococcal homolog, was unknown at the time that the NEM316 genome was published (Glaser *et al*., 2002). It has now been identified as a fibrinogen receptor, encoded by the ***fbsA*** gene in GBS. Sequencing of this gene from five different GBS isolates revealed variable numbers of contiguous copies of the motif LERRQRDAENR/KSQGNV (Schubert *et al*., 2002). NEM316 contains 16 copies of this repeat-encoding unit. However, 2603 V/R SAG1052 (corresponding to gbs1087) contained only 47 aa, corresponding to the C-terminal of FbsA, and no copy of motif LERRQRDAENR/KSQGNV (Tettelin *et al*., 2002), which further illustrated the significant variation in this gene (Schubert *et al*., 2002).

Each strain of *S. agalactiae* usually encodes one member of the surface protein family (Rib, alpha C or alpha C-like protein), all of which contain variable series of tandem repeat units (Heden *et al*., 1991; Michel *et al*., 1992; Wastfelt *et al*., 1996; Lachenauer *et al*., 2000). Variation in the number of repeats can change the antigenicity of these proteins, and is a mechanism to escape host immunity (Gravekamp *et al*., 1998; Lachenauer *et al*., 2000). However, the N and the C terminal parts of the protein are conserved. We studied the surface protein gene sequences in NEM316 and 2603 V/R in detail.

C• -like protein 2 (gbs0470) was identified in NEM316, which indicates that the strain belongs to the molecular subserotype III-3 (Kong *et al*., 2002b). Our genotyping study (see below) showed that besides in III-3, Alp2 is also found in a proportion of serotype Ia strains (6.0% of a total of 135 Ia human isolates), but

rarely in other serotypes. Compared with the GenBank sequence AF208158 – an Alp2 from a serotype V strain – the NEM316 Alp2 (gbs0470) contained a 340 aa. fragment duplication, giving an extra copy of U+A+BB, as designated by Lachenauer *et al*. (2000), which results in a significantly different protein structure between the two strains. Sequencing of *alp2* gene from multiple isolates, as previously reported for *alp3* (Lachenauer *et al*., 2000) would help to elucidate the significance of this finding.

Rib (SAG0433) was identified in 2603 V/R. It had 14 tandem repeats, two more than that in a previously published Rib sequence (GenBank number: U58333) (Wastfelt *et al*., 1996). Each tandem repeat encodes 79 aa. – not 67 aa. as reported by Tettelin *et al*. (2002). The Rib protein has previously been detected predominantly in *S. agalactiae* strains of serotypes II and III, whereas serotype V strains generally express a related member of the protein family, Alp3 (Lachenauer *et al*., 2000). This supports previous evidence from CGH and phylogeny studies (Tettelin *et al*., 2002) and multilocus sequence typing (MLST) (Jones *et al*., 2003) that indicates that 2603 V/R is closely related to serotype II. This was also supported by our genetic population analysis based on more than 1,000 isolates (see below).

Our analysis showed that the genome regions flanking the surface protein genes were conserved between the two genomes, especially the four genes at the 5'-end (SAG429-SAG432) (gbs466-gbs469) and five genes at the 3'-end regions (upstream of SAG0438-upstream of SAG0439) (gbs482-gbs486), respectively. The 5' and 3' ends are also conserved between members of the gene family (Lachenauer *et al*., 2000). This made it possible to design common primer pairs to amplify and sequence the other genes in the family (a, as etc.).

Table 7.3. NEM316 and 2603 V/R selected "virulence"-related proteins.

Proteins/descriptions	NEM316-ORF	2603 V/R-ORF	References
Sip	gbs0031	SAG0032	(Brodeur *et al*., 2000)
CAMP	gbs2000	SAG2043	(Lang & Palmer, 2003)
R5 (or BPS protein)	**-**	**SAG1331**	(Erdogan *et al*., 2002)
Enolase	gbs0608	SAG0628	(Hughes *et al*., 2002)
Hyaluronidase	gbs1270	SAG1197	(Pritchard *et al*., 1994)
Hemolysin/cytolysin	**gbs0651-Unknown**	**SAG0669-*cylE***	(Pritzlaff *et al*., 2001)
Lmb	gbs1307	SAG1234	(Franken *et al*., 2001)
ScpB	**gbs1308-1150aa.**	**SAG1236-NA**	(Franken *et al*., 2001)
ScpB-like	gbs0451	SAG0416	(Tettelin *et al*., 2002)
ScpB-like	gbs2008	SAG2053	(Glaser *et al*., 2002)
Rib	**-**	**SAG0433**	(Wastfelt *et al*., 1996)
Alp2	**gbs0470**	**-**	(Lachenauer *et al*., 2000)
Pullulanase	gbs1288	SAG1216	(Hytonen *et al*., 2003)
Neuraminidase	gbs1919	SAG1932	(Shakhnovich *et al*., 2002)
Adenylate kinase	gbs0079	SAG0079	(Bert *et al*., 1995)
Hsa-like	**gbs1529-1310aa.**	**SAG1462-970aa.**	(Takahashi *et al*., 2002)
FbsA	**gbs1087-410aa.**	**SAG2063-47aa.**	(Schubert *et al*., 2002)
Metalloprotease	gbs1279	SAG1206	(Blue *et al*., 2003)
Fibronectin-binding protein	gbs1263	SAG1190	(Holmes *et al*., 2001; Chhatwal, 2002)
NanA	gbs1919	SAG1932	(Tong *et al*., 2000)
Neuraminidase	gbs1919	SAG1932	(Tong *et al*., 2001)
SrtA	gbs0949	SAG0961	(Ilangovan *et al*., 2001)
SrtB	gbs0630	SAG0647	(Pallen *et al*., 2001)
SrtB	**gbs0631-283aa-c**	**SAG0648-260aa**	(Glaser *et al*., 2002)
SrtB	**gbs1476-292aa**	**SAG1406-293aa**	(Glaser *et al*., 2002)
SrtB	gbs1475	SAG1405	(Glaser *et al*., 2002)

Sortase pseudogene	**gbs0633-80aa**	**SAG0650-189aa**	(Ilangovan *et al*., 2001)
Prolipoprotein diacylglycerol transferase	gbs0758	SAG0737	(Petit *et al*., 2001)
Signal peptidase II	gbs1436	SAG1366	(Petit *et al*., 2001)
ClpX	gbs1383	SAG1312	(Mei *et al*., 1997)
ClpC	gbs1869	SAG1828	(Nair *et al*., 2000)
ClpL	gbs1367	SAG1294	(Kwon *et al*., 2003)
ClpE	gbs0535	SAG0488	(Nair *et al*., 1999)
ClpA ATPase paralogs	**gbs0718-610aa,**	**-**	(Glaser *et al*., 2002)
ClpA ATPase paralogs	**gbs0991-639aa**	**-**	(Glaser *et al*., 2002)
ClpA ATPase paralogs	**gbs0388-610aa**	**-**	(Glaser *et al*., 2002)
Rgg-like paralogs	gbs0230	SAG0239	(Kreikemeyer *et al*., 2003)
Rgg-like paralogs	gbs1555	SAG1490	(Kreikemeyer *et al*., 2003)
Rgg-like paralogs	gbs2117	SAG2158	(Kreikemeyer *et al*., 2003)
RofA/Nra-like paralogs	**gbs1426-503aa**	**SAG1356-503aa**	(Beckert *et al*., 2001)
RofA/Nra-like paralogs	**gbs1479-509aa**	**SAG1409-NA**	(Beckert *et al*., 2001)
RofA/Nra-like paralogs	**gbs1530-498aa**	**SAG1463-NA**	(Beckert *et al*., 2001)

Notes.

Bold characters indicate heterogeneity (length or binary diversity) between NEM316 and 2603 V/R.

Abbreviations: Sip – surface immunogenic protein; CAMP – (discovered by) Christie, Atkins, and Munch-Petersen; R5 – group B protective surface (BPS) protein; Lmb – laminin-binding protein; ScpB – C5a protease; Rib – resistance to proteases, immunity, group B; Alp2 – alpha-like protein 2; Hsa – (antigen that recognises) sialic acid-containing host receptors; FbsA – A fibrinogen receptor from group B; NanA – sialic acid lyase (catalyzes the hydrolysis of sialic acid into pyruvate and N-acetylmannosamine); Srt – sortase; Clp – Clp ATPase family of molecular chaperones; Rgg – encode a response regulator; Rof – encodes a response regulator; Nra – encodes a response regulator (no response to atmospheric conditions).

Table 7.4. The cell wall protein heterogeneity of NEM316 and 2603 V/R.

Cleavage motif	NEM316-ORF	NEM316-surface anchor proteins	NEM316-Size (a.a)	2603 V/R-Size (a.a)	2603 V/R-surface anchor proteins	2603 V/R-ORF
LPXTG	**gbs0391**	**Sec10**	**753**	**-**	-	-
LPXTG	gbs0392	Plasmid-encoded protein	**240**	**-**	-	-
LPXTG	gbs0393	**SpaA, Pas**	933	**-**	-	-
LPXTG	gbs0428	Cell surface protein	521	521	cell wall surface anchor family protein	SAG0392
LPXTG	**gbs0470**	**Alp2**	**1126**	**1389**	**Rib**	**SAG0433**
LPXTG	**gbs0479**	**Plasmid-encoded protein**	**253**	**-**	-	-
LPXTG	gbs0791	EaeH	512	512	cell wall surface anchor family protein	SAG0771
LPXTG	**gbs1087**	**Antigen p200**	**410**	**47**	**cell wall surface anchor family protein, putative**	**SAG1052**
LPXTG	**gbs1143**	**SpaA**	932	**-**	-	-
LPXTG	**gbs1144**	**Plasmid-encoded protein**	**236**	**-**	-	-
LPXTG	**gbs1145**	**Sec10**	743	**-**	-	-
LPXTG	gbs1288	PulA	1252	1252	pullulanase, putative	SAG1216
LPXTG	**gbs1356**	**Ssp-5, Pas**	**1634**	**1631**	**agglutinin receptor**	**SAG1283**
LPXTG	**gbs1420**	**Cell surface protein, CbpD**	**543**	**544**	**surface antigen-related protein**	**SAG1350**
LPXTG	gbs1474	Cell surface protein	308	308	cell wall surface anchor family protein	SAG1404

LPXTG	**gbs1529**	**Hsa, SrpA**	**1310**	**970**	**cell wall surface anchor family protein**	**SAG1462**
LPXTG	gbs1539	No homology in public databases Unknown	192	192	cell wall surface anchor family protein	SAG1473
LPXTG	gbs1540	AmiC, YbgE	680	680	amidase family protein	SAG1474
LPXTG	gbs1929	CpdB, YfkN	800	800	2`,3`-cyclic-nucleotide 2`-phosphodiesterase	SAG1941
LPXTG	gbs2008	PrtS	1570	1570	serine protease, subtilase family, putative	SAG2053
LPXTG	**gbs2018**	**M-like protein, PspC**	**643**	**630**	**pathogenicity protein, putative**	**SAG2063**
IPXTG	gbs0628	Hypothetical protein, Cell surface protein	554	554	cell wall surface anchor family protein	SAG0645
IPXTG	gbs0629	No homology in public databases Unknown	307	307	cell wall surface anchor family protein	SAG0646
IPXTG	**gbs1477**	**No homology in public databases Unknown**	**674**	**705**	**cell wall surface anchor family protein**	**SAG1407**
IPXTG	gbs1478	PFBP, Cell surface protein	901	901	cell wall surface anchor family protein	SAG1408
LPXTS	gbs0451	ScpB	1233	1233	protease, putative	SAG0416
LPXTS	gbs0456	SPy0843, BspA,	1055	1055	cell wall surface anchor family protein	SAG0421
LPXTN	**gbs1308**	**ScpB**	**1150**	**NA**	**C5a peptidase, authentic frameshift**	**SAG1236**
LPXTN	gbs1403	SPy0872	690	690	5'-nucleotidase family protein	SAG1333
LPSTG	**-**	**-**	**-**	**1062**	**hypothetical protein**	**SAG0677**

LPTTG	-	-	-	979	**R5 protein**	SAG1331
LPKTG	-	-	-	**263**	cell wall surface anchor family protein, putative	**SAG1996**
LPQTG	-	-	-	**826**	**cell wall surface anchor family protein, putative**	**SAG2021**
FPKTG	gbs0632	Cell surface protein	890	890	cell wall surface anchor family protein, putative	SAG0649

Notes.

Bold characters indicate some heterogeneity (length or binary diversity) between NEM316 and 2603 V/R.

Abbreviations:

Sec10 – Surface exclusion protein; SpaA – streptococcal protein antigen A of *Streptococcus sobrinus*; Pas – the surface protein antigen I/II of *Streptococcus intermedius*; Rib – resistance to proteases, immunity, group B; Alp2 – alpha-like protein 2; EaeH – EaeH of *Escherichia coli* O157:H7; PulA – Alkaline amylopullulanase; Ssp5 – agglutinin receptor; CbpD – choline binding protein D; Hsa – sialic acid-binding protein; SrpA – periplasmic linker protein; AmiC – amidase family protein; YbgE – putative branched-chain aminotransferase; CpdB – Cyclo-nucleotide phosphodiesterase; YfkN – 2`,3`-cyclic-nucleotide 2`-phosphodiesterase; PrtS – Serine proteinase, subtilase family; PspC – pneumococcal surface protein C; PFBP – *Streptococcus pyogenes* fibronectin-binding protein; ScpB – Serine protease and C5a peptidase; BspA – a cell surface associated leucine-rich repeat protein involved in adhesion to fibronectin and fibrinogen; R5 (or BPS protein) – group B protective surface protein.

7.4.2.3. Mobile genetic elements (mge) and possible pathogenicity islands (PIs).

S. agalactiae cps gene clusters, surface proteins, hemolysin, and several transcriptional regulators are believed to play a role in colonization or disease (Manning, 2003). Many mge, including bacteriophages, transposons and insertion sequences, are associated with acquisition of virulence traits from other species/strains. In the two *S. agalactiae* genomes, possible pathogenicity islands (PIs) were found, which contain the majority of known or putative virulence genes and many of the known mge (Glaser *et al.*, 2002; Tettelin *et al.*, 2002). An exciting evolutionary hypothesis is that pathogenic *S. agalactiae* have gradually evolved through successive acquisition of exogenous virulence factors carried by such islands (Glaser *et al.*, 2002). In particular, the emergence of hyper-virulent *S. agalactiae* clones might result from such horizontal gene transfer (Blumberg *et al.*, 1992; Musser *et al.*, 1989; Quentin *et al.*, 1995).

In genome 2603 V/R the following mge were present:

- Six copies of both IS*1381*-A (no heterogeneity) and IS*1381*-B (no heterogeneity);
- five copies of IS*1548* (one heterogeneity site: 3T/2C at bp 859);
- two copies of IS*861*-A (one heterogeneity site compared with a previous IS*861* sequence in GenBank, M22449: a ACATGATAA 9 bp repetitive at bases 223-232); three copies of IS*861*-B (two of which [SAG1068, SAG1527] have four heterogeneity sites: A/G at bp 97, T/C at bp 534, A/G at bp 678, T/C at bp 711; the third [SAG1256] has significant heterogeneity compared with the other two);
- two copies of both IS*Sag2* (IS*Sdy1*)-A (one heterogeneity: T/C at bp 263) and IS*Sag2* (IS*Sdy1*)-B (one heterogeneity: G/A at bp 511); and
- one copy each of IS*Sag1*, and GBSi1.

In genome NEM316, of the mge described so far in *S. agalactiae* (IS*861*, IS*1381*, IS*1548*, IS*Sa4*, IS*Sa4*, IS*Sag1*, IS*Sag2* and GBSi1), only two copies of IS*Sag2*, bracketing the *scpB* and *lmb* genes, were found (Franken *et al.*, 2001), with different length annotations between IS*Sag2* (IS*Sdy1*)-A and IS*Sag2* (IS*Sdy1*)-B compared with 2603 V/R. Six novel putative IS elements were identified and, of these, only transposon

gbs0208 does not seem to have been inactivated by frame shift mutations. Although no complete or cryptic prophage was identified in the NEM316 genome, a striking observation was the identification of a large number of plasmid- and phage-related genes. 12 genes encoding proteins related to plasmid functions (replication, partition or transfer), often in the vicinity of integrase genes and 12 genes encoding proteins similar to phage integrases were identified in the NEM316 genome (Glaser *et al.*, 2002).

CGH using DNA microarrays was performed between the sequenced 2603 V/R and 19 other *S. agalactiae* strains of multiple serotypes. 401 genes detected in strain 2603 V/R were not detected in one or more other strains, suggesting that they are absent or significantly divergent. 364 (91%) of the 401 varying genes were present in 15 regions containing five or more contiguous genes (Table 7.5.). Ten of these regions display an atypical nucleotide composition (compared with the average genome G+C content) in strain 2603 V/R, consistent with the possibility that they were laterally transferred into this strain (Hacker & Carniel, 2001). They contain many glycosyltransferases, cell-wall anchor proteins, and phage-related genes (Tettelin *et al.*, 2002). These findings suggest that some of these regions are virulence related or pathogenicity islands (Nakagawa *et al.*, 2003).

Of 945 genes without orthologs in *S. pyogenes* (the closest relative of *S. agalactiae)* 471 are clustered in 14 large islands containing 11-77 genes, including three copies of the integrative plasmid pNEM316-1 (Glaser *et al.*, 2002). These 14 islands contain all of the mge-related genes (including the two copies of IS*Sag2*), except the only intact novel IS element (gbs0208) and, most importantly, the majority of known or putative virulence genes of *S. agalactiae* (Table 7.5.). This means that some of these regions may be defined as pathogenicity islands (PIs). These islands also contain most of the pseudogenes identified, as well as genes probably mediating horizontal gene transfer, strongly suggesting that they undergo rapid evolution.

7.4.3. Genetic population studies.

We have studied 1,066 *S. agalactiae* isolates, using our genotyping system (Kong *et al*., 2003). Among 27 reference strains and 900 human GBS isolates, more than 99 genotypes were found if excluding *bac* sequence subtypes (Berner *et al*., 2002; Kong *et al*., 2002b). If subtypes identified according to *bac* gene sequence heterogeneity are included (Berner *et al*., 2002; Kong *et al*., 2002b), another 38 new genotypes were introduced into our new genotype database based on sequencing of 115 of 140 *bac* positive reference strains and clinical isolates. However, a recent study using PFGE (which should be very sensitive) (Dmitriev *et al*., 2002) showed that *S. agalactiae bac* gene-positive strains are genetically homogenous. Therefore, the significance of sequence variation in *bac* is doubtful and we did not sequence the other *bac* positive strains. Among 139 bovine *S. agalactiae* isolates, there were 50 polyphasic types (conventional serotype and genotypes [as discussed above] were considered together to identify "polyphasic types", which have greater discriminatory power for bovine *S. agalactiae* isolates). Generally, human and bovine *S. agalactiae* are two different populations, but a small minority (9/139; 6.5%) shared the same genotypes (three Ia-1, three III-3, one III-1 and two III-2 strains) as human isolates.

Our study showed that the two published *S. agalactiae* genomes are atypical among their corresponding serotypes. Among 900 human isolates, 228 were serotype III and, of these, only six (3%) were serosubtype III-3. Five of these isolates had IS*Sag1* but only one, an isolate from a German patient with early onset neonatal disease (EOD) (Berner *et al*., 1999), had the same genotype as NEM316. Interestingly, III-3/Ia *cps* sequence type was very common among bovine isolates (86 of 139 [61.9%]), and one of the above three bovine *S. agalactiae* III-3 strains was identical with NEM316 genotype. Our hypothesis is that NEM316 and other human III-3 strains may have originated from cattle strains.

Among 92 isolates belonging to serotype V, in our *S. agalactiae* collection, there were only six (6.5%) of the V-R serovariant (four V-RB, and two V-R), and only one was the same genotype as 2603V/R.

Table 7.5. Possible genomic islands in NEM316 and 2603 V/R genomes.

NEM316 PIs	NEM316 annotation	Gene identity	2603 V/R PIs	2603 V/R annotation
I	Integrase, Plasmid replication, Recombinase/resolvase	15/25	1	Integrase, Plasmid replication, Recombinase/resolvase
II	Integrase, Plasmid replication, DNA translocase	7/19	2	- - -
III	pNEM316-1: Plasmid replication, topoisomerase, Single strand binding protein, Plasmid transfer complex protein, Plasmid partition protein, Plasmid replication initiation	0/49	-	- - - - - -
IV	**Alp2**, Integrase, Phage related proteins, Plasmid related proteins	8/23	3	**Rib,** - - -
V	Integrase, Transposase, -	7/10	4	- Transposase (IS), **Prophage**
VI	***cyl* operon**, Transposase	41/58	5, 6	***cyl* operon**, Transposase
VII	pNEM316-1 (same as above III)	0/49	-	-
-	-	-	7	**Tn*916*,** **IS*1548***
VIII	pNEM316-1 (same as above III)	0/49	-	-
IX	DNA translocase	17/27	8	DNA translocase
X	Plasmid relaxase and mobilisation, Transfer complex proteins TrsK, Transfer complex proteins TrsE, Plasmid replication initiation	0/36	-	- - - -

XI	Integrase	3/11	-	-
-	III-3 capsule locus	-	9	V capsule locus
XII	**Lmb and ScpB**, Transposase, DNA polymerase, Exonuclease, Integrase, Plasmid replication, Type II DNA modification, Transposon relaxase, Helicase, Plasmid transfer complex proteins TraE, Plasmid transfer complex proteins TrsK	18/70	10	**Lmb and ScpB**, Transposase, DNA polymerase, - Integrase, Plasmid replication, - - Helicase, Plasmid transfer complex proteins TraE, Plasmid transfer complex proteins TrsK
-	Group B carbohydrate synthesis	-	11	Group B carbohydrate synthesis
-	-	-	12	IS*1381*, etc.
-	-	-	13	Prophage
XIII	**CAMP** factor, Integrase	26/47	14	**CAMP** factor, -
XIV	Plasmid replication protein, Integrase	13/22	15	Plasmid replication protein, Integrase

Notes.

See text for definition of genomic islands (GIs). Bold characters indicate some previous known virulence-related factors.

Abbreviations: Lmb – laminin-binding protein; ScpB – C5a protease; CAMP – (discovered by) Christie, Atkins, and Munch-Petersen; R5 (or BPS protein) – group B protective surface protein; Rib - resistance to proteases, immunity, group B; Alp2 – alpha-like protein 2; Hsa – (antigen that recognition of) sialic acid-containing host receptors; *cyl* operon – a genetic locus encoding the GBS beta-haemolysin/cytolysin activity.

We also noticed that the Ia strain (A909), whose genome is currently being sequenced, is a Ia-AaB serovariant strain, which is also rare in our collection (eight of 137 [5.8%] Ia isolates, of which three are reference strains).

It seems likely that these three atypical strains have arisen as a result of recombination events namely: for NEM316 (III-3 serovariant), between serotypes III and Ia-alp2as; for 2603 V/R (V-R serovariant), between serotype V and II-R; and for A909 (Ia-AaB serovariant), between Ia and Ib-AaB. Alternatively, these three atypical strains may indicate that *S. agalactiae* is more heterogenous than previously thought (Kong *et al*., 2003).

In future comparative genomic studies we need to consider the following factors:

1. Is "true" heterogeneity (excluding sequencing or annotation errors, as discussed above) reflected in the level at which difference occurs, e.g. between serotypes or strains?
2. If the recombination hypothesis is true, which serotypes are the parental strains? Are our suggested parental strains for the three sequenced strains correct?
3. We await, with interest the full sequence of the genotype III-2 (Kong *et al*., 2003), which represents the "main stream" of serotype III and has been shown, in various studies, to be highly virulent (Bohnsack *et al*., 2002; Jones *et al*., 2003).

The combination of all known genes from the four genomes into a microarray and its use to further study *S. agalactiae* population genetics would provide much more useful information for *S. agalactiae* evolution, heterogeneity, and pathogenesis/virulence.

7.5. CONCLUSION

We compared the general features of the two published genomes and analysed three sets of selected gene sequences. These *in silico* analyses revealed significant genetic heterogeneity between the two *S. agalactiae* genomes, but their backbones are conserved. Finally, based on the results of genotyping of 1,066 GBS isolates, we have shown that the

two sequenced *S. agalactiae* strains and one ongoing Ia genome are atypical human *S. agalactiae* isolates and suggest that genome sequence data analysis should be interpreted in the context of GBS genetic population structure.