Integrated study of group B streptococcus and human ureaplasmas – the paradigm shifts

Fanrong Kong

A thesis submitted to the University of Sydney in fulfilment of the requirement for the degree of Doctor of Philosophy.

Department of Medicine University of Sydney

2004

Declaration

This thesis is less than 100, 000 words in length and contains no material that has been accepted for the award of any other degree or diploma in any other universities.

I declare that the author, with the following exception, conducted all the practical work incorporated into this thesis:

- All the New Zealand GBS strains were provided by Dr Diana Martin (ESR, New Zealand) and GBS conventional serotyping was mainly performed by Sonia Gowan (ESR, New Zealand). Fourteen selected isolates were serotyped by Abbie Weisner (PHLS, United Kingdom).
- Ureaplasma and mycoplasma reference strain culture was partially performed by Susannah Gordon (CIDM, Westmead Hospital).
- Gregory James (CIDM, Westmead Hospital) provided all the needed molecular equipment, reagents, softwares and some protocols.
- Zhenfang Ma assisted in some ureaplasma PCR and data analysis.
- Mark Wheeler (WMI, Westmead Hospital) performed sequencing.
- Professor Lyn Gilbert (supervisor) supervised the overall project. Assisted in research design, in the analysis and interpretation of data, and made a significant contribution to the project.

Preface

Someone told me "a PhD thesis is just like a **story**". From my viewpoint, the **story** formed the major part of my past six-year life, and they will significantly affect my future research career and even my whole life. I tried to make the story something that would be interesting and useful to be read. On the other hand, I tried to cut its length to save the possible readers precious time.

As a PhD student, besides to **master** or to learn the techniques and methodologies as a **Master** student or a **technician**, to think about the **philosophy** or to explain the findings (including that of the others besides myself) would be also or even more important for the future related research fields. You would find the thesis in some extent (in particular, in section 3) reflected the above considerations.

From the teacher side, the selection of **good** teachers would be very important for a **good** student. I was so lucky to be a student of Professors Lyn Gilbert and Tania Sorrell. I also want to thank so many "informal teachers", I may only know many of them from their excellent publications (I tried my best to include their contributions in my reference sections though I might still have lost many for cutting the volume reason), and these teachers are also highly appreciated. Molecular lab – I located myself there in my past six years, all the staffs there were so kind and the conditions there were so good – the experience there will be kept in my mind as a very memorable life in my personal history.

The two **tiny** microorganisms that I selected, as my study objectives – GBS and ureaplasmas – were also my very **great** teachers! They showed me that they also had some kind of "wisdom", which I tried to expose partial in the thesis (especially in section 3) but far from enough. No doubt, their "wisdom" deserves to be further studied – especially based on their invaluable genome resources and benefit from the new bioinformatic concepts, theory and research tools.

The thesis contains three sections. The first section of general introduction provided the background and some useful concepts of the next two sections. The second section mainly contained those so called **“traditional”** molecular microbiology study of the two microorganisms. It included mainly the genotyping studies, which included most of my previous publications (also see appendices). My past six year publications (in PDF format) had been used as appendices and were put at the end of the thesis. Their inner relationship was also given at the beginning of the appendices. So only selected parts (after modification and reorganization) that were believed to be important to explain the outline of the **story** were put into the body of the thesis (mainly in section 2). The third section of the thesis contained some so called **“novel”** genomic and bioinformatics based study of the two organisms. Because we are facing the dramatic paradigm shift in their post genomic era, the ideas contained in the study may not be absolutely correct; even for the correct ones, some of them may still look a little bit **ugly**. But at least I would like to use them to show the potential values for doing this kind of **“novel”** study.

It was not easy to be a PhD student, especially considering the era of knowledge explosion, and the so many paradigm shifts. In this case, I really feel that what I had done was far from enough. However, to get the PhD or to pass the **milestone** is my long-term dream (even from my early childhood). So could I dare to use this thesis to have a **try**?

If the thesis can be seen as a **very tiny** drop of but **useful** water comparing with the endless sea of the **truth**, I will be more than happy**!**

Acknowledgments

The work described in this thesis was completed at:

Molecular Laboratory, Reference Laboratory Section, Centre for Infectious Diseases and Microbiology (CIDM), Institute of Clinical Pathology and Medical Research (ICPMR), Westmead Hospital, Westmead, NSW 2145, Australia. All the staffs in Molecular Laboratory and Reference Laboratory Section are highly appreciated.

The project was partially supported by Australian Postgraduate Awards 2000 (APA 2000) and Westmead Millennium Institute (WMI) Initiating Grant.

This thesis couldn't become true without:

Professor Lyn Gilbert, a great and very knowledgeable person, my highly respected supervisor – I was so lucky to be her student. Her excellent supervision, patient editing, and many other supports made my past six years productive and enjoyable. Professor Tania Sorrell, a very kind person, my associate supervisor – she always prepared to give her kind help to me when I needed.

I thank Mark Wheeler for his precious help in sequencing and Glenys Connor for some GBS data collection. Dr Hui Wang, in the third year of my PhD, while I was training him, gave me a great deal of help, was much appreciated.

I would like to thank all my family members; their continuous encouragements and supports made me feel full of energy. Zhenfang Ma, my deeply beloved wife, my another half, the most valuable **gift** I could get from my God – she gave me the greatest and invaluable supports, any words can't express my thanks to her – I **love** you! At the final stage of my PhD thesis writing, my lovely son (Alan, Xiangyu Kong) joined with us – welcome you my little honey and hope you can make my future **story** more exciting!

Abstract

Group B streptococcus (GBS, *S. agalactiae*) and human ureaplasmas (*U. parvum* and *U. urealyticum*) are two clinically and phylogenetically related, potential perinatal pathogens. Their relationships between genotypes and pathogenesis of GBS and ureaplasma infection were still not well understood, one of the reason is that both of them are still short of a very practical genotyping system. In the study, to solve the above problem we developed genotyping systems for the organisms (the second section). For human ureaplasmas, based on four genes/gene clusters (rRNA gene clusters, the elongation factor Tu genes, urease gene complexes and multiple banded antigen genes), we designed many primer pairs suitable for developing species identification assays for the two newly established human ureaplasma species (*U. parvum* and *U. urealyticum*). Further, based on the heterogeneity of ureaplasma multiple banded antigen gene (which contains species- and serovar-specific regions), we developed genotyping methods for each ureaplasma species. For GBS, based on three sets of molecular markers (capsular polysaccharide synthesis gene clusters, surface protein antigen genes and mobile genetic elements), we developed a genotyping system. The primary evaluation of the genotyping systems showed that the genotyping systems were practical alternative assays for the conventional serotyping and they will be useful to further explore the relationships between genotypes and pathogenesis of GBS and ureaplasma infection. In the study, we introduced novel data and tools into GBS and ureaplasma studies especially from genomic- and bioinformatics-based molecular microbiology

(the third section). For two newly established human ureaplasma species, based on the *U. parvum* serovar-3 genome, and using the above four important genes/gene clusters, we exposed some interesting problems in the understanding of new ureaplasma taxonomy especially in the post genomic era. For GBS, we studied the two published full genomes and exposed some new problems or possible future new research fields. In particular we found the two finished and one ongoing GBS genomes were all non-typical and suggest that future genomic project had better have genetic population structure viewpoint. Finally, we suggested that integrated studies of the two potential or conditional perinatal pathogens, from the viewpoint of evolution, would provide a new understanding angle of the pathogenesis of the two organisms. Studies suggested that during coevolution, human ureaplasmas (especially *U. parvum*) became friendlier than their ancestors to their human host (by losing most of its virulence genes); however, GBS tried to increase its invasive abilities (by getting more virulence genes) to fight against the human host attack.

Table of Contents

TITLE PAGE……………………………………………………………………………………….…I

DECLARATION..…………………………………………………...………………………..II

PREFACE…………………………………………………………………...………….……III

ACKNOWLEDGMENTS……………………………………………………………….……V

ABSTRACT…………………………………………………………………………………….VI

TABLE OF CONTENTS……………………………..……………………………………VIII

LIST OF FIGURES..………………………………………………………………….………X

LIST OF TABLES……………………...……………………………………………………XI

ABBREVIATIONS...………………………………………………………………………XIII

SECTION 1 GENERAL INTRODUCTION……………………………………………………….1

A. BRIEF INTRODUCTION………….……………………………………………………………1

B. WHY GBS AND UREAPLASMAS?…………………………..……………………………...2

B.1. GBS AND UREAPLASMAS…………………………………………………………………2

B.2. CLINICAL ASPECTS……………………………………………………………………...2

B.3. VIRULENCE AND PATHOGENESIS…………………………………………………….9

B.4. THE ROLE OF ANTIBODY……………………………………………………………...12

B.5. PREVENTION EFFORTS…………………………………………………...………14

C. WHY GENOTYPING?…………………..…………………………………………...…15

C.1. SEVERAL CONCEPTS OR GLOSSARY…………………………………………………15

C.2. GENOTYPING OF GBS AND UREAPLASMAS………………………………………16

C.3. MOLECULAR EPIDEMIOLOGY OF GSB AND UREAPLASMAS……………………..18

C.4. INTEGRATED STUDY OF INFECTION, POPULATION GENETICS AND EVOLUTION……………………………………………………………………………………...19

D. WHY GENOMICS AND COMPARATIVE GENOMICS?………………………………..21

D.1. SOME USEFUL CONCEPTS OR GLOSSARY……………………………………………22

D.2. UREAPLASMA AND GBS GENOMIC STUDIES……………………………………...26

E. THESIS ARRANGEMENT……………………………………………………………………..31

F. HYPOTHESIS TO BE TESTED BY THE STUDY………………………………………….31

SECTION 2 GENOTYPING SYSTEMS…………………………………………………………33

UREAPLASMAS…………………………………………………………………………………33

CHAPTER 1 *UREAPLASMA* SPECIES IDENTIFICATION……………………………….33

CHAPTER 2 *UREAPLASMA* GENOTYPING………………………………………………44

GBS……………………………………………………………………………………………65

CHAPTER 3 GBS MOLECULAR SEROTYPE IDENTIFICATION………………………65

CHAPTER 4 GBS PROTEIN GENE PROFILES………………………………………...…91

CHAPTER 5 GBS MOBILE GENETIC ELEMENTS…………………………………..…113

SECTION 3 GENOMIC- AND BIOINFORMATICS-BASED STUDIES………..……...…134

CHAPTER 6 POSTGENOMIC TAXONOMY OF HUMAN *UREAPLASMAS*………...…134

CHAPTER 7 *IN SILICO* GBS COMPARATIVE GENOMICS……………………………152

CHAPTER 8 GBS AND UREAPLASMA EVOLUTIONARY “WISDOM”……………..179

REFERENCES……..……………………………………………………………………...202

APPENDICES…………………………………..……………………………………………225

List of Figures

Figure 2.1. Multiple sequence alignment of the 5'-end and upstream of *mba*……………………….**54**

Figure 3.1. Molecular serotype identification based on the sequence heterogeneity of the 790-bp fragment at the 3'-end of *cpsE*-*cpsF*-and the 5'-end of *cpsG* ……………………………………….**80**

Figure 5.1. Genetic relationship of 194 invasive Australasian GBS strains (or 56 genotypes) …………………………………………………………………………………………..**124**

Figure 8.1. The branching order of the main groups within *Bacteria* based on conserved indels present in various proteins………………………………………………………………………….**187**

Figure 8.2. Supertrees of 45 species constructed with 730 trees…………………………………...**189**

List of Tables

Table 1.1. Primers targeting ureaplasma 3 different genes/regions……………………………….…**38**

Table 2.1. Primers for ureaplasma genotyping……………………………………………………**48**

Table 2.2. Specificity and expected lengths of ureaplasma genotyping primer pairs………………**60**

TABLE 3.1. REFERENCE STRAINS USED IN GBS MOLECULAR SEROTYPING STUDY……………………**70**

Table 3.2. Primers used in GBS molecular serotyping study……………………………………..**72**

Table 3.3. The heterogeneity of 8 GBS serotypes in the regions of the 3'-end of *cpsD* and the 5'-end of *cpsE*…………………………………………………………………………………………..**76**

Table 3.4. Comparison of the results of conventional serotyping (CS) and molecular serotype identification (MS)/subtyping of 206 clinical GBS isolates………………………………………**85**

Table 3.5. Algorithm for GBS molecular serotype (MS) identification by PCR and sequencing…..**87**

Table 4.1. Primers used in GBS protein gene profiling study………………………………………**96**

Table 4.2. Specificity and expected lengths of amplicons of GBS protein gene profiling primer pairs ……………………………………………………………………….………………**100**

Table 4.3. Genetic groups and subgroups of GBS *bac* based on amplicon length and sequence heterogeneity……………………………………………………………………………….….**104**

Table 4.4. The relationship between GBS protein gene profiles and capsular polysaccharide synthesis (*cps*) gene molecular sero/subtypes……………………………………………………**108**

Table 5.1. Primers used in GBS mobile genetic element study……………………………………**117**

Table 5.2. Distribution of mobile genetic elements among 194 invasive GBS isolates……………**123**

Table 5.3. Relationship between genotypes of invasive GBS isolates and patients' age-groups ……………………………………….…………………………………….……………**127**

Table 6.1. Primers used for sequencing 3 ureaplasma different genes/gene clusters………………**139**

Table 6.2. Comparison of inter-species, intra-species and inter-copy heterogeneity in rRNA gene complexes of *U. parvum* and *U. urealyticum*..……………………………………………………**146**

Table 6.3. Comparison of interspecies heterogeneity of DNA and amino acid sequences of the urease gene clusters of *U. parvum* and *U. urealyticum*…………………………………………………**149**

Table 7.1. General features of NEM316 and 2603 V/R genomes…………………………………..**157**

Table 7.2. *Sma*I restriction map of NEM316 and 2603 V/R genomes**.**……………………...……**160**

Table 7.3. NEM316 and 2603 V/R selected "virulence"-related proteins**.**…………………………**167**

Table 7.4. The cell wall protein heterogeneity of NEM316 and 2603 V/R……………………...…**169**

Table 7.5. Possible genomic islands in NEM316 and 2603 V/R genomes…………………………**175**

Table 8.1. General features of GBS and ureaplasma genomes.……………………………………**183**

Table 8.2. Some similar charizations of GBS, ureaplasma related organisms that support the concept of reductive evolution from *Streptococcus* to *Mollicutes*.……………………………………………..**185**

Table 8.3. Sequence comparison of human ureaplasma and GBS major surface protein antigens. ……………………………………..…………………………………………..……………**197**

Abbreviations

ABI	Applied Biosystems
Alp2	C alpha-like 2 antigen
alp2	C alpha-like 2 antigen gene
Alp3	C alpha-like 3 antigen
alp3	C alpha-like 3 antigen gene
Alp4	C alpha-like 4 antigen
alp4	C alpha-like 4 antigen gene
Alp5	C alpha-like 4 antigen
alp5	C alpha-like 4 antigen gene
AmiC	Amidase family protein
ANGIS	Australian National Genomic Information Service
ATCC	American Type Culture Collection (Manassas, VA, USA)
Bac	C beta antigen or IgA binding protein
bac	C beta antigen or IgA binding protein gene
Bca	C alpha antigen
bca	C alpha antigen gene
BPS	group B protective surface protein
bp	base pair
BspA	A cell surface associated leucine-rich repeat protein involved in adhesion to fibronectin and fibrinogen
CAMP	(discovered by) Christie, Atkins, and Munch-Petersen
CDC	Centers for Disease Control and Prevention
CDS	Coding sequences
CGH	Comparative genome hybridization
CI	Confidence interval
CLD	Chronic lung disease (of prematurity)
Clp	Clp ATPase family of molecular chaperones
CNS	Central nervus system
CbpD	Choline binding protein D
CpdB	Cyclo-nucleotide phosphodiesterase
CPS	Capsular polysaccharide
cps	Capsular polysaccharide synthesis (gene cluster)
CS	Conventional serotyping/serotype
CSF	Cerebrospinal fluid
DNA	Deoxyribonucleic acid

dNTP	Deoxynucleotide triphosphate
EaeH	EaeH of *Escherichia coli* O157:H7
EF-TU	Elongation factor Tu
EOD	Early onset disease (infection occurring within the first week of life)
FbsA	A fibrinogen receptor from group B streptococcus
G+C	Guanine plus Cytosine
GBS	Group B streptococcus or group B streptococci or *S. agalactiae*
g	Gram
GIs	genomic islands

Human ureaplasmas – *U. parvum* and *U. urealyticum*

h	Hour
Hsa	(antigen that recognition of) sialic acid-containing host receptors
IAP	Intrapartum antibiotic prophylaxis
Indels	Insertations and deletions
IS	Insertion sequence(s)
Lmb	Laminin-binding protein
lmb	Laminin-binding protein gene
LOD	Late onset disease (infection occurring between 1 week and 2 to 3 months of age)
LSU rRNA	Large submit rRNA
l	Liter/Litre
M	Molar
MBA	Multiple banded antigen
mba	Multiple banded antigen gene
mge	Mobile genetic element(s)
min	Minute
ml	Microlitre/Microliter
MLEE	Mutiple locus enzyme electrophoresis
MLST	Multiple locus sequence typing
MS	Molecular serotype(s)
NanA	Sialic acid lyase (catalyzes the hydrolysis of sialic acid into pyruvate and N-acetylmannosamine)
NCTC	National Culture Type Collection (Colindale, UK)
No.	Number(s)
NGU	Non-gonococcal urethritis
nra	Encode a response regulator (no response to atmospheric conditions)
OR	Odds ratio
ORF	Open reading frame(s)

Pas The surface protein antigen I/II of *Streptococcus intermedius*

PCR Polymerase chain reaction

PFBP *Streptococcus pyogenes* fibronectin-binding protein

PFGE Pulsed-field gel electrophoresis

pgp Protein gene profile(s)

PIs Pathogenicity islands

PrtS Serine proteinase, subtilase family

PspC Pneumococcal surface protein C

PulA Alkaline amylopullulanase

R28 *Streptococcus pyogenes* surface protein R28

R5 (or BPS protein) – group B protective surface protein

RAPD Random amplified polymorphic DNA

RFLP Restriction length fragment polymorphisms

rgg Encode a response regulator

Rib Rib antigen

rib Rib antigen gene

RNA Ribonucleic acid

rof Encode a response regulator

s Second

ScpB Serine protease and C5a peptidase

Sec10 Surface exclusion protein

SpaA Streptococcal protein antigen A of *Streptococcus sobrinus*

Sip Surface immunogenic protein

SrpA Periplasmic linker protein

Srt Sortase

SSH Suppressive subtractive hybridization

Ssp5 Agglutinin receptor

sst Serosubtype(s)

SSU rRNA Small subunit rRNA

***T*a** Annealing temperature

***T*m** Melting temperature

tuf Elongation factor Tu gene

U Unit

UAB University of Alabama at Birmingham

UP *Ureaplasma parvum*

Ureaplasmas – *U. parvum* and *U. urealyticum*

UU *Ureaplasma urealyticum*

Ureaplasma broth – 10B broth (Shepard, 1970)

WebANGIS – Website of Australian National Genomic Information Service (**http://www1.angis.org.au/pbin/WebANGIS/wrapper.pl**)

YbgE Putative branched-chain aminotransferase

YfkN 2`,3`-cyclic-nucleotide 2`-phosphodiesterase

GENERAL INTRODUCTION

"Pathogenesis and symbiosis are relatively rare among bacterial species." (Ochman & Moran, 2001).

"Although we will never conquer infectious diseases, we can certainly learn to live in greater harmony with them. This may perhaps be our most effective intervention strategy. Discovering how to do so will be the great challenge for molecular epidemiologists of the future." (Foxman & Riley, 2001).

A. BRIEF INTRODUCTION

In microbiology, we are in the era of and witness to many paradigm shifts (Bull, *et al.,* 2000). Group B streptococcus (GBS, *Streptococcus agalactiae*) and human ureaplasmas (*Ureaplasma parvum* and *U. urealyticum*) – are clinically and phylogenetically related potential urogenital tract and perinatal pathogens (Razin *et al*., 1998; Robertson *et al*., 2002; Schuchat, 1998). They were selected as examples to demonstrate the paradigm shifts – from traditional to molecular microbiology, then to genomic and bioinformatics-based microbiology (Glass *et al*., 2000; Glaser *et al*., 2002; Tettelin *et al*., 2002). In the post-genomic era, comparative, combinatorial and integrated studies, and bioinformatics have become popular strategies, because they can solve problems and generate new knowledge more efficiently than the other stratgies (Dobrindt & Hacker, 2001). When we look back at the history of microbiology, it is not difficult to find that significant progress in one defined specific area (for example, study of one bacterial species or even one test method, etc.) can be the basis for studies in related areas or even the whole microbiological field. For example, *S. pneumoniae* research has contributed significantly to studies of GBS serotyping, treatment, and prevention (Obaro & Adegbola, 2002). The aim of this thesis is to determine whether an integrated study

of GBS and human ureaplasma will reveal aspects of both that may not be apparent from studying them separately.

B. WHY GBS AND UREAPLASMAS?

B1. GBS and ureaplasmas.

GBS and ureaplasmas are among the most frequent bacteria associated with urogenital tract infection in women (Horner *et al*., 2003) and perinatal diseases (Goncalves *et al*., 2002; Razin *et al*., 1998; Schuchat, 1998). Both belong to the low G+C group of bacteria – 35.6% for *S. agalactiae* and 25.5% for human ureaplasmas (Glass *et al*., 2000; Glaser *et al*., 2002; Tettelin *et al*., 2002).

S. agalactiae is the only species of group B streptococcus (GBS). GBS is a facultative Gram-positive diplococcus and the leading cause of serious bacterial infections (bacteremia, pneumonia, and meningitis) in newborns (Chaffin *et al*., 2000; Manning, 2003). An indispensable GBS virulence determinant is its serotype-specific capsular polysaccharide (CPS), which prevents the deposition of host complement factor C3b and inhibits opsonophagocytosis (Chaffin *et al*., 2000).

Human ureaplasmas include two species: *U. parvum* and *U. urealyticum* and belong to the class *Mollicutes*. Mollicutes are the smallest free-living organisms, which lack cell walls, pass through "bacteria-retaining" filters, require cholesterol for membrane function and growth and use UGA codon for tryptophan. Their feature of without cell wall may be largely responsible for their biologic properties including lack of a Gram stain reaction and nonsusceptibility to many commonly prescribed antimicrobial agents, including beta lactams (Razin *et al*., 1998). Ureaplasmas have a requirement, unique among mollicutes, for urea. ATP is generated through an electrochemical gradient produced by ammonia liberated during the intracellular hydrolysis of urea by urease (Blanchard & Barile, 1989).

B2. Clinical aspects.

B2.1. GBS.

GBS was originally known as cause of bovine mastitis and was not shown to be a human pathogen until 1938 (Manning, 2003). Currently, GBS remains a common cause of neonatal disease despite prevention efforts (Schuchat, 1998). Clinical syndromes of GBS disease in newborns include sepsis, meningitis, pneumonia, cellulitis, osteomyelitis, and septic arthritis (Schuchat, 1998). In addition to acute illness, which is itself costly, GBS infections in newborns can result in death and disability like cerebral palsy (Schuchat, 1998; Russell & Buttery, 2003).

B2.1.1. Early-onset disease (EOD).

Approximately 80% of infant GBS infections occur in the first few days (1-6) of life, so-called EOD. Newborns with EOD acquire the organism, before or during delivery, from their mothers who are colonized with GBS in the genital tract. Most EOD results from ascending spread of the organism into the amniotic fluid, where aspiration of contaminated amniotic fluid leads to invasive disease in some infants (Schuchat, 1998). Clinically, infants with EOD are often premature and prone to develop sepsis, respiratory distress and leukopenia (Ho *et al*., 1999).

B2.1.2. Late-onset disease (LOD).

GBS LOD occurs in infants between 1 week and 2 to 3 months of age. The pathogenesis of LOD is less well understood, although some cases probably reflect acquisition of the organism during passage through the birth canal (Schuchat, 1998). Clinically, LOD often presents with meningitis and sepsis (Ho *et al*., 1999).

B2.1.3. GBS infections among pregnant women.

Among pregnant women, GBS can cause clinical illness ranging from mild urinary tract infection to life-threatening sepsis and meningitis; chorioamnionitis, endometritis and septic abortion also occur (Manning, 2003). Most invasive

maternal infections are bloodstream infections, and osteomyelitis; endocarditis, and meningitis have also been described (Schuchat, 1998).

B2.1.4. GBS infections among elderly adults.

Recent studies have shown that a substantial burden of GBS-related illness occurs outside the high-risk periods – pregnancy and early infancy. The clinical presentations of invasive GBS diseases among non-pregnant adults (especially in the elderly people, and with underlying medical conditions) most often take the form of primary bacteremia, and skin, soft tissue or bone infections (Schuchat, 1998).

B2.1.5. GBS infection incidence trends.

The introduction of intrapartum antibiotic prophylaxis (IAP) has lowered the incidence of early-onset GBS sepsis by 50-80%, and the incidence among pregnant women also decreased (Moore *et al*., 2003; Schuchat, 1998). In contrast, the number of GBS-attributable conditions among non-pregnant adults, particularly the elderly with underlying conditions, is increasing, ranging from 4.1 to 7.2 cases per 100,000 adults over the past decade (Manning, 2003; Schuchat, 1998).

B2.1.6. GBS colonization.

Many individuals are asymptomatically colonized with GBS. Depending on the characteristics of the study population and the detection methods utilized, 10 to 40% of individuals are colonized. Newborns over 48 hours old are most commonly colonized in the throat and rectum and can remain colonized throughout childhood. At the onset of sexual activity, colonization generally shifts to the genitourinary tract (Manning, 2003). Only a small proportion of people colonized with GBS suffer any ill effects (Feikin *et al*., 2001; McDonald & Chambers, 2000).

B2.2. Ureaplasmas.

Like GBS, human ureaplasmas are also associated with a variety of urogenital tract diseases, although they cause adverse effects in an even smaller proportion of those colonized (Razin *et al*., 1998; Horner *et al*., 2003).

B2.2.1. Adverse pregnancy outcomes.

Adverse pregnancy outcomes, including maternal complications, chorioamnionitis, preterm birth and respiratory complications of the preterm neonate have been attributed to ureaplasmas (Abele-Horn *et al*., 2000). Ureaplasmas have been isolated from the blood of women with postpartum or postabortal fever but not from afebrile women following abortion or from healthy pregnant women (Taylor-Robinson & Furr, 1997; Razin *et al*., 1998). Ureaplasmas have been isolated from maternal blood, umbilical cord blood, and the blood and cerebrospinal fluid (CSF) of neonates (Taylor-Robinson & Furr, 1997; Razin *et al*., 1998). But some others study showed that (1) infection of the cerebrospinal fluid by ureaplasma is infrequent, (2) ureaplasma organisms are frequently present in tracheal aspirate specimens but do not appear to be related to the presence or the subsequent development of respiratory disease, and (3) initiation of erythromycin treatment at 1 to 3 weeks of age does not alter the clinical course (Heggie *et al*., 1994).

B2.2.2. Infants ureaplasma infection.

Colonization of infants by ureaplasmas may occur by ascent of the microorganisms from the lower genital tract of the mother at the time of delivery or by direct invasion of the fetus in utero. The relationship between acute and chronic lung disease (CLD) of prematurity and ureaplasmas is still controversial (Abele-Horn *et al*., 1998; Castro-Alcaraz *et al*., 2002; Hannaford *et al*., 1999; Heggie *et al*., 2001). Colonization of amniotic fluid with ureaplasmas may be associated with subsequent meningitis and other central nervous system (CNS) complications (Taylor-Robinson

& Furr, 1997; Razin *et al*., 1998).

B2.2.3. Ureaplasma related other infections.

Ureaplasmas usually reside extracellularly in the respiratory and urogenital tracts and rarely penetrate the submucosa, except in association with immunosuppression or instrumentation, when they may invade the bloodstream and disseminate to numerous organs and tissues (Razin *et al*., 1998). Ureaplasmas are the most common "typical" bacteria that cause "culture-negative" infectious arthritis in persons with hypogammaglobulinemia (Taylor-Robinson & Furr, 1997; Razin *et al*., 1998). Ureaplasmas sometimes can be detected by culture or PCR in the synovial fluid of persons with rheumatoid arthritis but their precise contribution to this disease is uncertain (Taylor-Robinson & Furr, 1997).

B2.2.4. Ureaplasma morbidity and mortality.

Assessing morbidity and mortality of diseases specifically caused by ureaplasmas is difficult because few studies systematically evaluate them, and some conditions in which they are implicated are polymicrobial (Taylor-Robinson & Furr, 1997; Razin *et al*., 1998). Deaths occurr in some neonates with bloodstream invasion and meningitis due to ureaplasmas; but, in others, the organisms spontaneously disappear from cerebrospinal fluid (CSF) without treatment (Taylor-Robinson & Furr, 1997; Razin *et al*., 1998).

B2.2.5. Ureaplasma colonization.

Human ureaplasmas can establish colonization in utero or at birth; they have been found on the skin and in the oropharynx, lung, and lower genital tract of many newborns. The persistence of ureaplasma in the oropharynx and upper respiratory tract and in the male genital tract usually declines after birth, but as many as 20% of prepubertal girls harbor these organisms in the lower genital tract (Taylor-Robinson

& Furr, 1997; Razin *et al*., 1998). Human ureaplasmas are the most frequent bacteria isolated from the lower and upper genital tracts of women, especially pregnant women (Knox *et al*., 1997; Razin *et al*., 1998). Ureaplasmas are found in the cervix or vagina of 40-80% of asymptomatic, sexually active women (Cunningham *et al*., 1996) and in the urethras of 5-20% of asymptomatic, sexually active men (Razin *et al*., 1998; Taylor-Robinson, 1986).

The gastrointestinal tract may also be a potential reservoir: 9-18% of infants had rectal colonization with ureaplasmas (Sanchez & Regan, 1987; Sanchez & Regan, 1990); 57% of women attending a venereal disease clinic had positive anal or rectal swabs for ureaplasmas (Munday *et al*., 1981).

B2.3. Similarities and differences between GBS and ureaplasmas.

B2.3.1. Similarities.

- ♦ Both GBS and human ureaplasmas are common perinatal, urogenital, and opportunistic pathogens;
- ♦ Both GBS and human ureaplasmas have high “colonization” but low “infection” rates.

B2.3.2. Differences.

- ♦ GBS is more virulent than ureaplasmas – GBS causes more serious morbidity and a higher mortality; its pathogenic role is much more clearly defined than that of ureaplasma. Studies of ureaplasma pathogenesis are often controversial.
- ♦ Both organisms have significant age/time-patterns. The incidence of ureaplasma neonatal colonization and infection is inversely related to gestational age or chronological age and is rare in full term infants. GBS infections are more common in preterm infants but also occur at term and in the first few months of life. GBS also causes significant infections in elderly adult; whereas

ureaplasmas do not.

- Ureaplasma colonization rates (20-80%) are higher than GBS (15-40%); ureaplasma colonization rates in pregnant women (40-80%) are higher than in non-pregnant women (~20%) because of the possible effects of oestrogens (Reid *et al.*, 1993). GBS colonization rates are similar between pregnant and non-pregnant women (Hoshina *et al.*, 1991; Manning, 2003);

B2.4. Conditional pathogens.

According to traditional thinking, a pathogen was a microbe that was genetically endowed with a factor that, when expressed, caused disease. This postulate became central to the concept of the monomicrobial aetiology of infectious diseases (Larsen & Monif, 2001). However, the mere presence of potentially pathogenic species does not necessarily constitute disease when defined in terms of symptoms (Larsen & Monif, 2001).

"Conditional" pathogens could mean that only some strains are pathogenic or that virulence factors are only sometimes expressed. It also could mean that host susceptibility plays a role in disease/infection. We are confronted with having to explain why apparently commensal bacteria can cause disease (Larsen & Monif, 2001).

For GBS and ureaplasmas, their potential for causing disease does not guarantee that disease will occur with their presence (Larsen & Monif, 2001). In fact, they rarely do cause disease – most colonized people are healthy. Although GBS is a leading cause of perinatal and maternal postpartum septicaemia, the incidence of disease is grossly disproportional to that of colonization (Larsen & Monif, 2001). The high colonization rates and low morbidity/mortality rates of GBS and especially ureaplasmas suggest that they are not "pure" pathogens but conditional pathogens (Goncalves *et al.*, 2002; Razin *et al.*, 1998; Schuchat, 1998).

B2.4.1. Quantity and diseases.

B2.4.1.1. GBS.

There is evidence of an association between maternal vaginal GBS concentration and neonatal infections after birth, especially in women with premature ruptures of membranes (Itakura *et al*., 1996). Heavy GBS colonization at 23 to 26 weeks' gestation is associated with an increased risk of delivering a preterm, low-birth-weight infant but is not a reliable predictor of neonatal GBS sepsis (Regan *et al*., 1996).

B2.4.1.2. Ureaplasmas.

The degree of colonization with ureaplasma also possible correlates with adverse pregnancy outcomes, but two independent studies got controversial results (Abele-Horn *et al*., 2000; Heggie *et al*., 2001). But for the "conditional" nature of ureaplasmas, this point needs to be further studied before a definite conclusion.

B3. Virulence and pathogenesis.

B3.1. Difference in GBS and ureaplasma pathogenicity.

From the viewpoint of pathogenicity, GBS is more virulent than ureaplasmas; they can cause serious, invasive and potentially fatal infection (not only "morbidity" but also "mortality") (McDonald & Chambers, 2000; Schuchat, 1998). Of the two ureaplasma species, there are differences between *U. parvum* and *U. urealyticum* in distribution and even some possible pathogenicity in different populations (Deguchi *et al*., 2004; Domingues *et al*., 2002; Knox *et al*., 2003). For example, in young males with non-gonococcal urethritis (NGU) *U. urealyticum* is more common than *U. parvum* (Povlsen *et al*., 2002; Deguchi *et al*., 2004). Usually ureaplasma infection-related clinical presentations are more "benign" than GBS infection; and they are rarely directly implicated in patient death (only "morbidity" but no or very

rare “mortality”) (Razin *et al*., 1998). Future research is required to determine why some women develop ascending intrauterine infections with either GBS or ureaplasmas and others do not; further, what interventions may reduce the deleterious effect of systemic fetal inflammation (Finlay & Falkow, 1997; Goncalves *et al*., 2002).

B3.2. Common themes of GBS and ureaplasma pathogenesis.

Despite these differences in pathogenicity, GBS and ureaplasmas share many common themes in their pathogenesis.

B3.2.1. Adhesins.

There are only a few common types of adhesins – the first key step in causing disease (Baseman *et al*., 1996) – used by pathogens to adhere to host substrates. These include:

- **Major surface protein antigens:** The GBS surface protein family is defined by C alpha and Rib (Stalhammar-Carlemalm *et al*., 1999; Lachenauer *et al*., 2000). Ureaplasma multiple banded antigen (MBA) may also be related to adhesins (Smith *et al*., 1994; Zheng *et al*., 1995).
- **Other proteins:** GBS laminin binding protein (Lmb) mediates GBS attachment to human laminin (Spellerberg *et al*., 1999).

B3.2.2. Other virulence factors.

Several other common themes that are employed to avoid host immune defences:

- **Resistance to opsonophagocytosis** – GBS capsular polysaccharide and proteins (Harris *et al*., 2003),
- **Cleavage of human fibrinogen or host immune components** – GBS cell-surface-associated protein [*cspA*]) (Harris *et al*., 2003) and ureaplasma IgA1 protease (Robertson *et al*., 1984).

- **Antigenic variation –** both GBS and human ureaplasma major surface antigens (Zheng *et al*., 1994; Lachenauer *et al*., 2000),
- **Camouflage –** GBS by binding host molecules (for example, GBS IgA binding protein, C beta) (Heden *et al*., 1991) and ureaplasma MBA by possible mimicry (Ang *et al*., 2002; Jacobs *et al*., 1995).

B3.2.3. Mobile genetic elements (mge).

Virulence factors may be associated with mge or located within pathogenicity islands (PIs), which ensures that new strains of pathogens evolve constantly. Two GBS genomic studies have supported this theory (Glaser *et al*., 2002; Tettelin *et al*., 2002). Analysis of ureaplasma *tetM* and Tn*1545*-like (Tn*916*) transposon resistance determinants also support this point (de Barbeyrac *et al*., 1996; Taraskina *et al*., 2002).

B3.3. Capsular polysaccharide.

GBS capsular polysaccharide (CPS) is recognized as a key virulence factor of GBS (Schuchat, 1998), and its synthesis is determined by the capsular polysaccharide synthesis (*cps*) gene cluster (Chaffin *et al*., 2000). Ureaplasmas also have capsule-like polysaccharide structures that contain glucosyl-like resides (Whitescarver & Furness, 1975; Robertson & Smook, 1976).

B3.4. Proteins.

B3.4.1. GBS:

a) **GBS surface variable protein antigen family:** Bca, Rib, Alp2, Alp3 (Gravekamp *et al*., 1998; Lachenauer *et al*., 2000) and R1-R4 (Moyo *et al*., 2001a, b; Moyo & Maeland, 2003) are related to GBS virulence and are targets for serosubtyping.

b) **GBS C beta protein (Bac):** is a human IgA binding protein (Berner *et al*., 2002).

c) **Other proteins**. Many other proteins can contribute to GBS pathogenesis, which was described in two recent reviews (Manning, 2003; Spellerberg, 2000). For example, a cell surface protein designated Fbs, a R-like protein named BPS (group B protective surface protein) (Erdogan *et al*., 2002), a glutamine synthetase, the heat shock protein (Hsp70), and surface immunogenic protein (Sip) (Brodeur *et al*., 2000), hyaluronidase/hyaluronate lysase (gene *hylB*) (Granlund *et al*., 1998; Bohnsack *et al*., 2001), C5a-peptidase (ScpB) (Hauge *et al*., 1996; Adderson *et al*., 2000), GBS CAMP-factor (encoded by *cfb*) (Jurgens *et al*., 1987; Podbielski *et al*., 1994) and haemolysin/cytolysin (probable structural gene, *cylE*) (Pritzlaff *et al*., 2001) are all identified as probable virulence factors.

B3.4.2. Ureaplasmas.

B3.4.2.1. Multiple banded antigen (MBA).

MBA exposed on the cell surface are the major antigenic determinants in ureaplasmas. Antisera containing antibodies to these components inhibit growth and metabolism of ureaplasmas and, in the presence of complement, cause lysis of the organisms. These properties are used in various serologic tests that differentiate between ureaplasma species and serovars and detect antibodies to ureaplasmas in sera of patients (Razin *et al*., 1998).

B3.4.2.1. Human immuoglobulin A1 protease.

The 14 recognised human ureaplasma serovars express human immunoglobulin A1 (IgA1) protease activity (Robertson *et al*., 1984). Extensive cleavage of IgA1 could result in a local functional IgA1 deficiency that may facilitate colonization (Kilian *et al*., 1996).

B4. The role of antibody.

B4.1. GBS – protection.

The fact that serotype-specific anticapsular antibody opsonizes homologous GBS strains suggested the protective role of anticapsular antibody (Baker & Kasper, 1976; Schuchat, 1998). Seroype-specific antibody to capsular polysaccharide is protective, but such antibodies are rare (Hoshina, 1997; Campbell *et al*., 2000; Mikamo *et al*., 2000). A study of 1207 pregnant women in Canada showed that low level of protective serotype-specific capsular polysaccharide IgG antibody is one major risk factor for disease – in particular serotype III GBS is more invasive than other serotypes in this population may be due, at least in part, to poor maternal serotype III CPS-specific antibody response (Basham *et al*., 1996; Davies *et al*., 2001). GBS surface protein antibodies also can protect the host from infection (Hoshina, 1997; Schuchat, 1998). In one study, elevated levels of urinary antibodies were associated with an increased risk of preterm delivery, which suggests that a mucosal inflammatory response to urogenital infection may stimulate the onset of preterm labour. The results suggest that, unlike circulating serum antibody, mucosal GBS antibody may serve as adverse markers for preterm labour (McKenzie *et al*., 1994).

B4.2. Ureaplasmas – adverse markers.

For ureaplasmas, one hypothesis is that mucosal immunity mediates resistance to infection and systemic immunity contributes substantially to the adverse effects (Cunningham *et al*., 1996; Razin *et al*., 1998), which is different from the role of GBS antibody (Davies *et al*., 2001; McKenzie *et al*., 1994). The presence of antibody may be an adverse marker in ureaplasma infection, suggesting possible autoimmune or immunopathogenic effects (Horowitz *et al*., 1995; Abu-Shakra *et*

al., 1999). In such cases, molecular mimicry may play some role (Baseman *et al*., 1996).

B5. Prevention efforts.

B5.1. GBS.

In order to decrease the morbidity and mortality associated with GBS disease in newborns, the Centres for Disease Control and Prevention (CDC) and other organizations developed guidelines to provide women at risk of delivering a GBS-diseased infant with intrapartum prophylaxis. Penicillin and ampicillin are the primary agents used. Clindamycin and erythromycin are recommended for use in penicillin-allergic women. The consensus prophylaxis strategies are associated with frequent antibiotic use – at least 24% of women receive antibiotics during labour and delivery (Schuchat, 1998). Since the development of the GBS prevention program, the incidence of GBS neonatal sepsis has decreased significantly (Manning, 2003).

Because antibiotic resistance is a major public health concern, and antimicrobial prophylaxis cannot prevent most LOD, GBS-related stillbirths, or prematurity and does not address GBS disease in nonpregnant adults (Schuchat, 1998), alternative prevention protocols must be evaluated and implemented. For GBS, an alternative to intrapartum prophylaxis is vaccine, which should significantly reduce the number of women receiving antibiotics (Manning, 2003). Multiple GBS serotypes can cause perinatal and adult diseases, and multivalent GBS vaccines are being developed. Excellent immune responses to polysaccharides conjugated to tetanus toxoid were reported. Though the initial target of GBS vaccine development was the use of vaccines in pregnant women, during the second trimester, GBS vaccines could potentially be given to all women of reproductive age before pregnancy occurs, since conjugate vaccines are likely to lead to long-term protection in adults (Schuchat, 1998).

B5.2. Ureaplasmas.

It is not clear that any prevention strategy is justified for ureaplasma infection, considering the much lower pathogenicity compared with GBS (Gelfand, 1993). Moreover, any prevention strategy would be a challenge, since both intrapartum prophylaxis antibiotics and vaccines are associated with problems. The lack of ureaplasma cell wall makes them nonsusceptibe to many commonly prescribed antimicrobial agents, for example penicillin and ampicillin (Razin *et al*., 1998). The possible adverse effects of circulating antibody (Horowitz *et al*., 1995) because of molecular mimicry (Baseman *et al*., 1996) makes vaccine also a poor option.

C. WHY GENOTYPING?

C1. Several concepts or glossary.

C1.1. Genotyping.

Genotyping of microorganisms is the means to discriminate between and catalogue microbial nucleic acid molecules (van Belkum *et al*., 2001). It is widely used in basic and applied microbiological research.

- In basic microbiological research, genotyping is used in the fields of taxonomy, phylogenetic relationships, evolutionary dynamics, population genetics, and epidemiology (van Belkum *et al*., 2001).
- In applied microbiological research, genotyping clearly affects several areas: the epidemiological investigation of outbreaks of infectious diseases and the measurement of genetic diversity in relation to relevant biological properties such as pathogenicity, drug resistance etc. (van Belkum *et al*., 2001).

Genotyping techniques are considered superior to phenotyping methods, which are often less reproducible and discriminatory (Manning, 2003).

C1.2. Molecular epidemiology.

The term "molecular epidemiology" involves both "molecular" – the use of the techniques of molecular biology – and the “epidemiology” – the study of the distribution and determinants of disease occurrence in human populations (Foxman & Riley, 2001).

Molecular epidemiology can be used to examine disease patterns, investigate outbreaks, describe transmission and population dynamics, identify risk factors, understand evolution and disease pathogenesis, monitor the efficacy of control programmes (Foxman & Riley, 2001; Manning, 2003).

C2. Genotyping of GBS and ureaplasmas.

C2.1. Serotypes of GBS and ureaplasmas.

Human ureaplasmas include two species: *U. parvum* and *U. urealyticum*; *U. parvum* contains 4 serovars (1, 3, 6 and 14) and *U. urealyticum* contains 10 serovars (2, 4, 5, 7-13) (Robertson & Stemke, 1982).

S. agalactiae is the only species of group B streptococcus (GBS). Nine distinct capsular serotypes, Ia, Ib, and II to VIII, have been identified (Chaffin *et al*., 2000). GBS also can be serotyped or serosubtyped by surface protein antigens: C alpha, C beta, R and X proteins (Ferrieri, 1988; Lachenauer *et al*., 2000).

C2.2. Serotyping targets of GBS and ureaplasmas.

C2.2.1. Capsular polysaccharide.

GBS capsular polysaccharide antigens have been studied intensively – they are GBS major serotyping target antigens. The capsular polysaccharide synthesis (*cps*) gene

clusters of six serotypes (Ia, Ib, II, III, IV, V, and VI) have been sequenced and analysed (Chaffin *et al*., 2000).

Ureaplasmas also have capsule-like structures (extramembranous carbohydrates) and contain glucosyl-like residues (Whitescarver & Furness, 1975; Robertson & Smook, 1976), but their encoding genes have not been studied as systemically as those of GBS (Chaffin *et al*., 2000; Glass *et al*., 2000).

C2.2.2. Protein antigens.

The major surface antigens of GBS are a family of variable proteins (Bca/Rib/Alp2/Alp3) (Lachenauer *et al*., 2000). All the family members contain N- and C-terminal conserved regions and the middle repetitive unit region (Lachenauer *et al*., 2000). The repetitive units occupy 60-80% of the gene length (Lachenauer *et al*., 2000). GBS C beta protein is an IgA binding protein, which also contains repetitive units (Berner *et al*., 2002). These GBS surface protein antigens are the basis for serosubtyping separate from, but related to CPS serotyping (Lachenauer *et al*., 2000).

Ureaplasma major surface proteins are the multiple banded antigens (MBA) , which contain both species and serotype definition sites (Zheng *et al*., 1995). Their gene repetitive units also occupy over 60% of the gene length (Zheng *et al*., 1996).

C2.3. Genotyping is required and important for GBS and ureaplasmas.

From the point of methodology, the traditional serotyping methods for both GBS and ureaplasmas are not very practical (Stemke & Robertson, 1985; Manning, 2003). The reason is that both of them are multiple serotype species and the antisera are expensive and not widely available (Stemke & Robertson, 1985; Manning, 2003).

The accurate genetic bases of GBS (capsular polysaccharide and protein antigen) (Chaffin *et al*., 2000; Lachenauer *et al*., 2000) and ureaplasma (MBA) serotypes (Zheng *et al*., 1995) are still not very clear.

The relationship between GBS and ureaplasma serotype and pathogenesis has not been defined partly because of the impractical nature of conventional serotyping (Manning, 2003; Stemke & Robertson, 1985) and partly because of the relatively poor discriminatory ability of conventional serotyping (Manning, 2003). A prerequisite for solving the problem is to develop practical species identification and typing (for serotype, serosubtype and genotype) assays (Abele-Horn *et al*., 1997; Manning, 2003).

C3. Molecular epidemiology of GBS and ureaplasmas.

Traditionally, serotyping has been used to study GBS and ureaplasmas (Robertson *et al*., 1986; Harrison *et al*., 1998). Among the nine distinct capsular polysaccharide serotypes (Ia, Ib, and II to VIII) (Chaffin *et al*., 2000), serotypes Ia, III, and V are currently the most common isolates from the United States and many other countries. Serotype III is the most prevalent serotype associated with neonatal disease and serotype V is an important emerging GBS serotype worldwide (Manning, 2003). There are four *U. parvum* and ten *U. urealyticum* serovars (Robertson *et al*., 1986). The distribution of species (previous biovars) and serovars differs in different populations, but generally the four serovars of *U. parvum* are more common than ten *U. urealyticum* serovars (Knox *et al*., 2003; Ren & Zhu, 2003).

The phenotypic serotyping methods have contributed to our understanding of GBS and ureaplasma infections, but they are limited since they do not reveal information regarding genetic identity (Bidet *et al*., 2003; Echahidi *et al*., 2002). Many investigators now use genotypic techniques alone or in conjunction with certain phenotypic techniques (Manning, 2003; Martinez *et al*., 2001). GBS and ureaplasma

studies demonstrate that more molecular epidemiological information is needed to understand the pathogenesis of disease and to design alternative prevention strategies (Bidet *et al*., 2003; Zheng *et al*., 1992). For instance, comparing strains that cause disease with common colonization strains is important to identify bacterial factors that may be important in disease pathogenesis (Lopardo *et al*., 2003; Povlsen *et al*., 2002).

C4. Integrated study of infection, population genetics and evolution.

C4.1. Population genetics.

The analysis of bacterial collections, which accurately represent the natural population, by genotyping provides data that can be used both to investigate the population structure of bacterial pathogens and for the molecular characterization of bacterial isolates. A spectrum of possible population structures exists – with few bacterial species occupying the extremes of highly clonal and completely non-clonal, most containing both clonal and non-clonal elements (Spratt & Maiden, 1999).

At present, two main kinds of population structure can be distinguished in natural microbial populations:

(a) Species that are not subdivided into discrete phylogenetic lineages (panmictic species or basically sexual species with occasional bouts of short-term clonality fall into this category);

(b) Species that are strongly subdivided by either cryptic speciation or clonal evolution.

Improvements in available statistical methods are required to refine these distinctions and to better quantify the actual impact of gene exchange in natural microbial populations. Moreover, a codified selection of markers with appropriate molecular clocks (in other words: adapted levels of resolution) is sorely needed to

answer distinct questions that address different scales of time and space: experimental, epidemic, and evolutionary. The problems raised by natural genetic diversity are very similar for all microbial species, in terms of both basic and applied science (Tibayrenc, 1996).

C4.2. Evolutionary genetics.

Evolutionary studies address how genetic changes are induced, persist, and become fixed. At least four fundamental mechanisms can give rise to variation used to define evolutionary genetics: mutation, hypermutation, genetic recombination, and selection (van Belkum *et al.*, 2001). Evolutionary genetics, apart from improving our basic knowledge of the taxonomy and evolution of microbes, can also greatly contribute to applied research in microbiology. Evolutionary genetics provides convenient guidelines for better interpreting genetic and molecular data dealing with microorganisms.

The three main potential applications of evolutionary genetics in microbiology are:
(a) Epidemiological follow-up (with the necessity of evaluating the stability of microbial genotypes over space and time);
(b) Taxonomy in the broad sense (better definition and sharper delimitation of presently described taxa, research into hidden genetic subdivisions);
(c) Evaluation of the impact of the genetic diversity of microbes on their relevant properties (pathogenicity, resistance to drugs, etc).

C4.3. Virulence clones.

A common theme demonstrated by these analyses is that distinct bacterial clones are responsible for disease outbreaks and increases in infection frequency. Unique combinations of virulence genes or alleles of virulence genes characterize many of these clones. Because substantial interclonal variance results in relative differences in virulence, molecular population genetic studies have led to the concept that the unit of bacterial pathogenicity is the clone or cell line. Continued new insights into

host parasite interactions at the molecular level will be achieved by combining clonal analysis of bacterial pathogens with large-scale comparative sequencing of virulence genes (Musser, 1996).

Understanding the population structure and evolution of pathogens is important (Spratt & Maiden, 1999; McDonald & Linde, 2002). In many bacterial infections there is a variety of disease manifestations and clinical outcomes. One factor that can contribute to differences in severity is the variation in virulence that exists among strains of a bacterial population. Genomic studies reveal that bacterial populations are comprised of isolates that show a surprisingly wide spectrum of genetic diversity at the DNA level (Whittam & Bumbaugh, 2002). It is clear that during bacterial evolution, gene acquisition and loss contributes substantially to the variation in virulence and to the overall genetic variation harboured in pathogen populations (Whittam & Bumbaugh, 2002). It is possible that the integrated study of GBS and ureaplasmas from the viewpoint of infection, population genetics and evolution could provide us some new clues for their further study (van Belkum *et al*., 2001).

Infectious microbial epidemiology, population genetics of microorganisms, and evolutionary dynamics all rely on genotyping data for discrimination between microbial genotypes, which made the integrated study of infection, population genetics and evolution of microorganisms possible (Tibayrenc, 1999; van Belkum, 2003). It is believed that these integrated studies will be widely used in future microbiological studies because of their efficiency in solving problems (Gurtler & Mayall, 2001; Wiedmann, 2003).

D. WHY GENOMICS AND COMPARATIVE GENOMICS?

D1. Some useful concepts or glossary.

D1.1. Genomic.

The capacity to sequence and assemble entire genomes of bacteria, pathogenic protozoans, and fungi in a rapid and cost-effective way has energized every aspect of microbial science (Fraser *et al*., 2000). Genomic technologies could be defined as those used to manipulate and analyse genomic information (Galas & McCormack, 2003). So far, more than 200 bacterial genomes have been fully sequenced, including most important human pathogenic species (http://www.tigr.org). The genomic era brings many advantages and new hopes for many research areas; however, the acquisition and analysis of genome sequence data is not an end in itself, but a starting point for generating testable hypotheses for future research (Wren, 2000).

D1.2. Comparative genomics.

Recently, the trend in the sequencing of closely related genomes, including the sequencing of more than one strain from a single pathogenic species means that the comparative genomics has truly come of age (Whittam & Bumbaugh, 2002; Wren, 2000). While analysis of a single genome provides tremendous biological insights into any given organism, comparative analysis of multiple genomes provides substantially more information (Fraser *et al*., 2000).

Comparative genomic studies of selected important human pathogens have contributed to our understanding of bacterial diversity, pathogenesis/virulence, evolution and the development of improved disease therapeutics (Fraser *et al*., 2000).

D1.3. Bioinformatics.

Bioinformatics is composed of many different interrelated scientific fields such as genomics, proteomics, and transcriptional profiling (Bull *et al*., 2000; Luscombe *et*

al., 2001). For bioinformatics, the search strategy is based upon data collection and storage, and mining of databases in order to generate knowledge (Bull *et al*., 2000). Bioinformatic databases include DNA (genes and genomes), RNA, and protein sequences, proteomes, macromolecular structures, chemical diversity, biotransformations, metabolic pathways (metabolomes), biodiversity, and systematics, etc. (Bull *et al*., 2000). Thus, innovative "experiments" can be made *in silico* rather than *in vivo* or *in vitro* and bioinformatic technology will accelerate the paradigm shift from bench-top to desk-top (Bull *et al*., 2000; Katoh, 2002). Bioinformatics is more successful than its competitors in solving problems in search and discovery (Bull *et al*., 2000).

D1.4. Genomic islands (GIs).

Many genomic islands have been discovered in a variety of pathogenic as well as non-pathogenic bacteria (Hentschel & Hacker, 2001). Depending on the functions which are encoded by genomic islands, these may also be called pathogenicity islands (PIs), symbiosis, metabolic or resistance islands (Osborn & Boltner, 2002). GIs are acquired by horizontal gene transfer and are now recognised as important contributors to bacterial adaptation and evolution (Hentschel & Hacker, 2001).

D1.4.1. GIs share a set of unifying but highly conserved features (Osborn & Boltner, 2002):

(i) GIs are present in the genomes of many bacteria but absent from the genomes of closely related strains.

(ii) They are often large (10-500 kb); however, smaller inserts (1-10 kb) have also been found and may be termed genomic islets.

(iii) They usually differ in G+C content and codon usage from the rest of the chromosome.

(iv) They are flanked by specific sequences (direct repeats), which may be generated following integration into the host genome via recombination.

(v) They are usually associated with tRNA loci, which presumably act as targets for the integration of foreign DNA.

(vi) They often possess genes or cryptic pseudogenes coding for genetic mobility such as phage genes, insertion sequences (IS) (Mahillon *et al*., 1999), integrases, transposases and origins of replication.

(vii) They are frequently but not necessarily unstable.

D1.4.2. Pathogenicity islands (PIs).

PIs have been discovered in many pathogens of humans, animals and plants (Osborn & Boltner, 2002). PIs encode clusters of genes whose products contribute to virulence (Hentschel & Hacker, 2001; Strauss & Falkow, 1997). They range from adherence factors, to mechanisms required for entry into the host cell or acquisition of limiting metabolites, to overt toxins. Frequently, PIs encode type III and IV secretion systems, which are responsible for modulation of host contact, interference with the host signal transduction pathways, promotion of apoptosis and entry into non-phagocytic cell.

D1.5. Taxonomy.

Taxonomy (*taxis*: arrangements or order; *nomos*: law), also known as (bio)systematics, is the study of diversity of organisms and their relationships, comprising classification, nomenclature and identification (Hofling *et al*., 1997). Theoretical studies of organism classification, which involves the clustering of units into groups, identification and development of coherent nomenclature, are essential (van Belkum *et al*., 2001). Classification means the arrangement of organisms in groups; nomenclature is the attribution of correct international scientific names to organisms; identification is the inclusion of unknown strains in groups derived from classification (Moreno, 1997). Vandamme *et al*. advocate so-called polyphasic taxonomy – taxonomy based on a combination of data obtained by various laboratory techniques (Vandamme *et al*., 1996; Moreno, 1997), which should be

widely accepted in future (Moreno, 1997).

D1.5.1. Species of bacteria.

The definition of species is a primary underlying concept; however, it is controversial and is undergoing continuous refinement (Moreno, 1997). Among microbial taxonomists, there is general agreement that the species concept currently in use – a bacterial species is defined as an entity in which members have a DNA-DNA homology value of at least 70% – is useful, pragmatic and universally applicable within the prokaryotic world (Murray & Schleifer, 1994; Moreno, 1997). A species can be seen as a "group with a common origin that is composed of the smallest diagnosable cluster of individual organisms within which there is a parental pattern of ancestry and descent". A species can also be considered as "condensed nodes" in an "otherwise cloudy, confluent taxonomic space" (Moreno, 1997).

Phylogeny – Development of systems based on evolutionary relationships rather than general resemblance (*phylon*: race or tribe; *genesis*: generation or origin) (van Belkum *et al*., 2001).

D1.6. Microarray.

Recently developed GeneChip technology provides efficient access to genetic information using miniaturized, high-density arrays of DNA or oligonucleotide probes. Such microarrays are powerful tools to study the molecular basis of many aspects on a scale that would be impossible using conventional analysis (Gabig & Wegrzyn, 2001; Magee *et al*., 2001).

D1.6.1. Microarrays have been used:

- to disclose differences in gene content between taxonomically related strains (Schoolnik, 2002) and infer their phylogeny (Whittam & Bumbaugh, 2002);

- to quantify the extent of genetic variability within natural populations at the gene level of resolution (Fitzgerald & Musser, 2001);
- to identify crucial differences between pathogen and commensal (Schoolnik, 2002);
- to study host-pathogen interactions, mainly by identifying genes from pathogens that may be involved in pathogenicity and by surveying the scope of the host response to infection (Israel *et al*., 2001);
- to analyse genetic polymorphisms of specific loci associated with resistance to antimicrobial agents; this has the potential to identify targets for drug design (Hamels *et al*., 2001; Israel *et al*., 2001).

Although many challenges lie ahead, the usability of microarrays means that they should accelerate the advance of genetic epidemiology on multiple fronts (Dalma-Weiszhausz *et al*., 2002).

D2. Ureaplasma and GBS genomic studies.

There is one overarching lesson: completion of the genomic sequence of any species answers many questions, while at the same time it invites totally new ones (Subramanian *et al*., 2001). The acquisition and analysis of ureaplasma (Glass *et al*., 2000) and GBS genome sequence data (Glaser *et al*., 2002; Tettelin *et al*., 2002) are not ends in themselves, but starting points for generating testable hypotheses relating to their pathogenesis (Subramanian *et al*., 2001).

D2.1. *U. parvum* genomic study.

D2.1.1. *U. parvum* genome.

The complete genome sequence of *U. parvum* was released in 2000; it was the third mycoplasma to be sequenced (Glass *et al*., 2000). The genome size is 751,719 bp, which is smaller than any other sequenced microbial genome except *M. genitalium*

(Fraser *et al*., 1995). The genome contains 613 predicted protein-coding genes and 39 genes that code for RNAs (Glass *et al*., 2000). *U. parvum* G+C nucleotide content is 25.5%; it is much more A+T-rich than any of the other prokaryotic genomes sequenced to date (Glass *et al*., 2000). The feature that almost all (95%) ATP synthesis is the result of urea hydrolysis make ureaplasmas unique among mycoplasmas and all bacteria (Glass *et al*., 2000) and its urease is 30- to 180-fold more efficient than reported for other bacterial ureases (Glass *et al*., 2000).

D2.1.2. Combinatorial and comparative genomics.

Five ureaplasma proteins, urease, immunoglobulin-A1 (IgA1) protease, phospholipases A and C, and MBA have been proposed as virulence factors (Robertson *et al*., 2002). The pathogenic effect of urease is caused by its generation of ammonia. IgA1 protease might give ureaplasmas the capacity to invade the upper urogenital tract by degrading human IgA1 (Robertson *et al*., 1994). Ureaplasma phospholipases may be responsible for premature labour by altering prostaglandin biosynthesis. However, neither an IgA1 protease nor phospholipase A or C genes could be identified in the *U. parvum* genome (Glass *et al*., 2000). Pollack identified the similar problem when comparing genomic and enzymatic activity data (Pollack, 2001). He suggested a combinatorial analysis involving available evidence of genomic sequence, transcription, translational phenomena, structure and enzymatic activity, which will give the best picture of the organism (Pollack *et al*., 2002). Other *U. parvum* genome based studies are limited and often as a part of a larger study or as a comparison with the other genomes (Razin *et al*., 1998). For example, the comparison of the genomes of *M. genitalium*, *M. pneumoniae* and *U. parvum*, has helped define the essential functions of a self-replicating minimal cell (Razin *et al*., 1998). Rocha and Blanchard established a bioinformatic strategy to detect the major recombination hot-spots (in particular genomic repeats) in the genomes of *U. parvum* and other 3 mycoplasma species (Rocha & Blanchard, 2002).

D2.1.3. *U. urealyticum* genome.

To clarify some problems in *U. parvum* genomic studies and to make comparative genomic studies possible, the *U. urealyticum* genome will be valuable. This would help in better understanding of the new taxonomy of human ureaplasmas (Robertson *et al*., 2002). A study by Robertson *et al*. showed that genome sizes were 760 kbp for the four serovars of *U. parvum* and 840-1,140 kbp for ten serovars (eleven strains) of *U. urealyticum* (Robertson *et al*., 1990). If *U. urealyticum* genome were available, it would help explain the extra 80 or more kbp genomic fragments. However, because *U. urealyticum* genome will not be sequenced in the foreseeable future, we must adopt alternative strategies to help understand their new taxonomy, for example, a study of core genes has proved to be promising (Daubin *et al*., 2002).

D2.2. GBS genomic studies.

D2.2.1. GBS genomes.

GBS genome projects have recently moved from the study of distantly related organisms to within-species comparisons of multiple strains (Glaser *et al*., 2002; Tettelin *et al*., 2002; Whittam & Bumbaugh, 2002) (http://www.tigr.org/). Complete sequences of two GBS genomes have been completed and a third is about to be released (Glaser *et al*., 2002; Tettelin *et al*., 2002) (http://www.tigr.org/).

D2.2.1.1. Serotype III strain NEM316 genome.

A serotype III strain NEM316 genome that was responsible for a fatal case of septicaemia was sequenced at the Pasteur Institute by Dr Glaser's group (http://www.pasteur.fr/english.html) (Glaser *et al*., 2002). The most interesting finding is 14 genomic islands, which contain known and putative virulence genes, mostly encoding surface proteins, as well as a number of genes related to mobile elements. Some of these islands could therefore be considered to be pathogenicity

islands (PIs) and they may have an important role in virulence acquisition and genetic diversity (Glaser *et al*., 2002).

D2.2.1.2. Serotype V strain 2603 V/R genome.

Tettelin *et al.* from The Institute for Genomic Research (http://www.tigr.org/) sequenced the complete genome sequence of a *S. agalactiae* serotype V strain 2603 V/R. Their comparative genome hybridization (CGH) experiments between the 2603 V/R and 19 *S. agalactiae* strains from several serotypes using whole-genome microarrays, revealed the genetic heterogeneity among *S. agalactiae* strains, even of the same serotype (Tettelin *et al*., 2002). They also found 15 genomic regions as possible GIs.

D2.2.2. GBS comparative genomic.

Besides these two completed GBS genomes, a serotype Ia reference strain A909 genome at The Institute for Genomic Research (http://www.tigr.org/) is about to be released (http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=NC_004128). While a single genome analysis provides tremendous biological insights into GBS, intraspecies comparative genomics of multiple serotypes or strains will provide substantially more information (Fraser *et al*., 2000). Strains often differ in their ability to cause disease. DNA microarrays, and bioinformatics, as powerful comparative genomic research tools, are uncovering novel virulence determinants, hidden aspects of pathogenesis, and new targets for vaccine development (Fraser *et al*., 2000; Whittam & Bumbaugh, 2002).

D2.3. Link GBS genomic data with other study findings.

Dmitriev *et al.* constructed physical chromosomal maps of *S. agalactiae* serotype II/(alpha + beta) and III/alpha strains. Their estimated genome sizes varied from 2,030 to 2,290 kb (Dmitriev *et al*., 2002), which proved to be very accurate when

compared with the sizes of the three full GBS genomes (http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=NC_004128) (Glaser *et al*., 2002; Tettelin *et al*., 2002). But their study found only six copies of ribosomal operons, which was incorrect. The possible reason was that two copies are directly connected together, which made them difficult or impossible to identify as two copies by the PFGE methods they used (Dmitriev *et al*., 1998).

Bohnsack *et al*. constructed the complete physical map of RDP III-3, which was a virulent subtype of serotype III. Their findings suggest that the genetic variation that distinguishes the RDP type III-3 from other serotype III strains occurs largely within localized areas of the genome containing known or putative virulence genes (Bohnsack *et al*., 2002). Their finding was supported by analysis of the two known GBS genomes (Glaser *et al*., 2002; Tettelin *et al*., 2002) and will be further confirmed in future by BM110 genome (also RDP III-3).

Schubert *et al.* recently described the *fbsA* gene, which encodes FbsA, a fibrinogen receptor in GBS. Their study suggested that FbsA is the major fibrinogen receptor in this strain (Schubert *et al*., 2002). Their sequencing of *fbsA* from five different GBS strains revealed significant variation in the number of repeat-encoding units (Schubert *et al*., 2002). Although it was not specifically annotated in NEM316 genome, closer examination of the genome should reveal its presence (Glaser *et al*., 2002; Schubert *et al*., 2002).

These are only very limited examples to show the “link” studies. No doubt, this kind of study in future will be more fruitful, efficient, and also mutually beneficial for both genome mining and conventional bench-top study than either alone (Pollack, 2001).

E. THESIS ARRANGEMENT

The first section above is the general introduction of the thesis.

The second section relates to molecular microbiology and includes genotyping studies of ureaplasma and GBS, which will cover most of my previous ureaplasma and GBS publications with re-organization and/or modification (see appendices). The thesis-related publications (in PDF format) have been added as appendices. The inner relationship of the publications is explained at the beginning of the appendices. Only selected parts, which are believed to be important to explain the "story"**,** have been included in the body of the thesis.

The third section relates to genomic- and bioinformatics-based microbiology and contains the genomic-based ureaplasma taxonomy and GBS intra-species comparative genomic studies. I also would examine evolutionary aspects of ureaplasma and GBS at the end of the thesis.

F. HYPOTHESIS TO BE TESTED BY THE STUDY

The relationship between GBS and ureaplasma types (serotype, subtype, genotype) and disease is still not well understood. One hypothesis is that there are variations in virulence among strains of a bacterial population, some of which more easily cause disease (Whittam & Bumbaugh, 2002). In order to test the hypothesis, well-developed genotyping systems for both of them are needed in addition to conventional serotyping methods. Hopefully, the genotyping systems are closely related to conventional serotyping, but more practical and with higher discriminatory abilities (van Belkum *et al*., 2001).

Now that both GBS and urealasmas are in the post genomic era, my hypothesis is that the genome resources and bioinformatic research tools and theories together can

help resolve previous problems more efficiently. For ureaplasmas, we would like to test whether *U. parvum* genome based core gene analysis could help better understanding of the new ureaplasma taxonomy. For GBS, we hope to further expose the pathogenicity-related GBS heterogeneity, based on intraspecies comparative genomic study.

Thus far, the most exciting work on the genomics of pathogens has clearly been in the area of evolution (Sassetti & Rubin, 2002). Both GBS and ureaplasmas are common potential perinatal pathogens and phylogenetically related. GBS is more virulent than ureaplasmas. We would like to test whether the integrated study of GBS and ureaplasmas together can help us understand them as perinatal pathogens better than study either alone. To test the hypothesis, we will examine them from an evolutionary viewpoint.

CHAPTER 1

UREAPLASMA SPECIES IDENTIFICATION

Fanrong Kong, Zhenfang Ma, Gregory James, Susannah Gordon, Gwendolyn L. Gilbert

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

This chapter has been published in *Journal of Clinical Microbiology* 2000; 38:1175-1179. (see Appendices section)

Statement of Joint Authorship

Kong, F. (candidate)

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

Ma, Z.

Assisted in some PCR.

James, G.

Provided all the needed molecular equipment, reagents and software.

Gordon, S.

Assisted in culture of related reference strains and some clinical isolates.

Gilbert G. L. (supervisor)

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

1.1. SUMMARY

Ureaplasma urealyticum have been divided into two species *U. parvum* (previously *U. urealyticum* parvo biovar) and *U. urealyticum* (previously *U. urealyticum* T-960 biovar). In this study, we designed a series of primers, targeting the 16S rRNA gene and 16S-23S rRNA intergenic spacer regions, the urease gene subunits and adjoining regions, and the 5'-ends of the multiple-banded antigen genes (*mba*) and upstream regions, to identify the two ureaplasma species. All of the species-specific primer pairs could distinguish the two species. A selection of primer pairs was used to identify 78 clinical ureaplasma isolates from vaginal swabs of pregnant women and 185 vaginal swabs. *U. parvum* was identified in 228 (87%) of 263 isolates or specimens, *U. urealyticum* was identified in 50 (19%) and both were present in 6%. The ureaplasma species identification PCR methods will facilitate future studies of the relationship between individual ureaplasma species and human diseases.

1.2. INTRODUCTION

Human ureaplasmas had been separated into two new species, namely, *U. parvum* (previously *U. urealyticum* parvo biovar) and *U. urealyticum* (previously *U. urealyticum* T-960 biovar) (Robertson *et al*., 2002). Ureaplasmas are commensals in the genital tract, recognized causes of some diseases (Taylor-Robinson & Furr, 1997), and suspected contributors to a number of other pathological conditions (Abele-Horn *et al*., 1997). Because they are commonly found in healthy people, their pathogenic role can be difficult to prove (Ollikainen *et al*., 1998; Tully, 1993). The majority of ureaplasma isolates belong to the proposed new species *U. parvum* (Abele-Horn *et al*., 1997), which includes serovars 1, 3, 6, and 14 (Robertson *et al*., 2002; Knox *et al*., 2003). *U. urealyticum* is isolated less often but is not uncommon (Abele-Horn *et al*., 1997). *U. urealyticum* has been associated with some disease syndromes more commonly than with normal flora (Povlsen *et al*., 2002; Deguchi *et al*., 2004), but data are limited because of difficulties with conventional serotyping

methods (Povlsen *et al*., 2002; Deguchi *et al*., 2004). Rapid molecular methods for ureaplasma species identification would be of great value in studies of the epidemiology and pathogenesis of infections with *U. parvum* and *U. urealyticum* (Robertson *et al*., 2002).

Recently, PCR-based methods have been used successfully to distinguish the two ureaplasma species, but there is a need to improve their specificity and sensitivity (Robertson *et al*., 2002). Target sequences of the 16S rRNA gene (Robertson *et al*., 1993), 16S-23S rRNA intergenic spacer regions (Harasawa & Kanamoto, 1999), the urease gene subunits (Blanchard, 1990; Povlsen *et al*., 1998), and the 5'-ends of the *mba* and upstream regions (Kong *et al*., 1999a; Teng *et al*., 1995) have all been used in PCR-based assays to differentiate *U. parvum* from *U. urealyticum*. Previously, we have sequenced portions of these genes from all 14 ureaplasma serovars (Kong *et al*., 1999a, b). In the present study, we evaluated the specificity of a large range of primers and used a small subset to develop ureaplasma species identification assays.

1.3. MATERIALS AND METHODS

1.3.1. Bacterial strains.

Two sets of reference strains were used. Ureaplasma serovar 1 (ATCC 27813), serovar 2 (ATCC 27814), serovar 3 (ATCC 27815), serovar 4 (ATCC 27816), serovar 5 (ATCC 27817), serovar 6 (ATCC 27818), serovar 7 (ATCC 27819), serovar 8 (ATCC 27618), serovar 9 (ATCC 33175), serovar 10 (ATCC 33699), serovar 11 (ATCC 33695), serovar 12 (ATCC 33696), serovar 13 (ATCC 33698), serovar 14 (ATCC 33697) were obtained directly from the American Type Culture Collection, Manassas, VA, U.S.A. (ATCC reference strains) (Robertson *et al*., 2002). In addition, a set of reference strains of serovars 1 to 14, were kindly provided by Dr. H. L. Watson, Department of Microbiology, University of Alabama at Birmingham, Alabama (UAB reference set). These had been obtained originally

from E. A. Freundt, Institute of Medical Microbiology, University of Aarhus, Aarhus, Denmark (serovars 1-8) and J. A. Robertson, Department of Medical Microbiology and Infectious Diseases, University of Alberta, Edmonton, Alberta, Canada (serovars 9-14) (Robertson *et al*., 2002).

Additional reference strains from ATCC were used to test the specificity of primers: *Mycoplasma pneumoniae* strains M129 (ATCC 29342) and FH (ATCC 15531), *Mycoplasma genitalium* (ATCC 33530), *Mycoplasma fermentans* (ATCC 19989), *Mycoplasma hyorhinis* (ATCC 17981), and *Acholeplasma laidlawii* (ATCC 23206). *Mycoplasma hominis* isolates were grown from clinical specimens on A7 agar in our laboratory and identified by colonial morphology and partial sequencing of the 16S rRNA gene and the 16S rRNA-23S rRNA intergenic spacer regions.

1.3.2. Clinical isolates and specimens.

A total of 78 ureaplasma isolates obtained from vaginal swabs of pregnant women recently cultured in our laboratory, and 185 vaginal swabs obtained from pregnant women and women attending a sexually transmitted disease clinic, in which ureaplasmas had been previously detected (Kong *et al*., 1999a) were used in this study.

1.3.3. Oligonucleotide primers.

The 19 individual primers used in this study to amplify portions of three genes of all 14 serovars are shown in Table 1.1. They include 4 primers that have been previously described (Robertson *et al*., 1993; Teng *et al*., 1995) and 15 new primers designed by us. The nomenclature of our primers is based on specificity (e.g., UU and UP for *U. urealyticum* and *U. parvum*, respectively), gene target (e.g., UM for *mba*), the direction of the sequence (S, sense; A, antisense), and the numbered base position at which the primer sequence starts (for *mba*) (Kong *et al*., 1999b).

1.3.4. DNA preparation.

DNA preparation was performed as previously described (Kong *et al.*, 1999a, b). Cells from 0.5 ml of ureaplasma broth cultures (10B) of each ureaplasma servovar were harvested from late logarithmic growth by centrifugation at 14, 000x g for 20 minutes; clinical specimens were processed directly. DNA was isolated from both cultures and clinical specimens by treatment with 500 µl of digestion buffer (10 mM Tris-HCl [pH 8.0], 0.45% Triton X-100 and 0.45% Tween 20) and proteinase K, 100 µg/ml, at 55°C for 1 hour and then extraction with pheno-chloroform-isoamyl alcohol (25:24:1) and chloroform-isoamyl alcohol (24:1). DNA was precipitated with 1/10 volume of 3 M sodium acetate (pH 5.2) and 2 volumes of ethanol. The washed and dried pellets were hydrated in 200 µl ultrapure and sterile water.

1.3.5. PCR.

PCR was performed as previously described (Kong *et al.*, 1999a, b). The 25 µl amplification reaction mixtures contained 2.5 µl of 10x PCR buffer (1x is 10 mM Tris-HCl [pH 8.8] at 25°C, 1.5 mM $MgCl_2$, 50 mM KCl, and 0.1% Triton X-100), 0.5 U of *Taq* polymerase (Finnzymes OY, Finland), 200 µM of each deoxynucleotide triphosphate (dATP, dCTP, dGTP, dTTP, Boehringer Mannheim, Germany), 10 ρmol of each primer, 5 µl of sample DNA and added ultrapure sterile water to 25 µl. In each reaction, positive and negative controls were processed in parallel with the tested samples to detect false-negative results or contamination. Melting temperature (*T*m) values are shown in Table 1.1. The denaturation, annealing and elongation temperatures and times used were 96°C for 10 seconds, 58-70°C (according to the primer *T*m values) for 10 seconds and 72°C for 30-60 seconds (according to the lengths of the amplicons), respectively, for 40 cycles using a Perkin Elmer thermocycler 9600.

Table 1.1. Primers targeting ureaplasma 3 different genes/regions.

Primers	**Target**	**Specificity**	***T*m °C[a]**	**GenBank numbers**	**Sequence[b]**
UPS1	16S rRNA	*U. parvum*	61.6	AF073456	**154**ATG AGA AGA TGT AGA AAG TCG CTC**177**
UPA	"	"	65.8	"	**831**TTA GCT ACA ACA CCG ACC CAT TC**809**
UPA1	"	"	67.9	"	**833**CGT TAG CTA CAA CAC CGA CCC A**812**
UPS	"	"	80.1	"	**780**CGTAAACGAT CAT CAT TAA ATG TCG GCC CGA ATG G**814**
*UPSA	16S-23S intergenic spacers	"	71.0	AF059323	**138**AAACTCTCAAAACTAAA TAG AAT CCG ACC ATA TGA ATT TTT A**97**
U8[c]	16S rRNA	*U. urealyticum*	66.7	AF073450	**158**GAA GAT GTA GAA AGT CGC GTT TGC**181**
UUA	"	"	65.8	"	**823**CTA CAA CAC CGA CTC GTT CGA G**802**
UUS	"	"	76.5	"	777GTAAACGATCAT CAT TAA ATG TCG GCT CGA ACG AG**811**
*UUSA	16S-23S intergenic spacers	"	72.3	AF059330	**138**ACTCTCAAAACTAAAT AGA GTC CGA CCA TAT GAA CTT TTG**99**
*UPS2	urease gene clusters	*U. parvum*	65.1	AF085732	**486**GATTATATGT CAG GAT CAT CAA GTC AAT TTA G**517**
*UPA2	"	"	65.4	"	**925**GAAATTTTA AAC ATA ATG TTC CCC TTT TTA TC**894**
*UUS2	"	*U. urealyticum*	67.3	AF085724	**489**GATTATATGT CAG GAT CAT CAA ATC AAT TCA C**520**

*UUA2	"	"	68.1	"	**921**GAAATTTTAAA CAT AAT GTT CCC CTT CGT CTA**890**
*UMS-125[d] *UMS-170[e]	*mba* and upstream	*Ureaplasmas*	60.5	L20329	**1**GTA TTT GCA ATC TTT ATA TGT TTT CG**26**
UMA226[d]	"	"	64.0	"	**403**CAG/A CTG ATG TAA G/TTG CAG/A CAT TA/GA ATT C**376**
*UMS-57	"	*U. parvum*	59.9	"	**64**GATTA/GA/C T/CAA ATC TTA GTG TTC ATA TTT TTT AC**95**
UMA222	"	"	63.6	"	**396**T GTA AGT GCA GCA TTA AAT TCA ATG**372**
*UMS-61	"	*U. urealyticum*	61.2	AF055366	**103**GAAAAATA TTT GCA AAA CTA TAA ATA GAC AC**135**
UMA263[e]	"	"	70.8	"	**485**AAGTGACCT TTT GTT GTT GCG TTT TCT G**458**

Notes:

* Primers based on intergenic spacer regions

a. The melting temperatures (*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 points "1" of correspondent gene GenBank accession numbers). Underlined sequences show bases added to modify previously published primers (Teng *et al*., 1994; Kong *et al*., 2000). Letters behind "/" indicate alternative nucleotides in different serovars.

c. From Robertson, *et al.*, 1993.

d. From Teng, *et al.*, 1994.

e. From Teng, *et al.*, 1995.

8 µl of PCR products were analysed by electrophoresis on 2.0% agarose gels, which were stained with 0.5 µg/ml ethidium bromide. A visible band with appropriate size on ultraviolet translumination was considered a positive result.

1.4. RESULTS

1.4.1. Specificity of *U. parvum* and *U. urealyticum* specific primer pairs.

Five primer pairs (UPS1-UPA, UPS1-UPA1, UPS-UPSA, UPS2-UPA2, and UMS-57-UMA222) were specific for and amplified all four serovars of *U. parvum*; and four (U8-UUA, UUS-UUSA, UUS2-UUA2, and UMS-61-UMA263) were specific for and amplified all ten serovars of *U. urealyticum*; UMS-125-UMA226 were specific for and amplified all 14 serovars of ureaplasmas and showed different length amplicons between *U. parvum* and *U. urealyticum*.

All the ATCC and UAB reference strains of *U. parvum* and *U. urealyticum* were correctly identified with the species-specific primers. No amplification occurred when DNA extracted from *Mycoplasma* or *Acholeplasma* species was tested with any of the above ten primer pairs.

1.4.2. *U. parvum* and *U. urealyticum* identification results for clinical isolates and clinical specimens.

Of the 78 clinical isolates, 62 (80%) were identified as *U. parvum*, 15 (19%) were identified as *U. urealyticum*, and 1 (1%) was mixed. Of 185 vaginal swabs that had been shown previously to contain ureaplasmas, 151 (82%) contained *U. parvum* only, 20 (11%) contained *U. urealyticum* only, and 14 (8%) contained both.

1.5. DISCUSSION

Recently, human ureaplasmas had been divided into two species *U. parvum* and *U. urealyticum* (Robertson *et al*., 2002). Several primer pairs have been described for the identification of the two species (previously biovars). However, they were not based on the sequences of all 14 serovars (Blanchard, 1990; Robertson *et al*., 1993) (Teng *et al*., 1994, 1995), and some lacked specificity or the ability to detect all serovars (Blanchard, 1990; Robertson *et al*., 1993). Better PCR-based methods for ureaplasma species identification are needed to facilitate studies of the relationship between ureaplasma species and disease (Cunliffe *et al*., 1996).

Many of the primer targets used in this study were based on our previous observation that the heterogeneity of the intergenic spacer regions is greater than that within the genes (Kong *et al*., 1999b). We believed that primers based on these regions would be more discriminatory for the identification of ureaplasma species (Kong *et al*., 1999b). These included primers UPSA and UUSA (16S-23S rRNA intergenic spacer regions), UPS2-UPA2 and UUS2-UUA2 (*ureA-ureB* and *ureB-ureC* intergenic spacer regions, respectively); and UMS-57, UMS-61 (upstream of *mba*) (Table 1.1.).

Differences between the two human ureaplasma species in the 16S rRNA genes have been described previously and were used to design species-specific primers (Robertson *et al*., 1993). The primer pair U8-P6, designed to amplify *U. urealyticum*, also amplified DNA from *M. pneumoniae* (Robertson *et al*., 1993; Sharma *et al*., 1998). We designed a new primer, UUA, which was paired with U8. This pair was specific for *U. urealyticum* and did not amplify DNA from two *M. pneumoniae* ATCC strains or the other mollicute species tested. Two new *U. parvum* specific primer pairs, UPS1-UPA and UPS-UPA, also based on the 16S rRNA gene, did not amplify DNA from either *U. urealyticum* or the other mollicute species tested.

The 16S-23S rRNA intergenic spacer regions are normally more heterogeneous than the 16S rRNA genes (Harasawa *et al*., 1993; Kong *et al*., 1999b). Primers spanning these regions UPS-UPSA for *U. parvum* and UUS-UUSA for *U. urealyticum* were specific for all serovars within the corresponding species for both ATCC and UAB reference strains.

Primers based on urease gene sequences used to differentiate *U. parvum* and *U. urealyticum* have been described previously (Blanchard, 1990). We designed two additional species-specific primer pairs, UPS2-UPA2 (*U. parvum* specific) and UUS2-UUA2 (*U. urealyticum* specific), targeting the *ureA-ureB* and *ureB-ureC* intergenic spacer regions, respectively. They were specific for all serovars within the corresponding species for both ATCC and UAB reference strains.

The *mba* contains both species- and serovar-defining regions (Zheng *et al*., 1992, 1995). Several primer sets based on 5'-ends of *mba* and upstream region sequences have been described previously for differentiating *U. parvum* and *U. urealyticum* (Teng *et al*., 1994, 1995). We noticed that UMS-125-UMA226 were specific for human ureaplasmas and could differentiate *U. parvum* and *U. urealyticum* in a single PCR reaction according to their different amplicon lengths; however it was difficult to identify mixtures of the two species, especially in clinical specimens, in which mixtures were not rare. To improve sensitivity and specificity and to provide more choice, we designed additional species-specific primers based on our sequencing results. The primer pair UMS-57-UMA222 was specific for *U. parvum*, and UMS-61-UMA263 was specific for *U. urealyticum*.

Having confirmed the sensitivity and specificity of the new primer pairs, we identified *U. parvum* and *U. urealyticum*, using a selection of the most suitable primers. Its utility was evaluated with ATCC and UAB reference strains, stored clinical isolates, and ureaplasma-positive clinical specimens. Identification of clinical isolates and specimens confirmed the previous finding that *U. parvum* is found much more commonly (87% of isolates and specimens overall) than

U. urealyticum (19% overall) among vaginal flora (Cheng *et al*., 1994). Both ureaplasma species were detected in 1 clinical isolate (1%) and 14 vaginal swabs (8%) (6% overall).

In summary, we have designed a series of new primer pairs based on previously reported sequences of three important ureaplasma genes and adjoining regions and modified some previously published primers to improve their sensitivity and specificity. Our initial evaluation of the identification of *U. parvum* and *U. urealyticum*, using a selected set of primers with ATCC and UAB reference strains, clinical isolates, and clinical specimens, confirmed their specificity and sensitivity. We believe that, in the future, they will assist in studies of the epidemiology, pathogenicity, and clinical significance of ureaplasma species in humans and will provide significant advantages over conventional serotyping methods.

CHAPTER 2

UREAPLASMA GENOTYPING

Fanrong Kong, Zhenfang Ma, Gregory James, Susannah Gordon, Gwendolyn L. Gilbert

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

This chapter has been published in *International Journal of Systematic and Evolutionary Microbiology* 2000; 50:1921-1929. (see Appendices section)

Statement of Joint Authorship

Kong, F. (candidate)
Did all the molecular work, interpreted the data and wrote the manuscript.
Ma, Z.
Assisted in some PCR.
James, G.
Provided all the needed molecular equipment, reagents and software.
Gordon, S.
Assisted in culture of related reference strains.
Gilbert G. L. (supervisor)
Supervised the overall project, assisted in research design, analysis and interpretation of data, and made a significant contribution to the manuscript.

2.1. SUMMARY

U. parvum comprises four serovars (1, 3, 6, 14) and *U. urealyticum*, ten serovars (2, 4, 5, 7-13). The multiple banded antigen genes (*mba*) of ureaplasmas contain both species and serovar specific sequences. To further elucidate the relationships between serovars and establish serovar identification assays, we sequenced the 5'-ends (including partial repetitive regions) and upstream regions of *mba* for all 14 serovars of ureaplasmas. All four serovars of *U. parvum* were clearly differentiated from each other. Ten serovars of *U. urealyticum* were divided into five *mba* genotypes, as follows: *mba* genotype A comprises serovars 2, 5, 8; *mba* genotype B, serovar 10 only; *mba* genotype C, serovars 4, 12, 13; *mba* genotype D, serovar 9 only; and *mba* genotype E comprises serovars 7 and 11. There were no sequence differences between members within each *mba* genotype. Further work is required to identify other genes or sequences in other regions of *mba* that may be used to differentiate *U. urealyticum* serovars within *mba* genotypes A, C and E. A better understanding of the molecular basis of serovar differentiation will help to improve serovar identification methods for use in studies of the pathogenesis and epidemiology of human ureaplasmas.

2.2. INTRODUCTION

Human ureaplasmas are recognised causes of urethritis (Taylor-Robinson *et al*., 1985), and have been associated with complications of pregnancy and prematurity (Cassell *et al*., 1988; Kundsin *et al*., 1996). However, as common genital tract commensals, their pathogenic roles in individual cases are difficult to confirm (Robertson & Stemke, 1982; Heggie *et al*., 2001).

U. parvum comprises four serovars (1, 3, 6, 14) and *U. urealyticum*, ten serovars (2, 4, 5, 7-13) (Robertson & Stemke, 1982). The relationships between serovars and disease syndromes needs to be studied further (Naessens *et al*., 1988; Knox *et al*.,

2003). However, this has been limited by technical difficulties and cross-reactions associated with serotyping (Stemke & Robertson, 1985), even when monoclonal antibodies were used (Naessens *et al*., 1998).

Better understanding of the genetic basis of conventional ureaplasma serotyping will assist in development of a practical molecular serovar identification assay and allow further investigation of the pathogenic potential of individual serovars (Robertson & Stemke, 1982; Kong *et al*., 1999a). In our previous study, we showed that the sequences of the 16S rRNA genes and 16S-23S rRNA intergenic spacer regions, the urease gene subunits and adjoining regions were generally conserved for serovars within the two ureaplasma species. Only the 5'-ends (including partial repetitive regions) and upstream regions of *mba* showed heterogeneity between the 4 serovars of *U. parvum* and the 10 serovars of *U. urealyticum* (Kong *et al*., 1999b).

It has been suggested, previously, that the repetitive regions of the *mba* would contain serovar-specific sites (Watson *et al*., 1990; Zheng *et al*., 1996). In this study, we sequenced the 5'-ends of *mba* and upstream regions of all 14 serovars of ureaplasmas to determine whether these genes could be used to develop alternative serovar identification methods (Kong *et al*., 1999a, 2000).

2.3. MATERIALS AND METHODS

2.3.1. Bacterial strains.

Two sets of reference strains of all 14 serovars of ureaplasmas were used as previously described (Kong *et al*., 1999b) and as listed in chapter 1. One set was obtained directly from the American Type Culture Collection (ATCC reference strains) and the other was kindly provided by Dr. H. L. Watson, Department of Microbiology, University of Alabama at Birmingham, Alabama (UAB reference strains).

2.3.2. Oligonucleotide primers.

The oligonucleotide primers used in this study are shown in Table 2.1. Previously published oligonucleotide primers UMS-125, UMA226, UMA1213, UMA1586 (Teng *et al*., 1994; Zheng *et al*., 1995), and new primers designed by us – UMS-57, UMA222, UMSPS1, UMSPS2, UMAUA – based on previously published sequences (Zheng *et al.*, 1999 submitted to GenBank, accession numbers: U50459, U50460, U50461) (Zheng *et al*., 1995) were used to sequence *mba* of the four *U. parvum* serovars. Previously published oligonucleotide primers UMS-170, UMA226, UMA263 (Teng *et al*., 1994, 1995), and new primers designed by us – UMS-61, UMSUS, UMSUS1, UMSUS2, UMAUA, UMAUA1, UMAUA2 – based on the previously published sequences (Zheng *et al.*, 1999 submitted to GenBank, accession numbers: U50459, U50460, U50461) (Zheng *et al*., 1995) were used to sequence *mba* of all the ten *U. urealyticum* serovars.

Additional new primers – UMS3S, UMA314A, UMA314A', UMS14S, UMA1A, UMA1A', UMA6A, UMA6A' – based on sequences determined in this and previous studies (Kong *et al*., 1999b), and previously published primers designed by us UMS-83, UMS-54, UMA269, and UMA269' were designed specifically to amplify and differentiate *mba* of four *U. parvum* serovars 3, 14, 1 and 6. New primers UMA2A1, UMA2A2; UMA2A; UMA10A; UMA4A1, UMA4A2; UMA9A1, UMA9A2; UMA7A1, UMA7A2 and UMA7A3; UMS-112, UMS-112', UMA194, UMA194', UMA219 based on the sequences obtained in this study were designed to amplify and differentiate *mba* genotypes A to E (Table 2.1.).

2.3.3. DNA preparation and PCR.

DNA preparation and PCR system were used as previously described (Kong *et al*., 1999a, b).

Table 2.1. Primers for ureaplasma genotyping.

Primer	Specificity[a]	*T*m °C[b]	GenBank numbers	Sequence[c]
UMS-125[d] UMS-170[e]	Ureaplasmas	60.5	L20329	**1**GTA TTT GCA ATC TTT ATA TGT TTT CG**26**
UMA226[d]	"	64.0	"	**403**CAG/A CTG ATG TAA G/TTG CAG/A CAT TA/GA ATT C**376**
UMS-57	*U. parvum*	59.9	"	**64**GATTA/GA/C T/CAA ATC TTA GTG TTC ATA TTT TTT AC**95**
UMA222	"	63.6	"	**396**T GTA AGT GCA GCA TTA AAT TCA ATG**372**
UMS-61	*U. urealyticum*	61.2	AF055366	**103**GAAAAATA TTT GCA AAA CTA TAA ATA GAC AC**135**
UMA263[e]	"	70.8	"	**485**AAGTGACCT TTT GTT GTT GCG TTT TCT G**458**
UMSPS1	"	64.2	L20329	**469**CCT CGT GAA CCA AAA CCT AAT G**490**
UMSPS2	"	61.5	"	**517**GGA TTA ATC AAG ACT TCA GGT TTG**540**
UMA1213	UP 3/14	56.0	"	**1389**CTA AAG TAA TTA TTT TCC AGT AGT TTC**1363**
UMA1586	"	60.1	AE002134	**8480**GAT AAT CAT TCA TCT TCT CTT AAT TGT C**8507**
UMS3S	UP 3	58.9	L20329	**38**AATTAA TTA CTG TAG AAA TTA TGT AAG ATT ACC**70**
UMS14S	UP 14	58.9	AF056982	**36**AAAATT AAT TAC TGT AGA AAT TAT GTA AGA TTA AT**70**
UMA269	UP 3/14	61.7	L20329	**445**C AAC TAA ATG ACC TTT TTC AAG TGT AC**419**
UMA314A	"	60.4	"	**613**GTT GTT CTT TAC CTG GTT GTG TAG**590**
UMA314A'	"	63.2	"	**615**TG/TG TTG TTC TTT ACC TGG TTG TGT A**591**
UMS-54	UP 6	58.2	AF056984	**66**TTAATA AAT CTT AGT GTT CAT ATT TTT TAC TAG**98**

UMA6A	"	65.8	"	**615**CCT GGT TCT TGA GTT TTC GGA G**594**
UMA6A'	"	65.9	"	**619**TTT ACC TGG TTC TTG AGT TTT CGG**596**
UMS-83	UP 1	60.2	AF056983	**38**AATTAA TTACT GTA GAA ATT ATG TAA GAT TGC**69**
UMA1A	"	65.1	"	**619**TTT CTT TTG GTT CTT CAG TTT TTG AAG**593**
UMA1A'	"	64.5	"	**621**ATT TTC TTT TGG TTC TTC AGT TTT TGA**595**
UMA269'	UP 1/6	63.0	"	**445**CA ACC AAA TGA CCT TTT GTA ACT AGA T**419**
UMSUS	*U. urealyticum*	64.0	AF055366	**358**GTT TAC GAC ATT GAA AAT TTC GAT G**382**
UMAUA	"	66.4	"	**666**GGG G/TA/TG TTG/T A/C/TAC CAC/T TG/TC CTG GTT**638**
UMSUS1	"	53.1	"	**573**AAC TGC ATC TC/TT AGC/T ATT ACC T**594**
UMSUS2	"	56.7	"	**592**CCT GAT AAT TTG/T AAT TAT CAA ACA G**616**
UMAUA1	"	63.1	U50459	**1687**GCC CAA TTC ATA GGC TAT TAA TTG**1664**
UMAUA2	"	63.8	"	**1696**AAA AAA ATA GCC CAA TTC ATA GGC**1673**
UMA2A1	UU A/B	64.2	AF055366	**646**TTC CTG GTT TTG TTT CAA AAC CTA T**622**
UMA2A2	"	64.4	"	**649**CAC TTC CTG GTT TTG TTT CAA AAC**626**
UMA2A	UU A	62.5	"	**674**CCA CTT CCT GGT TTT GTA GTT TC**652**
UMA10A	UU B	65.0	AF055358	**674**CCA CTT CCT GGT TGT GTA GTT GA**652**
UMA4A1	UU C	69.1	AF055363	**647**TT GCC TGG TTG TGT TTC GAA CTC**625**
UMA4A2	"	69.3	"	**649**CAT TGC CTG GTT GTG TTT CGA AC**627**
UMA9A1	UU D	68.3	AF055367	**655**CTG GAG TTG GTG TAG GCG CAT**635**

UMA9A2	"	68.4	"	**657**TTC TGG AGT TGG TGT AGG CGC**637**
UMA7A1	UU E	67.2	AF055365	**440**GTA ATT GCA ACA TGG AAT TCA GTT TCA**415**
UMA7A2	"	63.0	"	**653**GGT TCT GGT GTA TGA GTG CTT TT**631**
UMA7A3	"	63.3	"	**656**GTT GGT TCT GGT GTA TGA GTG C**635**
UMA219	UU A–D	70.3	AF055366	**440**GTA ATT GCA ACA TGG AAT TCA GCT TCG**414**
UMS-112	UU A/D/E	56.3	"	**59**GAT TAA ACA AAA TCT TAA TGT TGT TA**84**
UMA194	UU A/D/E	63.3	"	**415**CGT TTA ATG CTT TTT TAT CAT TTT CAG**389**
UMS-112'	UU B/C	58.6	AF055363	**59**GAT TAA ACA AAA TCT TAA TGT TGT TG**84**
UMA194'	UU B/C	62.8	"	**415**CGT TTA ATG CTT TTT TAT CAT TTT CAT**389**

Notes.

a. UP 3: *U. parvum* serovar 3; UP 14: *U. parvum* serovar 14; UP 3/14: *U. parvum* serovars 3 and 14; UP 1: *U. parvum* serovar 1; UP 1/ 6: *U. parvum* serovars 1 and 6. UP 6: *U. parvum* serovar 6. UU A: *U. urealyticum mba* genotype A, includes serovars 2, 5, and 8; UU B: *U. urealyticum mba* genotype B, includes serovar 10; UU C: *U. urealyticum mba* genotype C, includes serovars 4, 12, and 13; UU D: *U. urealyticum mba* genotype D, includes serovar 9; UU E: *U. urealyticum mba* genotype E, includes serovars 7 and 11.

b. Primer melting temperatures (*T*m) were provided by the primer synthesiser (Sigma-Aldrich, Castle Hill, NSW, Australia).

c. Numbers represent the numbered base positions at which primer sequences start and finish (numbering start point "1" refer to the start points "1" of correspondent gene GenBank accession numbers). Underlined sequences show bases added to modify previously published primers (Teng *et al*., 1994; Kong *et al*., 2000). Letters behind "/" indicate alternative nucleotides in different serovars.

d. From Teng, *et al.*, 1994.

e. From Teng, *et al.*, 1995.

To amplify the 5'-ends and repetitive regions of *mba* genes of *U. parvum* serovars for sequencing, a nested PCR was developed, using UMS-125-UMA1586 as outer primers and UMS-57-UMA1213 (for serovars 3 and 14) and UMS-57-UMAUA (for all four serovars of *U. parvum*) as inner primers. Nested PCR was also used to amplify the 5'-ends and repetitive regions of the *mba* of the ten *U. urealyticum* serovars for sequencing. The outer primers were UMS-170-UMAUA2 and inner primers were UMS-61-UMAUA and UMSUS-UMAUA2 (or UMSUS-UMAUA1) (Table 2.1.).

The denaturation, annealing and elongation temperatures and times used for the first step PCR were 95°C for 30 seconds, 50°C for 30 seconds and 72°C for 3 minutes, respectively, for 30 cycles. For the second step PCR, the denaturation, annealing and elongation temperatures and times used were 95°C for 30 seconds, 55°C for 30 seconds and 72°C for 2 minutes, for 30 cycles. For the serovar- or genotype-specific PCR, the denaturation, annealing and elongation temperatures and times used were 95°C for 30 seconds, 55-62°C (according to the *T*m value) for 30 seconds and 72°C for 1 minute, respectively, for 40 cycles.

12.5 µl of PCR products were analysed by electrophoresis on 2.0% agarose gels, which were stained with 0.5 mg/ml ethidium bromide. For sequencing, PCR products of appropriate size that produced visible bands on ultraviolet illumination were further purified. For identification of individual subtypes, the presence of PCR amplicons of expected length on ultraviolet transillumination were accepted as positive.

2.3.4. Sequencing and sequence analysis.

The PCR products were sequenced using Applied Biosystems (ABI) *Taq* DyeDexoy terminator cycle-sequencing kits according to standard protocols. Primer UMSPS1 (or UMSPS2) was used as sequencing primer for the amplicons of UMS-57-UMA1213 (for serovars 3 and 14), and UMS-57-UMAUA (for serovars 1 and 6);

UMSUS, UMSUS1 (or UMSUS2) were used as sequencing primers for the amplicons of UMS-61-UMAUA, and UMSUS-UMAUA2 (or UMSUS-UMAUA1).

Multiple sequence alignments were performed using *Pileup* and *Pretty* programs from the Multiple Sequence Analysis program group, provided in WebANGIS, ANGIS (Australian National Genomic Information Service), 3rd version.

2.3.5. Nucleotide sequence accession numbers.

The new sequence data in the study were deposited into GenBank Nucleotide Sequence Databases with the following accession numbers: AF055358-AF055367, AF056982-AF056984 (*mba*). The following GenBank sequences were used as references: NC_002162 (*U. parvum* serovar 3 genome), U50462 (serovar 14 *mba*, partial coding sequences [cds]), L20329 (serovar 3 *mba*, complete cds), U50459 (serovar 10 *mba*, complete cds), U50460 (serovar 2 *mba*, partial cds) and U50461 (serovar 4 *mba*, partial cds).

2.4. RESULTS

2.4.1. PCR and sequencing.

As predicted, the inner primer pair UMS-57-UMA1213 produced amplicons only from serovars 3 and 14 and UMS-57-UMAUA produced amplicons from all four serovars of *U. parvum*. From *U. urealyticum,* inner primers UMS-61-UMAUA produced amplicons from all ten serovars whereas UMSUS-UMAUA2 (or UMSUS-UMAUA1) produced amplicons from seven (all except serovars 9, 7 and 11). UMS-125-UMA226 produced amplicons from all 14 serovars of *U. parvum* and *U. urealyticum.* The relevant portions of sequencing results for each serovar were submitted to GenBank after further analysis (see below).

2.4.2. Comparative study of the sequences of the 5'-ends and upstream of *mba*.

There were base differences at 45 (45/601=7.5 %) sites at the *mba* regions (1 ~ 650) among the four serovars of *U. parvum* (Kong *et al*., 1999b) (Figure 2.1.). There were base differences at 22 (22/634=3.5%) sites at the *mba* regions (1 ~ 639) among the ten serovars of *U. urealyticum* (Figure 2.1.).

Nucleotide sequences of the *mba* repetitive units of *U. parvum* and *U. urealyticum* are shown in Figure 2.1. They begin in the vicinity of nucleotide 651. There were differences between sequences from all four *U. parvum* serovars. Sequences from serovars 2, 5 and 8 of *U. urealyticum* were identical and grouped as *mba* genotype A. The serovar 10 sequence was the same length but differed from *mba* genotype A by 3/24 nucleotide bases and it was classified as *mba* genotype B. Serovar 4, 12 and 13 sequences were longer than those of *mba* genotypes A and B, but identical with each other and were grouped together as *mba* genotype C. No repetitive units were identified for serovars 9, 7 and 11. However, there were differences between serovar 9 and serovars 7/11 in 58 (58/391=14.8%) nucleotide bases in the region 640-1033 and 14 (14/634=2.2%) in the region 1-639 (Figure 2.1.). These differences defined two additional *mba* genotypes, D (serovar 9) and E (serovars 7 and 11).

2.4.3. The specificity of *Ureaplasmas* genotyping primers.

All the ATCC and UAB reference strains of *U. parvum* and *U. urealyticum* were correctly identified by the *mba* serovar- or genotype-specific primer pairs. The PCR results for all the serovars of *U. parvum* and *U. urealyticum*, using the 36 serovar- or genotype-specific primer pairs to amplify the 5'-ends of the *mba* are summarised in Table 2.2.

Figure 2.1. Multiple sequence alignment of the 5'-end and upstream of *mba.*

```
            1                                                 50
 serovar 1  ---------- ---------- ---------- ---------- -----a--g-
 serovar 3  ---------- ---------- ---------- ---------- -----a--g-
serovar 14  ---------- ---------- ---------- ---------- -----a--g-
 serovar 6  ---------- ---------- ---------- ---------- -----a----
serovar 11  ---------- ---------- ---------- ------.... .---------
 serovar 7  ---------- ---------- ---------- ------.... .---------
 serovar 9  ---------- ---------- ---------- ------.... .---------
serovar 12  ---------- ---------- ---------- ------.... .---------
serovar 13  ---------- ---------- ---------- ------.... .---------
 serovar 4  ---------- ---------- ---------- ------.... .---------
 serovar 2  ---------- ---------- ---------- ------.... .---------
 serovar 5  ---------- ---------- ---------- ------.... .---------
 serovar 8  ---------- ---------- ---------- ------.... .---------
serovar 10  ---------- ---------- ---------- ------.... .---------
 Consensus  GTATTTGCAA TCTTTATATG TTTTCGTTAA AATTAAAAAT TAATTCCTAT

            51                                               100
 serovar 1  -g---tt-tg -a-----..g -t-------- g----ca--- -t-t--c---
 serovar 3  -g---tt-tg -a-----..- -c-------- g----ca--- -t-t--c-t-
serovar 14  -g---tt-tg -a-----..- at-------- g----ca--- -t-t--c-t-
 serovar 6  -----tt-tg -a-----..- at-------- g----ca--- -t-t--ctag
serovar 11  ---------- ---------- ---------- ---------- ----------
 serovar 7  ---------- ---------- ---------- ---------- ----------
 serovar 9  ---------- ---------- ---------- ---------- ----------
serovar 12  ---------- ---------- ---------- --------g- ----------
serovar 13  ---------- ---------- ---------- --------g- ----------
 serovar 4  ---------- ---------- ---------- --------g- ----------
 serovar 2  ---------- ---------- ---------- ---------- ----------
 serovar 5  ---------- ---------- ---------- ---------- ----------
 serovar 8  ---------- ---------- ---------- ---------- ----------
serovar 10  ---------- ---------- ---------- --------g- ----------
 Consensus  AAAAACAACA TGAGATTAAA CAAAATCTTA ATGTTGTTAT TATCTATACA

            101                                              150
 serovar 1  -at----.t- --g....... .......... .......... ..........
 serovar 3  -at----.t- ---....... .......... .......... ..........
serovar 14  -at----.t- ---....... .......... .......... ..........
 serovar 6  -at----tt- ---....... .......... .......... ..........
serovar 11  ---------- ---------- ---------- ---------- ----------
 serovar 7  ---------- ---------- ---------- ---------- ----------
 serovar 9  ---------- ---------- ---------- ---------- ----------
serovar 12  ---------- ---------- ---------- ---------- ----------
serovar 13  ---------- ---------- ---------- ---------- ----------
 serovar 4  ---------- ---------- ---------- ---------- ----------
 serovar 2  ---------- ---------- ---------- ---------- ----------
 serovar 5  ---------- ---------- ---------- ---------- ----------
 serovar 8  ---------- ---------- ---------- ---------- ----------
serovar 10  ---------- ---------- ---------- ---------- ----------
 Consensus  TTCTAAAGAA AAATATATTT GCAAAACTAT AAATAGACAC AAAAAACAAT

            151                                              200
 serovar 1  ........-c --ta----ga -------tt- a--------- -..a-c----
 serovar 3  ........-c --ta----ga -------tt- a--------- -..a-c----
serovar 14  ........-c --ta----ga -------tt- a--------- -..a-c----
 serovar 6  ........-c --ta----ga -------tt- a--------- -..acc----
serovar 11  ---------- ---------- -g-------- ---------- ----------
 serovar 7  ---------- ---------- -g-------- ---------- ----------
 serovar 9  ---------- ---------- ---------- ---------- ----------
serovar 12  ---------- ---------- -g-------- ---------- ----------
serovar 13  ---------- ---------- -g-------- ---------- ----------
 serovar 4  ---------- ---------- -g-------- ---------- ----------
 serovar 2  ---------- ---------- ---------- ---------- ----------
 serovar 5  ---------- ---------- ---------- ---------- ----------
 serovar 8  ---------- ---------- ---------- ---------- ----------
serovar 10  ---------- ---------- -g-------- ---------- ----------
 Consensus  AGAATAATAA AACTAAATTT CATATTTAGT TTATTAGGAG ATCGTTATAA
```

```
             201                                                250
 serovar 1   ---------- ---------- a-----c--- --t--g---t ----a--t--
 serovar 3   ---------- ---------- a-----c--- --t--g---t ----t--t--
serovar 14   ---------- ---------- a-----c-t- --t--g---t ----t--t--
 serovar 6   ---------- ---------- a-----c--- --t--g---t ----a--t--
serovar 11   ---------- ---------- ---------- ---------- ----------
 serovar 7   ---------- ---------- ---------- ---------- ----------
 serovar 9   ---------- ---------- ---------- ---------- ----------
serovar 12   ---------- ---------- ---------- ---------- ----------
serovar 13   ---------- ---------- ---------- ---------- ----------
 serovar 4   ---------- ---------- ---------- ---------- ----------
 serovar 2   ---------- ---------- ---------- ---------- ----------
 serovar 5   ---------- ---------- ---------- ---------- ----------
 serovar 8   ---------- ---------- ---------- ---------- ----------
serovar 10   ---------- ---------- ---------- ---------- ----------
 Consensus   ATGAAATTAT TAAAAAATAA GAAATTTTGA GCAATTACAC TAGGGGTAAC

             251                                                300
 serovar 1   c-----t--- --t--aa--- -----a-a-- ---------- ----at----
 serovar 3   c-----t--- --t--aa--- -----a-a-- ---------- ----at----
serovar 14   c-----t--- --t--aa--- -----a-a-- ---------- ----at----
 serovar 6   c-----t--- --t--aa--- -----a-a-- g--------- ----at----
serovar 11   ---------- ---------- ---------- ---------- ----------
 serovar 7   ---------- ---------- ---------- ---------- ----------
 serovar 9   ---------- ---------- ---------- ---------- ----------
serovar 12   ---------- ---------- ---------- ---------- ----------
serovar 13   ---------- ---------- ---------- ---------- ----------
 serovar 4   ---------- ---------- ---------- ---------- ----------
 serovar 2   ---------- ---------- ---------- ---------- ----------
 serovar 5   ---------- ---------- ---------- ---------- ----------
 serovar 8   ---------- ---------- ---------- ---------- ----------
serovar 10   ---------- ---------- ---------- ---------- ----------
 Consensus   TTTAGTGGGA GCAGGGGTAG TTGCTGTGGC AGCTTCATGT TCTAGCTCAA

             301                                                350
 serovar 1   cc-------- ---------- -ac---t--- c--------c ------t--a
 serovar 3   c--------- ---g------ -ac---t--- c--------c -----gt--a
serovar 14   c--------- ---g------ -ac---t--- c--------c -----gt--a
 serovar 6   c--------- ---g------ --c---t--- ---------c ---t--t--a
serovar 11   ---------- ---------- ---------- ---------- ----------
 serovar 7   ---------- ---------- ---------- ---------- ----------
 serovar 9   ---------- ---------- ---------- ---------- ----------
serovar 12   ---------- ---------- ---------- ---------- ----------
serovar 13   ---------- ---------- ---------- ---------- ----------
 serovar 4   ---------- ---------- ---------- ---------- ----------
 serovar 2   ---------- ---------- ---------- ---------- ----------
 serovar 5   ---------- ---------- ---------- ---------- ----------
 serovar 8   ---------- ---------- ---------- ---------- ----------
serovar 10   ---------- ---------- ---------- ---------- ----------
 Consensus   ATGTTAAATC TAAATTAAGT AGTCAACTTG TTAAATCAAA AGACGAAAAG

             351                                                400
 serovar 1   --t-----t- -g-------- a--------c --ta-a---c ---g---t--
 serovar 3   --t-----t- -g-------- a--------c --ta-a---c ---g-a-tg-
serovar 14   --t-----t- -g-------- a--------c --ta-a---c ---g-a-tg-
 serovar 6   --t-----t- -a-------- a--------c --ta-a---c ---g---t--
serovar 11   ---------- ---------- ---------- ---------- ----------
 serovar 7   ---------- ---------- ---------- ---------- ----------
 serovar 9   ---------- ---------- ---------- ---------- ----------
serovar 12   ---------- ---------- ---------- ---------- ---a------
serovar 13   ---------- ---------- ---------- ---------- ---a------
 serovar 4   ---------- ---------- ---------- ---------- ---a------
 serovar 2   ---------- ---------- ---------- ---------- ----------
 serovar 5   ---------- ---------- ---------- ---------- ----------
 serovar 8   ---------- ---------- ---------- ---------- ----------
serovar 10   ---------- ---------- ---------- ---------- ---a------
 Consensus   AGCTTTTACG CTGTTTACGA CATTGAAAAT TTCGATGATT TAACTGAAAA

             401                                                450
 serovar 1   ---------- t-------t- -cat------ t----c---- c---------
```

```
 serovar 3  ---------- t------gta -cat------ t----c---- c---------
serovar 14  ---------- t------gta -cat------ t----c---- c---------
 serovar 6  ---------- t-------t- -cat------ t----c---- c---------
serovar 11  ---------- --------t- --a------- ---------- ----------
 serovar 7  ---------- --------t- --a------- ---------- ----------
 serovar 9  ---------- ---------- ---------- ---------- ----------
serovar 12  ---------- ---------- ---------- ---------- ----------
serovar 13  ---------- ---------- ---------- ---------- ----------
 serovar 4  ---------- ---------- ---------- ---------- ----------
 serovar 2  ---------- ---------- ---------- ---------- ----------
 serovar 5  ---------- ---------- ---------- ---------- ----------
 serovar 8  ---------- ---------- ---------- ---------- ----------
serovar 10  ---------- ---------- ---------- ---------- ----------
 Consensus  TGATAAAAAA GCATTAAACG AAGCTGAATT CAATGTTGCA ATTACATCAG

            451                                                500
 serovar 1  t------c-- ---------t ct-gtt---- -------t-- gg--ggtg--
 serovar 3  -------c-- --------gt a--cttga-- -------t-- -g--ggtg--
serovar 14  -------c-- --------gt a--cttga-- -------t-- -g--ggtg--
 serovar 6  t------c-- ---------t ct-gtt---- -------t-- gg--ggtg--
serovar 11  t--------- ---------- -----t-t-- ---------- ---------g
 serovar 7  t--------- ---------- -----t-t-- ---------- ---------g
 serovar 9  ---------- ---------- ---------- ---------- ----------
serovar 12  ---------- ---------- ---------- ---------- ----------
serovar 13  ---------- ---------- ---------- ---------- ----------
 serovar 4  ---------- ---------- ---------- ---------- ----------
 serovar 2  ---------- ---------- ---------- ---------- ----------
 serovar 5  ---------- ---------- ---------- ---------- ----------
 serovar 8  ---------- ---------- ---------- ---------- ----------
serovar 10  ---------- ---------- ---------- ---------- ----------
 Consensus  CTGAAAATAA AACAGAAAAC GCAACAACAA AAGGTCACTT ACTTAACAAA

            501                                                550
 serovar 1  -----t--c- ---------- t--------- ---c----t- ----------
 serovar 3  -----t--c- ---------- t--------- ---c----t- ----------
serovar 14  -----t--c- ---------- t--------- ---c----t- ----------
 serovar 6  -----t--c- ---------- t--------- ---c----t- ----------
serovar 11  --------c- ---------- ---------- ---------- -g--------
 serovar 7  --------c- ---------- ---------- ---------- -g--------
 serovar 9  ---------- ---------- ---------- ---------- ----------
serovar 12  ---------- ---------- ---------- ---------- ----------
serovar 13  ---------- ---------- ---------- ---------- ----------
 serovar 4  ---------- ---------- ---------- ---------- ----------
 serovar 2  ---------- ---------- ---------- ---------- ----------
 serovar 5  ---------- ---------- ---------- ---------- ----------
 serovar 8  ---------- ---------- ---------- ---------- ----------
serovar 10  ---------- ---------- ---------- ---------- ----------
 Consensus  AAAATCTATG TTAAATTACC ACGTGAACCA AAAGCTAAAG AACAATTAAC

            551                                                600
 serovar 1  ---------- ---a----ac --a-c--g-- -t--gg---g t--a------
 serovar 3  --------g- ---a----a- --a-c--g-- -t--gg---g t--a--t---
serovar 14  --------g- ---a----a- --a-c--g-- -t--gg---g t--a--t---
 serovar 6  ---------- ---a----a- --a-c--g-- -t--gg---g t--a-----a
serovar 11  ---------- --------a- ---------- -------c-- -c--------
 serovar 7  ---------- --------a- ---------- -------c-- -c--------
 serovar 9  ---------- ---------- ---------- ---------- ----------
serovar 12  ---------- ---------- ---------- ---------- ----------
serovar 13  ---------- ---------- ---------- ---------- ----------
 serovar 4  ---------- ---------- ---------- ---------- ----------
 serovar 2  ---------- ---------- ---------- ---------- ----------
 serovar 5  ---------- ---------- ---------- ---------- ----------
 serovar 8  ---------- ---------- ---------- ---------- ----------
serovar 10  ---------- ---------- ---------- ---------- ----------
 Consensus  TATTATTAAT AAAGGTGGCT TACTAAAAAC TGCATCTTTA GTATTACCTG

            601                                                650
 serovar 1  ---------- ---------- --------ga ---------- -gctt--aa-
 serovar 3  ---------- ---------- --------ga ---------- t-c--a-cc-
serovar 14  ---------- ---------- --------ga ---------- t-c--a-cc-
```

```
 serovar 6  ---------- ---------- --------ga ---------- -gct--gaa-
serovar 11  ---------- ---------- --------a- -----a---g c-ct-atac-
 serovar 7  ---------- ---------- --------a- -----a---g c-ct-atac-
 serovar 9  -------t-- ---------- --------a- ---------a tgcg--tac-
serovar 12  ---------- ---------- --------g- -g--c----- -ca------c
serovar 13  ---------- ---------- --------g- -g--c----- -ca------c
 serovar 4  ---------- ---------- --------g- -g--c----- -ca------c
 serovar 2  -------t-- ---------- -----ta-a- gt-------- --a-------
 serovar 5  -------t-- ---------- -----ta-a- gt-------- --a-------
 serovar 8  -------t-- ---------- -----ta-a- gt-------- --a-------
serovar 10  ---------- ---------- -----ta-a- gt-------- -ca-------
 Consensus  ATAATTTGAA TTATCAAACA GAAAAAGT-G ACTTTGAAAC AA-ACCAGGA

            651                                                 700
 serovar 1  -c--aa-a-c ---...a--a ---tgtt--- g-acaac--g g--a-g---a
 serovar 3  g--aaa-a-c a-c...---- -ggtaa--a- c-acca---g g--a-g---a
serovar 14  g--aaa-a-c a-c...a-c- -gc-ggtaa- g-ac...... ......---a
 serovar 6  -c-caa-a-c --g...gtaa -g-------t --a-aac--g g--a-g----
serovar 11  cca-aacc-- -gccaa-tc- --c---t-c- ccaaaa-a-g a--...--g-
 serovar 7  cca-aacc-- -gccaa-tc- --c---t-c- ccaaaa-a-g a--...--g-
 serovar 9  ccaac-cc-g a-ccta-tc- --c---tac- ccaaaa-a-g a--...--g-
serovar 12  -a----ac-- ---g-c---a ---------c --t---a--- -a-gc-c-ga
serovar 13  -a----ac-- ---g-c---a ---------c --t---a--- -a-gc-c-ga
 serovar 4  -a----ac-- ---g-c---a ---------c --t---a--- -a-gc-c-ga
 serovar 2  -------a-- -t-......- ---------- -gt----a-- -----a----
 serovar 5  -------a-- -t-......- ---------- -gt----a-- -----a----
 serovar 8  -------a-- -t-......- ---------- -gt----a-- -----a----
serovar 10  ------tc-- -t-......- -c-------- -gt---t--- ----------
 Consensus  AGTGGTG-AA CAA-C-CAGC AAAACCAGGA AA-GGTGCAA CTACACAACC

            701                                                 750
 serovar 1  -ca-c----- ----aaca-c -------ta- a--aca-caa c--ggt-a--
 serovar 3  -cc-g----- ----aaca-c c-g----ta- a--aca-c-a g--ggt-a--
serovar 14  -cc-g----- ----aaca-c ------c-gg -a-a-a-c.. ..........
 serovar 6  ---taa--aa cc--g--a-g -------ta- a--ac--ggt -a-g--c---
serovar 11  --ttgt-a-- --t-t-ga-t ttagc-atgt ---t----aa g--a------
 serovar 7  --ttgt-a-- --t-t-ga-t ttagc-atgt ---t----aa g--a------
 serovar 9  -attgt-a-- --t-t-gagt ttagc-a-gt -a-t---caa ---a------
serovar 12  -aa-c----c --t-g--c-- c-agccc-g- aa-ac--ggc -atggt---a
serovar 13  -aa-c----c --t-g--c-- c-agccc-g- aa-ac--ggc -atggt---a
 serovar 4  -aa-c----c --t-g--c-- c-agccc-g- aa-ac--ggc -atggt---a
 serovar 2  ----agt--- g--ac--c-- --------ag --gt-a---t ---a--c---
 serovar 5  ----agt--- g--ac--c-- --------ag --gt-a---t ---a--c---
 serovar 8  ----agt--- g--ac--c-- --------ag --gt-a---t ---a--c---
serovar 10  ----agt--- tc-ac--c-c --------ag --gtt----t ---c--c---
 Consensus  AGGA-CAGGT AAAG-TA-AA AACCAGGA-A TGA-GCAAC- ACA-AAACAG

            751                                                 800
 serovar 1  a-ca-caacc -gg--a---- -a-ca-c--- ......gt-- -g---a-ca-
 serovar 3  a-ca-c-a-c -gg--a---- -a-cc-g--- ......gt-- -g---a-cc-
serovar 14  a-ca-c-a-c -gg--a---- -a-ca-c--- caggtaaag- -c---a-cc-
 serovar 6  gt-a--aacc -gg--a---- -----t-a-- aacc-ggt-- -g-------t
serovar 11  ---a--tta- -tta---ttt g-ctt-gt-- t-c---ta-- -g-cga-a-t
 serovar 7  ---a--tta- -tta---ttt g-ctt-gt-- t-c---ta-- -g-cga-a-t
 serovar 9  ---a--tta- -tta---ttt g-ctc-g--- t-c---ta-- -g-cga-a-c
serovar 12  ---gcc-a-- --aac---gc aat--t---a caagc--ag- -a-------c
serovar 13  ---gcc-a-- --aac---gc aat--t---a caagc--ag- -a-------c
 serovar 4  ---gcc-a-- --aac---gc aat--t---a caagc--ag- -a-------c
 serovar 2  g--gt-gt-- --c----a-- -------gt- g-g--a-t-c -a--------
 serovar 5  g--gt-gt-- --c----a-- -------gt- g-g--a-t-c -a--------
 serovar 8  g--gt-gt-- --c----a-- -------gt- g-g--a-t-c -a--------
serovar 10  g--gt-gttc --c----c-- -------gt- g-tc-a-t-c -c--------
 Consensus  CAA-AGC-GA AA-TACAGAA CCAGGAACAG -T-AACC-AA A-AACCAGGA

            801                                                 850
 serovar 1  cca----a-g a-c------- --......-- a--gaaca-c a-c----t--
 serovar 3  gca----a-g a-c---c-g- --......-- a--gaaca-c --g----t--
serovar 14  gca----a-g a-c------- --c-gg-aaa ---caaca-c --g----t--
 serovar 6  --a-aacc-g gt---g---- ---t-aa-aa cc-g---a-g a-c----t--
serovar 11  c-aaaatt-t t----tt-a- ttt----aaa --t--cga-- --aa--a-gt
```

```
  serovar 7  c-aaaatt-t t----tt-a- ttt----aaa --t--cga-- --aa--a-gt
  serovar 9  c-aaaatc-t tg---tt-a- ttt----aaa --t---ga-- --aa--a-gt
 serovar 12  --t----c-- ---gc-c-ga -aa-c-a--c a-tg---c-- --agccc-g-
 serovar 13  --t----c-- ---gc-c-ga -aa-c-a--c a-tg---c-- --agccc-g-
  serovar 4  --t----c-- ---gc-c-ga -aa-c-a--c a-tg---c-- --agccc-g-
  serovar 2  -gt---ga-- -t-c-a---- -----g---- ----c--c-- a-c------g
  serovar 5  -gt---ga-- -t-c-a---- -----g---- ----c--c-- a-c------g
  serovar 8  -gt---ga-- -t-c-a---- -----g---- ----c--c-- a-c------g
 serovar 10  -gt---tc-- -t-c------ -----g---- tc--c--c-c a-c------g
  Consensus  AA-GGTA-AA CAAAACAACC AGGAACTGGT GAAAGTA-AA CA-CAGGAAA

             851                                              900
  serovar 1  a------ca- c---...... -t-a--a-c- -ca-c----- ----a-caac
  serovar 3  a------c-- g---...... -t-a--a-c- -c--g----- ----a-caac
 serovar 14  a------ca- c---caggta a-ga-ca-c- -c--g----- ----a-caac
  serovar 6  a----c--gt -a--aa---- -t-a--a-cc -ggt-a--aa cc---t-aa-
 serovar 11  t--ttt--t- tt-a-tga-- at-a-tt-ag tg-g----c- g-ttt---t-
  serovar 7  t--ttt--t- tt-a-tga-- at-a-tt-ag tg-g----c- g-ttt---t-
  serovar 9  t--ctt--t- tt-a-t-a-- at-a-tt-ag tg------c- g-ttt--at-
 serovar 12  aa---c--gc -at--ta--a c--gccc--- --a-c----c --t--t-caa
 serovar 13  aa---c--gc -at--ta--a c--gccc--- --a-c----c --t--t-caa
  serovar 4  aa---c--gc -at--ta--a c--gccc--- --a-c----c --t--t-caa
  serovar 2  t-gtg--a-t ---aaa---- ---gt-gt-- ---t---aaa cc------t-
  serovar 5  t-gtg--a-t ---aaa---- ---gt-gt-- ---t---aaa cc------t-
  serovar 8  t-gtg--a-t ---aaa---- ---gt-gt-- ---t---aaa cc------t-
 serovar 10  t-gttc-a-t ---caa---- ---gt-gttc ---t---caa cc------t-
  Consensus  -GAACAAGCA ACAGG-CCAG GAA-AG-AGA AACAACAGGT AAAGGAAG-G

             901                                              950
  serovar 1  --c--g.... ..gtaa--a- c-acaac--g g--a-g---- -c--c-a-..
  serovar 3  c-g--g.... ..gtaa--a- c-acca---g g--a-g---- -cc-g-a-..
 serovar 14  --c--g-a-g t---ga-ca- c-acca---g g--a-g---- -c--c-a-ca
  serovar 6  --c--ggta- -g-------t --a-aac--g g--a-g---c -ggt-aa-aa
 serovar 11  -gtt--aa-- -ggtatttat --a-t-t-t- aatt--c-tt ---c-at-t-
  serovar 7  -gtt--aa-- -ggtatttat --a-t-t-t- aatt--c-tt ---c-at-t-
  serovar 9  tgtt--at-- -gg-a-ttat --a-taa-t- aatt--cttt ---tggtaa-
 serovar 12  c-agcc-a-- ---------c --t---a--- -a-gccc-g- ----c-a--c
 serovar 13  c-agcc-a-- ---------c --t---a--- -a-gccc-g- ----c-a--c
  serovar 4  c-agcc-a-- ---------c --t---a--- -a-gccc-g- ----c-a--c
  serovar 2  gtga---tac ---------- -gt----a-- ---c-----c -gg--gt---
  serovar 5  gtga---tac ---------- -gt----a-- ---c-----c -gg--gt---
  serovar 8  gtga---tac ---------- -gt----a-- ---c-----c -gg--gt---
 serovar 10  gtt----tac -c-------- -gt---t--- ---c-c---c -gg--gt---
  Consensus  AA-CAAC-GA AAAACCAGGA AA-GGTGCAA CTA-AAAACA AAAAAC-GGT

             951                                             1000
  serovar 1  ....---a-g ---a-ca-cc -------g-a c--c------ ......-ta-
  serovar 3  ....---a-g ---a-cc--c -------g-a c--cc-g--- ......-ta-
 serovar 14  -gtaaaga-c ---a-cc--c -------g-a c--c------ c-g--aa---
  serovar 6  cc-----a-g ---c--gta- --aacc-gg- ---------- -t-aa-a-cc
 serovar 11  ----t--gtt t-agt-ac-- -at----a-- ------tt-a a-gta-a-tt
  serovar 7  ----t--gtt t-agt-ac-- -at----a-- ------tt-a a-gta-a-tt
  serovar 9  ----t--gtt t-aat-ac-- -at----a-- ------tt-a a-gta-a--c
 serovar 12  a-t------- c-agccc--- -aaacc-ggc --t-gta--a c---ccc---
 serovar 13  a-t------- c-agccc--- -aaacc-ggc --t-gta--a c---ccc---
  serovar 4  a-t------- c-agccc--- -aaacc-ggc --t-gta--a c---ccc---
  serovar 2  ---ac----- ---c--g-ag t---g--ac- -c-a------ -------t--
  serovar 5  ---ac----- ---c--g-ag t---g--ac- -c-a------ -------t--
  serovar 8  ---ac----- ---c--g-ag t---g--ac- -c-a------ -------t--
 serovar 10  tc-ac----c ---c--g-ag t---tc-ac- -c-c------ -------ttc
  Consensus  GAAGGTACAA AAC-AG-AGA AGGTAAA-AT AAAGAACCAG GAAGTGGAGA

             1001                                            1050
  serovar 1  -gaaca-c-- -----t-a-- a-ca--a-c- -ggtaa--a- c-acaac--g
  serovar 3  -gaaca-cc- g----t-a-- a-ca------ -ggtaa--a- c-acca---g
 serovar 14  -caaca-cc- g----t-a-- a-ca--a-c- -gc-ggtaa- g-a.......
  serovar 6  -gg-aa-g-- -----t-a-- a-........ .......... ..........
 serovar 11  ----t---c- t-tcaac--- -c-gtggt-- ---....... ..........
  serovar 7  ----t---c- t-tcaac--- -c-gtggt-- ---....... ..........
  serovar 9  tt--aa---- --t-aa---- -t-gca---a -gg....... ..........
```

```
serovar 12  --aac--ggc aat--t---a c--gc----a ---------c ------a--.
serovar 13  --aac--ggc aat--t---a c--gc----a ---------c ------a--.
 serovar 4  --aac--ggc aat--t---a c--gc----a ---------c ------a--.
 serovar 2  ----a----- -----a-gt- -tga-a-ta- ---------- -g-----a--
 serovar 5  ----a----- -----a-gt- -tga-a-ta- ---------- -g-----a--
 serovar 8  ----a----- -----a-gt- -tga-a-ta- ---------- -g-----a--
serovar 10  ----a--c-- -----a-gt- -ttc-a-ta- -c-------- -g----t---
 Consensus  AACT-CAAAA CCAGG-ACAG GAA-ACCAGC AAAACCAGGA AATGGTGCAA

            1051
 serovar 1  g-------
 serovar 3  g-------
serovar 14  ........
 serovar 6  ........
serovar 11  ........
 serovar 7  ........
 serovar 9  ........
serovar 12  ........
serovar 13  ........
 serovar 4  ........
 serovar 2  --......
 serovar 5  --......
 serovar 8  --......
serovar 10  --......
 Consensus  CTAAAGAA
```

Table 2.2. Specificity and expected lengths of ureaplasma genotyping primer pairs.

Primer pairs*	Specificity	Length of amplicons (base pairs)
UMS-125-UMA226	Ureaplasmas	403/404 for UP; 448 for UU
UMS-125-UMA222	UP	396/397
UMS-170-UMA263	UU	485
UMS3S-UMA269	UP 3	408
UMS3S-UMA314A	UP 3	576
UMS3S-UMA314A’	UP 3	578
UMS14S-UMA269	UP 14	410
UMS14S-UMA314A	UP 14	578
UMS14S-UMA314A’	UP 14	580
UMS-83-UMA1A	UP 1	582
UMS-83-UMA1A’	UP 1	584
UMS-83-UMA269’	UP 1	408
UMS-54-UMA6A	UP 6	550
UMS-54-UMA6A’	UP 6	554
UMS-54-UMA269’	UP 6	380
UMS-61-UMA2A	UU A	572
UMSUS-UMA2A	UU A	317
UMS-61-UMA10A	UU B	572
UMSUS-UMA10A	UU B	317
UMS-61-UMA4A1	UU C	545
UMSUS-UMA4A1	UU C	290
UMS-61-UMA4A2	UU C	547
UMSUS-UMA4A2	UU C	292
UMS-61-UMA9A1	UU D	553
UMSUS-UMA9A1	UU D	298
UMS-61-UMA9A2	UU D	555
UMSUS-UMA9A2	UU D	300

UMS-61-UMA7A1	UU E	338
UMS-61-UMA7A2	UU E	551
UMSUS-UMA7A2	UU E	296
UMS-61-UMA7A3	UU E	554
UMSUS-UMA7A3	UU E	299
UMS-61-UMA2A1	UU A/B	544
UMSUS-UMA2A1	UU A/B	289
UMS-61-UMA2A2	UU A/B	547
UMSUS-UMA2A2	UU A/B	292
UMS-61-UMA219	UU A–D	338
UMS-112-UMA194	UU A/D/E	357
UMS-112’-UMA194’	UU B/C	357

Notes.

* See Table 2.1. for primer sequences and genotype explainations.

2.4.4. Algorithm for serovar or genotype identification by sequencing and PCR.

A. *U. parvum* serovar identification:

- A) By sequencing:
  - a) Sequencing UMS-125-UMA222 amplicons;
  - b) Using Figure 2.1. or multiple sequence aligment software (*Pileup* and *Pretty*) to identify *U. parvum* serovars.
- B) By serovar-specific PCR, using primer pairs listed in Table 2.1.

B. *U. urealyticum* genotypes identification:

- A) By sequencing:
  - a) Sequencing UMS-170-UMA263 amplicons;
  - b) Using Figure 2.1. or multiple sequence aligment software (*Pileup* and *Pretty*) to identify *U. urealyticum* genotypes.
- B) By genotype-specific PCR, using primer pairs listed in Table 2.1. – only when necessary.

C. For most isolates or clinical specimens (suitable for most studies):

- a) Sequencing UMS-125-UMA226 amplicons;
- b) Using Figure 2.1. or multiple sequence aligment software (*Pileup* and *Pretty*) to identify *U. parvum* serovars or *U. urealyticum* subtypes (for 3 subtypes) (Kong *et al*., 2000).

2.5. DISCUSSION

We have described methods previously that distinguish the two ureaplasma species (Kong *et al*., 2000; chapter 1). Our previous study showed that homology between sequences of the 16S rRNA genes, 16S-23S rRNA intergenic spacer regions and urease gene subunits of serovars within each proposed species was high and these regions were not suitable for further genotyping (Kong *et al*., 2000). In this study, we sequenced the 5'-ends (including partial repetitive regions) and upstream regions of *mba* for all 14 ureaplasma serovars. Our aim was to define sequence differences

that would allow further molecular serovar or genotype identification of *U. parvum* and *U. urealyticum* (Kong *et al*., 1999a, 2000).

Our previous studies showed only three base differences between sequences of the partial 5'-ends and upstream regions of *mba* of *U. parvum* serovars 3 and 14 (Kong *et al*., 1999a, b). In this study we showed more numerous differences in nucleotide sequences immediately upstream of the repetitive regions and in the repetitive units themselves, between *U. parvum* serovars, which allowed all of them, including serovars 3 and 14, to be differentiated. Based on our previous study of partial 5'-ends and upstream regions of *mba* of *U. urealyticum* (Kong *et al*., 1999b), serovar 10 is closely related to serovars 4, 12, and 13. However, differences immediately upstream and in the repetitive units allowed serovar 10 to be separated from serovars 4/12/13. Similarly, serovar 9 was closely related to serovars 2, 5, and 8, based on sequences of the 5'-ends of *mba* (Kong *et al*., 1999b) but deletion of the repetitive region in serovar 9 allowed it to be differentiated from serovars 2/5/8. This finding is supported by the recent development of a monoclonal antibody against *U. urealyticum* serovar 9, that cross-reacts only minimally with serovar 2 (Naessens *et al*., 1998).

The present study also showed that there were 22 bases at the 5'-ends and upstream regions of *mba* of the ten serovars of *U. urealyticum,* upstream of the repetitive regions, which helped to differentiate the five *mba* genotypes. More than half of these differences were between *mba* genotype E (serovars 7/11) and the other four *mba* genotypes. Serovars 7/11 were similar to serovar 9 in that the repetitive sequences were deleted. However, their sequences differed by 14 bases at the 5'-ends of *mba* and the *mba* upstream (1-639), and 58 within sequences that corresponded with those of the repetitive regions of *mba* (640-1033), of the other serovars (Figure 2.1.). Serovars in *mba* genotypes A (serovars 2, 5 and 8), C (serovars 4, 12 and 13), and E (serovars 7 and 11) could not be differentiated further on the basis of these sequences.

Further work is required to identify other genes or other regions of the *mba* that may be used to differentiate *U. urealyticum* serovars within *mba* genotypes A, C and E. However, on the basis of our data, we suggest that genetic and antigenic differences between some serovars are so minor that further subdivision into serovars might be artificial and/or unnecessary. These data provide a better understanding of the molecular basis of serotype differentiation. Based on the sequence analysis, we designed a series of serovar- or *mba* genotype-specific primer pairs, and a practical algorithm which, after further evaluation, could be used for further study of the relationships between serovar/genotypes and diseases, if there are any.

CHAPTER 3

GBS MOLECULAR SEROTYPE IDENTIFICATION

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; 40:216-226. (see Appendices section)

Statement of Joint Authorship

Kong, F. (candidate)

Did all the molecular serotype identification 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.

3.1. SUMMARY

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. Ongoing surveillance, to monitor GBS serotype distribution, will be needed to guide the development and use of GBS conjugate vaccines. Based on previously published sequences of the capsular polysaccharide synthesis (*cps*) gene clusters, we designed sequencing primers to further define partial *cps* gene clusters for eight of the nine GBS serotypes (Ia to VII). Subsequently, we designed and evaluated primers to identify serotypes Ia, Ib, III, IV, V and VI directly by PCR and all eight serotypes (Ia to VII) by sequence heterogeneity. 206 clinical GBS isolates were used to compare our molecular serotype (MS) identification method with conventional serotyping (CS). All clinical isolates were assigned an MS, compared with 188 of 206 (91.3%) using antisera. A small number of isolates (serosubtypes III-3 and III-4) showed different serotype specificities between PCR and sequencing, but PCR results correlated with CS. Overall agreement between MS and CS, for isolates for which results of both were available, was 100% (188/188). MS is a specific and practical alternative to conventional GBS serotyping and will facilitate epidemiological studies.

3.2. 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; Tyrrell *et al*., 2000). The incidence of neonatal GBS sepsis has been reduced in recent years by the use of intrapartum antibiotic prophylaxis (Schrag *et al*., 2000), but there are many problems with this approach (Levine *et al*., 1999). In future, vaccination is likely to be preferred and there has been considerable progress in development of conjugate polysaccharide GBS vaccines (Paoletti *et al*., 1999).

Before the introduction of vaccines, extensive epidemiological studies will be required to assess, not only the burden of disease, but also the distribution of GBS serotypes to determine the optimal formulation of vaccine antigens (Harrison *et al*., 1998). Serotype distribution based on one geographic location or small numbers of patients may not be generally applicable (Hickman *et al*., 1999; Lin *et al*., 1998). Continued monitoring will be necessary to assess the suitability of combinations of GBS vaccine antigens for different target populations in different geographic locations (Harrison *et al*., 1998; Hickman *et al*., 1999).

Nine capsular polysaccharide GBS serotypes have been described (Harrison *et al*., 1998; Hickman *et al*., 1999). Various serotyping methods have been used, including immuno-precipitation (Wilkinson & Moody, 1969), enzyme immunoassay (Holm & Hakansson, 1988), coagglutination (Hakansson *et al*., 1992), counter-immunoelectrophoresis, and capillary precipitation (Triscott & Davis, 1980), latex agglutination (Zuerlein *et al*., 1991), fluorescence microscopy (Cropp *et al*., 1974) and inhibition-ELISA (Arakere *et al*., 1999). These methods are labour-intensive and require high-titer serotype-specific antisera, which are expensive and difficult to make and commercially available for only six serotypes, serotypes Ia to V (Arakere *et al*., 1999). Molecular genotyping, such as pulsed-field gel electrophoresis (Rolland *et al*., 1999), restriction endonuclease analysis (Nagano *et al*., 1991) are useful for epidemiological studies but do not generally identify serotypes. Molecular serotype (MS) identification methods are theoretically attractive because of their potentially high discriminatory power and reproducibility (Sellin *et al*., 2000). PCR-based assays have been used to detect and genotype GBS, but further development is needed to make them practicable for use in serotype identification (Ke *et al*., 2000; Sellin *et al*., 2000).

In this study, we used published sequences of the capsular polysaccharide synthesis (*cps*) gene clusters of GBS serotypes Ia and III (Chaffin *et al*., 2000; Yamamoto *et al*., 1999), and our own sequencing results, to analyse partial *cps* gene cluster sequences of eight serotypes (serotypes Ia to VII) and to develop a MS method.

Recently published sequences of *cps* gene clusters of serotypes Ib (Miyake *et al.,* submitted into GenBank, 2001 [GenBank accession number AB050723]), IV, V and VI (McKinnon *et al.,* submitted into GenBank, 2001 [GenBank accession numbers AF355776, AF349539, AF337958, respectively]) confirmed our sequence analysis. Subsequently, we used these sequences to design PCR primers specific for serotypes Ib, IV, V and VI.

3.3. MATERIALS AND METHODS

3.3.1. GBS reference strains and clinical isolates.

A panel of nine GBS serotypes (Ia to VIII) was kindly provided by Dr Lawrence Paoletti, Channing Laboratory, Boston, USA (reference panel 1). Dr Diana Martin, Streptococcus Reference Laboratory, at Environmental Science and Research (ESR), Wellington, New Zealand, provided another panel of nine international reference GBS type-strains including serotypes Ia to VI (reference panel 2) (Table 3.1.). In addition, we tested isolates from 205 clinical cases including 146 which had been referred from various laboratories in New Zealand for serotyping and 59 isolated from normally sterile sites over a period of ten years in one diagnostic laboratory in Sydney. One culture was subsequently shown to be mixed, so 206 different isolates were examined. Conventional serotyping (CS) was performed at the Streptococcus Reference Laboratory, at ESR, Wellington, New Zealand, and MS at the Centre for Infectious Diseases and Microbiology (CIDM), Insititute of Clinical Pathology and Medical Research (ICPMR), Sydney, Australia.

The two panels of GBS reference strains and 63 selected clinical isolates were studied in more detail, by sequencing >2200 base pairs (bp) of each to identify appropriate sequences for use in MS. These and the remaining clinical isolates were then used to evaluate the MS method and compare results with those of CS. Typing by both methods was done initially without knowledge of results of the other.

Bacterial isolates were retrieved from storage by subculture on blood agar plates (Columbia II agar base supplemented with 5% horse blood) and incubated overnight at 37°C.

3.3.2. Conventional serotyping (CS).

CS was performed using standard methodology (Wilkinson & Moody, 1969). Briefly, an acid-heated (56°C) extract was prepared for each isolate and the serotype determined by immuno-precipitation of type-specific antiserum in agarose. An isolate was considered positive for a particular serotype when the precipitation occurring formed a line of identity with that of the control strain. Antisera used were prepared at ESR in rabbits against serotypes Ia, Ib, II, III, IV and V. Fourteen selected isolates, including six that were nontypable using antisera against serotypes I-V, six that initially gave discrepant results between CS and MS and two separate isolates from a mixed culture, were kindly tested using antisera against all serotypes by Abbie Weisner and Dr Androulla Efstratiou at Central Public Health Laboratory, Colindale, London, UK.

3.3.3. Oligonucleotide primers.

Four previously published oligonucleotide primers, and a series of new primers designed by us were used to sequence the genes of interest, namely 16S-23S rRNA intergenic spacer region and partial *cps* gene cluster, or to amplify unique sequences of GBS *cps* clusters of each serotype. The sequences, target sites and melting temperatures (*T*m) of primers used in this study are shown in Table 3.2. Some were designed with high melting temperatures to be used in rapid cycle PCR (Ke *et al*., 2000).

Table 3.1. Reference strainss used in GBS molecular serotype identification study.

Lab strain number	Source	Serotype	MS/subtype	GenBank numbers
Reference panel 1[a]				
090	Channing	Ia	Ia	AF332893
H36B	Channing	Ib	Ib	AF332903
18RS21	Channing	II	II	AF332905
M781	Channing	III	III-2[c]	AF332896
3139	Channing	IV	IV	AF332908
CJB 111	Channing	V	V	AF332910
SS1214	Channing	VI	VI	AF332901
7271	Channing	VII	VII	AF332913
JM9 130013	Channing	VIII	VIII	-
Reference panel 2[b]				
NZRM 908 (NCDC SS615)	ESR	Ia	Ia	AF332894
NZRM 909 (NCDC SS618)	ESR	Ib	Ib	AF332904
NZRM 910 (NCDC SS700)	ESR	Ia	Ia	AF332914
NZRM 911 (NCDC SS619)	ESR	II	II	AF332906
NZRM 912 (NCDC SS620)	ESR	III	III-3[c]	AF332897
NZRM 2217 (Prague 25/60)	ESR	Non-typable (R)	II	AF332907
NZRM 2832 (Prague 1/82)	ESR	IV	IV	AF332909
NZRM 2833 (Prague 10/84)	ESR	V	V	AF332911
NZRM 2834 (Prague 118754)	ESR	VI	VI	AF332902

Notes.

a. Reference panel 1: supplied by Dr Lawrence Paoletti, Channing Laboratory, Boston, USA.

b. Reference panel 2: supplied by Dr Diana Martin, ESR, Porirua, Wellington, New Zealand.

c. MS III subtypes based on sequence heterogeneity; see text for more detail.

3.3.4. DNA preparation and PCR.

Five individual GBS colonies or a sweep of culture were sampled using a disposable loop and resuspended in 200 μl of digestion buffer (10mM Tris-HCl [pH 8.0], 0.45% Triton X-100 and 0.45% Tween 20) in 2 ml Eppendorf tubes. The tubes containing GBS suspension were heated at 100°C (dry block heater or water bath) for ten minutes then quenched on ice and centrifuged at 14,000 rpm (16,000x *g*) in an Eppendorf Centrifuge 5415C for two minutes to pellet the cell debris. 5 μl of each supernatant containing extracted DNA was used as template for PCR (Mawn *et al*., 1993).

PCR systems (25 μl for detection only, 50 μl for detection and sequencing) were used as previously described (Kong *et al*., 1999). The denaturation, annealing and elongation temperatures and times used were 96°C for 1 second, 55-72°C (according to the primer *T*m values or as previously described) for 1 second and 74°C for 1 to 30 seconds (according to the length of amplicons), respectively, for 35 cycles.

10 μl of 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 serotype identification, the presence of PCR amplicons of expected length, shown by ultraviolet transillumination, were accepted as positive. For sequencing, 40 μl of PCR products were further purified by polyethylene glycol precipitation method (Ahmet *et al*., 1999).

3.3.5. Sequencing.

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.

Table 3.2. Primers used in GBS molecular serotype identification study.

Primer	Target gene name	*Tm* °C[a]	GenBank numbers	Sequence[b-d]
CFBS	*cfb*	56.7	X72754	**328**GAT GTA TCT ATC TGG AAC TCT AGT G**352**
Sag59[e]	*cfb*	77.4	X72754	**350**GTGGCTGGTGCATTGTTAT TTT CAC CAG CTG TAT TAG AAG TA**391**
Sag190[e]	*cfb*	76.8	X72754	**545**CATTAACCGGTTTTTCATAATCT GTT CCC TGA ACA TTA TCT TTG AT**500**
CFBA	*cfb*	63.2	X72754	**568**TTT TTC CAC GCT AGT AAT AGC CTC**545**
16SS	16S rRNA	69.3	AB023574	**1441**GCC GCC TAA GGT GGG ATA GAT G**1462**
23SA	23S rRNA	65.7	X68427	**70**CGT CGT TTG TCA CGT CCT TC**51**
DSF2[f]	16S rRNA	75.9	AB023574	**975**CATCCTTCTGACC GGC CTA GAG ATA GGC TTT CT**1007**
DSR1[f]	16S rRNA	81.5	AB023574	**1250**CGTCACCGG CTT GCG ACT CGT TGT ACC AA**1222**
cpsES3	*cpsE*	71.5	AB028896 (Ia), AF163833 (III)	**6410/6020**GTT AGA TGT TCA ATA TAT CAA TGA ATG GTC TAT TTG GTC AG**6450/6060**
cpsFS	*cpsF*	75.0	AB028896 (Ia), AF163833 (III)	**6777/6387**CAT CTG GTG CCG CTG TAG CAG TAC CAT T**6804/6414**
cpsFA	*cpsF*	73.2	AB028896 (Ia), AF163833 (III)	**6859/6469**GTC GAA AAC CTC TAT A/GT A AAC/T GGT CTT ACA A/GCC AAA TAA CTT ACC**6815/6425**
cpsGA	*cpsG*	54.7	AB028896 (Ia), AF163833 (III)	**7162/6772**AAG/C AGT TCA TAT CAT CAT ATG AGA G **7138/6748**
cpsGA1	*cpsG*	74.5	AB028896 (Ia), AF163833 (III)	**7199/6809**CCG CCA/G TGT GTG ATA ACA ATC TCA GCT TC**7171/6781**
IacpsHS1	*cpsH*	77.9	AB028896 (Ia)	**8463**GGC CTG CTG GGA TTA ATG AAT ATA GTT CCA GGT TTG C**8499**

cpsIA	*cpsI*	70.3	AB028896 (Ia), AF163833 (III)	**8816/8312**GTA TAA CTT CTA TCA ATG GAT GAG TCT GTT GTA GTA CGG**8778/8274**
IbcpsIS	*cpsI*	71.1	AB050723 (Ib)	**4116**GAT AAT AGT GGA GAA ATT TGT GAT AAT TTA TCT CAA AAA GAC G**4158**
IbcpsIA1	*cpsI*	78.6	AB050723 (Ib)	**4638**CCT GAT TCA TTG CAG AAG TCT TTA CGA TGC GAT AGG TG**4601**
IIIcpsHS	*cpsH*	72.1	AF163833 (III)	**7672**GAA TAC TAT TGG TCT GTA TGT TGG TTT TAT TAG CAT CGC**7710**
IVcpsHS1	*cpsH*	71.2	AF355776 (IV)	7887CCC AAG TAT AGT TAT GAA TAT TAG TTG GAT GGT TTT TGG7925
IVcpsMA	*cpsH*	80.7	AF355776 (IV)	**8265**GGG TCA ATT GTA TCG TCG CTG TCA ACA AAA CCA ATC AAA TC**8225**
VcpsHS2	*cpsH*	74.0	AF349539 (V)	**7871**CCC AGT GTG GTA ATG AAT ATT AGT TGG CTA GTT TTT GG**7908**
VcpsMA	*cpsM*	73.1	AF349539 (V)	**8244**CCC CCC ATA AGT ATA AAT AAT ATC CAA TCT TGC ATA GTC AG**8204**
VIcpsHS1	*cpsH*	77.2	AF337958 (VI)	**7767**CCT TAT TGG GCA AGG TAT AAG AGT TCC CTC CAG TGT G**7803**
VIcpsIA	*cpsI*	74.5	AF337958 (VI)	**8126**GAA GCA AAG ATT CTA CAC AGT TCT CAA TCA CTA ACT CCG**8088**

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" refer to the start points "1" of correspondent gene GenBank accession numbers).

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

d. Letters behind "/" indicate alternative nucleotides in different serotypes.

e. From Ke *et al*., 2000.

f. From Ahmet *et al*., 1999.

3.3.6. Multiple sequence alignments.

Multiple sequence alignments were performed with *Pileup* and *Pretty* programs in Multiple Sequence Analysis program group. Both programs are provided in WebANGIS (Website: http://www1.angis.org.au/pbin/WebANGIS/wrapper.pl), ANGIS (Australian National Genomic Information Service).

3.3.7. Nucleotide sequence accession numbers.

The new sequence data reported in this chapter were deposited into the GenBank Nucleotide Sequence Databases with the following accession numbers: AF332893-AF332917, AF363032-AF363060, AF367973, AF381030 and AF381031 for partial *cps* gene clusters for two panels of reference strains (Table 3.1.) and selected representative clinical isolates. Previously reported sequence data used in this chapter have appeared in the GenBank Nucleotide Sequence Databases with the following accession numbers: AB023574 for the 16S rRNA gene, U39765, L31412 for the 16S-23S rRNA intergenic spacer regions, X68427 for the *S. oralis* 23S rRNA gene, X72754 for the *cfb* gene, AB028896 for the *cps* gene cluster for serotype Ia, AB050723 for the partial *cps* gene cluster for serotype Ib, AF163833 for the *cps* gene cluster for serotype III, AF355776 for the *cps* gene cluster for serotype IV, AF349539 for the *cps* gene cluster for serotype V, and AF337958 for the *cps* gene cluster for serotype VI.

3.4. RESULTS

3.4.1. PCR.

With two exceptions, all GBS-specific primer pairs produced amplicons of the expected size from all reference strains and clinical isolates tested (Table 3.2.). The exceptions were Sag59-Sag190 and CFBS-CFBA. Both target the *cfb* gene, but

failed to produce amplicons from one clinical isolate, despite repeated attempts. We assumed that this isolate either lacked the *cfb* gene or that the gene was present in a mutant form. It has been suggested previously that PCR targeting the *cfb* gene will not identify all GBS isolates (Hassan *et al*., 2000) and that another primer pair based on 16S rRNA gene, DSF2-DSR1 (Ahmet *et al*., 1999) was not entirely specific. Therefore, in this study, we used both primer pairs (DSF2-DSR1 and Sag59-Sag190) to confirm all the isolates were GBS.

3.4.2. Sequence heterogeneity at the 3'-end of *cpsD-cpsE-cpsF*-and the 5'-end of *cpsG*.

Using a series of primers targeting the 3'-end of *cpsD-cpsE-cpsF*-and the 5'-end of *cpsG*, we amplified and sequenced 2226- or 2217 bp − depending on the presence or absence of a nine-base repetitive sequence − from both panels of reference strains (serotypes Ia to VII) and 63 selected clinical isolates. Representative sequences were deposited into GenBank. See Table 3.1. for GenBank accession numbers of reference panel strains.

3.4.3. Repetitive sequence.

At the 3'-end region of *cpsD*, we found a nine-base repetitive sequence (TTA CGG CGA) in most isolates of MS Ia and II, some of MS III, all of MS IV, V, and VII, but none of the isolates of MS Ib or VI examined. (Table 3.3.). The presence or absence of this repetitive sequence can be used to further subtype MS Ia, II and III (see below).

Table 3.3. The heterogeneity of eight GBS serotypes in the regions of the 3'-end of *cpsD*- and the 5'-end of *cpsE*.

Sites[a]	Ia	Ib	II/III-4[d]	III	IV	V	VI	VII	Specificity
cpsD									
62	G	A	G[e]	A	A	A	A	G	Ia, II, VII
78-86 repetitive sequence - TTACGGCGA *cpsD*-*cpsE* spacer	- Ia-2[b]; + Ia-1[b]	-	- II-2[c, e]; + II-1[c]	- III-2[d]; + III-1[c], III-3[c]	+	+	-	+	See text
138	G	G	G	G	G	A[f]	G	G	V
139	G	G	G	A III-2; G III-1, III-3	G	G	G	G	III-2
144	T	T	T	G III-2; T III-1, III-3	T	T	T	T	III-2
cpsE									
198	A	C	A[e]	A	C	C[f]	A	A	Ib, IV, V
204	G	G	G	A III-2, III-3; G III-1	G	G	G	G	III-2, III-3
211	T	T	T	T	T	T	G	T	VI
218	C	C	C	C	C	C	T	C	VI
240	T	T	T	T	T	T	C	T	VI

249	T	C	T[e]	T	C	C[f]	T	T	Ib, IV, V
300	C	C	C	T III-2; C III-1, III-3	C	C	C	C	III-2
321	C	C	C	T III-1; C III-2, III-3	C	C	C	C	III-1
419	T	C	T[e]	T	T	T	T	T	Ib
429	A	T	A[e]	T	T	T	T	A	Ia, II, VII
437	C	C	C; T III-4	C	C	C	C	T	VII, III-4
457	T	A	C[e]	A	A	A	A	C	Ia, II, VII
466	G	G	G	G	A	G	G	A	IV
486	G	A	A	G III-3; A III-2, III-1	A	A	A	A	Ia, III-3
602	G	G	A[e]	G	G	G	G	A	II, VII
606	T	T	T	T	T	T	C	T	VI
627	T	C	C	C	C	C	C	C	Ia
636	C	T	T	C III-1; T III-2, III-3	T	T	T	T	Ia, III-1
645	C	T	C[e]	C	T	T	C	C	Ib, IV, V
803	A	A	A	A	A	A	T	A	VI

971	C	T	T	C	C	C	T	T	Ia, III, IV, V
1026	A	G	G	G III-2, III-1; A III-3	A	A	G	G	Ia, III-3, IV, V
1044	T	T	T	T	T	T	C	T	VI
1173	A	G	A	A	A	A	A	A	Ib
1194	C	C	C	A	A	C	A	C	III, IV, VI
1251	G	G	G	G	G	G	A	G	VI
1278	A	A	A	A	A	G	A	A	V

Notes.

a. Numbering start point “1” refers to the start point “1” of GenBank accession number AF332908 (for serotype IV reference strain 3139).

b. Repetitive sequence: Ia-1 present (+); Ia-2 absent (-) (see text).

c. Repetitive sequence: II-1 present (+); II-2 absent (-) (see text).

d. Repetitive sequence: III-1 and III-3 present (+); III-2 absent (-); III-4 variable (see text)

e. One CS II strain has mutations at the 9 sites (see text).

f. At positions 138, 198, and 249, one CS V reference strain (Prague 10/84) is identical with corresponding sequence in GenBank (GenBank accession number AF349539), the sequences are G, A and T, respectively; another CS V reference strain (CJB 111) and all the other sequenced CS V strains are identical, the sequences are A, C and C, respectively.

3.4.4. Intraserotype heterogeneity.

In general, intraserotype heterogeneity was low; there were minor random variations in a few isolates of all serotypes except MS III, in which the intraserotype heterogeneity was more complex. MS III could be divided into four sequence subtypes on the basis of heterogeneity at 22 positions (62, 139, 144, 204, 300, 321, 429, 437, 457, 486, 602, 636, 971, 1026, 1194, 1413, 1501, 1512, 1518, 1527, 1629 and 2134) and the presence or absence of the repetitive sequence (positions 78-86) (Figure 3.1. and Table 3.3.).

Among 60 MS III isolates (58 clinical isolates and two reference strains), subtypes III-1 (30 isolates) and III-2 (22 isolates) were predominant. The repetitive sequence was present in subtype III-1 but not III-2; there were differences at seven other sites (positions 139, 144, 204, 300, 321, 636 and 1629) (Table 3.3.).

There were five isolates belonging to subtype III-3, which contained the repetitive sequence and were identical with subtype III-1 at three variable sites (positions 139, 144 and 300) and with subtype III-2 at four (positions 204, 321, 626 and 1629). Subtype III-3 differed from both subtypes III-1 and III-2 at seven sites (positions 486, 1026, 1413, 1512, 1518, 1527 and 2134). These seven sites in subtype III-3 were identical with the corresponding sites of MS Ia.

There were three subtype III-4 isolates, whose sequences were nearly identical with the corresponding sequence of MS II. The only exception was at position 437, where the nucleotide was T in subtype III-4 (as in MS VII), and C in MS II. This difference can be used (in addition to PCR, see below) to differentiate subtype III-4 from MS II. Two subtype III-4 isolates contained the repetitive sequence, and the other did not. Because of the small number of subtype III-4 isolates, we did not use the repetitive sequence to subtype them further.

Figure 3.1. Molecular serotype identification based on the sequence heterogeneity of the 790 bp fragment (positions 1437-2226).

```
                   1401                                                1450
   Subtype III-2   ---------- ---------- ---------- ---------- ----------
     Serotype VI   ---------- ---------- ---------- ---------- ----------
     Serotype Ib   ---------- ---------- ---------- ---------- ----------
Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
    Serotype VII   ---------- ---------- ---------- ---------- ----------
Serotype Ia/III-3  ---------- --c------- ---------- ---------- ----------
   Subtype III-1   ---------- ---------- ---------- ---------- ----------
     Serotype IV   ---------- ---------- ---------- ---------- ----------
      Serotype V   ---------- ---------- ---------- ---------- ----------
       Consensus   TAGAAATAAT ATTACTGATT TTGATGAAAT CGTAAAGTTA GATGTTCAAT

                   1451                                                1500
   Subtype III-2   ---------- ---------- ---------- ---------- ----------
     Serotype VI   ---------- ---------- ---------- ---------- ----a-----
     Serotype Ib   ---------- ---------- ---------- ---------- ---------g
Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
    Serotype VII   ---------- ---------- ---------- ---------- ----------
Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
   Subtype III-1   ---------- ---------- ---------- ---------- ----------
     Serotype IV   ---------- ---------- ---------- ---------- ----------
      Serotype V   ---------- ---------- ---------- ---------- ----------
       Consensus   ATATCAATGA ATGGTCTATT TGGTCAGATA TTAAGATTAT TCTCCTAACA
                            cpsES3
                   1501                                                1550
   Subtype III-2   ---------- -t-----c-- ---------- ---------- ----------
     Serotype VI   ---------- -t-----c-- ---------- ---------- ----------
     Serotype Ib   ---------- -t-----c-- ---------- ---------- ----------
Serotype II/III-4  t--------- ---------- ---------- ---------- ----------
    Serotype VII   t--------- ---------- ---------- ---------- ----------
Serotype Ia/III-3  ---------- ---------- ------t--- ---------- ----------
   Subtype III-1   ---------- -t-----c-- ---------- ---------- ----------
     Serotype IV   ---------- ---------- ------t--- ---------- ----------
      Serotype V   ---------- -t-----c-- ---------- ---------- ----------
       Consensus   CTAAAGGTAG TCTTACTTGG GACAGGAGCT AAGTAAAGGT AAGGTTTGAA
                                                    cpsE |
                   1551                                                1600
   Subtype III-2   ---------- ---------- ---------- ---------- ----------
     Serotype VI   ---------- ---------- ---------- ---------- ----c-----
     Serotype Ib   ---------- ---------- ---------- ---------- ----c-----
Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
    Serotype VII   ---------- ---------- ---------- ---------- ----------
Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
   Subtype III-1   ---------- ---------- ---------- ---------- ----------
     Serotype IV   ---------- ---------- ---------- ---------- ----------
      Serotype V   ---------- ---------- ---------- ---------- ----------
       Consensus   AGGAATATAA TGAAAATTTG TCTGGTTGGT TCAAGTGGTG GTCATCTAGC
                            | cpsF
                   1601                                                1650
   Subtype III-2   ---------- ---------- ---------- ---------- ----------
     Serotype VI   ---------- ---------- ---------- ---------- ----------
     Serotype Ib   ---------- ---------- ---------- ---------- ----------
Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
    Serotype VII   ---------- t--------t ---------- ---------- ----------
Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
   Subtype III-1   ---------- ---------- --------a- ---------- ----------
     Serotype IV   ---------- ---------- ------t--- ---------- ----------
      Serotype V   ---------- ---------- ---------- ---------- ----------
       Consensus   ACACTTGAAC CTTTTGAAAC CCATTTGGGA AAAAGAAGAT AGGTTTTGGG

                   1651                                                1700
   Subtype III-2   ---------- ---------- ---------- ---------- ----------
     Serotype VI   ---------- ---------- ---------- ---------- ----------
     Serotype Ib   ----t----- ---------- ---------- ---------- ----------
Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
```

```
     Serotype VII  ---------- ---------- ---------- ---------- ----------
Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
     Subtype III-1 ---------- ---------- ---------- ---------- ----------
      Serotype IV  ---------- ---------- ---------- ---------- ----------
       Serotype V  ---------- ---------- ---------- ---------- ----------
        Consensus  TAACCTTTGA TAAAGAAGAT GCTAGGAGTA TTCTAAGAGA AGAGATTGTA

                   1701                                              1750
    Subtype III-2  ---------- ---------- ---------- ---------- ----------
      Serotype VI  ---------- ---------- ---------- ---------- ----------
      Serotype Ib  ---------- ---------- ---------- ---------- ----------
Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
     Serotype VII  ---------- ---------- ---------- ---------- ----------
Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
    Subtype III-1  ---------- ---------- ---------- ---------- ----------
      Serotype IV  ---------- ---------- ---------- ---------- ----------
       Serotype V  ---------- ---------- ---------- ---------- ----------
        Consensus  TATCATTGCT TCTTTCCAAC AAACCGTAAT GTCAAAAACT TGGTAAAAAA

                   1751                                              1800
    Subtype III-2  ---------- ---------- ---------- ---------- ----------
      Serotype VI  ---------- ---------- ---------- ---------- ----------
      Serotype Ib  ---------- ---------- ---------- ---------- ----------
Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
     Serotype VII  ---------- ---------- ---------- ---------- ----------
Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
    Subtype III-1  ---------- ---------- ---------- ---------- ----------
      Serotype IV  ---------- ---------- ---------- ---------- ----------
       Serotype V  ---------- ---------- ---------- ---------- ----------
        Consensus  TACTATTCTA GCTTTTAAGG TCCTTAGAAA AGAAAGACCA GATGTTATCA

                   1801                                              1850
    Subtype III-2  ---------- ---------- ---------- ---------- ----------
      Serotype VI  ---------- ---------- ---------- ---------- ----------
      Serotype Ib  ---------- ---------- ---------- ---------- ----------
Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
     Serotype VII  ---------- ---------- ---------- ---------- ----------
Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
    Subtype III-1  ---------- ---------- ---------- ---------- ----------
      Serotype IV  ---------- ---------- ---------- -t-------- ----------
       Serotype V  ---------- ---------- ---------- ---------- ----------
        Consensus  TATCATCTGG TGCCGCTGTA GCAGTACCAT TCTTTTATAT TGGTAAGTTA
                                 cpsFS
                   1851                                              1900
    Subtype III-2  ---------- ---------- ---------- ---------- ----------
      Serotype VI  ---------- ---------- ---------- ---------- ----------
      Serotype Ib  -----c---- ---------- ---------- ---------- ----------
Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
     Serotype VII  ---------- -----a---- ---------- ---------- ----------
Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
    Subtype III-1  ---------- ---------- ---------- ---------- ----------
      Serotype IV  ---------- ---------- ---------- ---------- ----------
       Serotype V  ---------- ---------- c--------- ---------- -g--------
        Consensus  TTTGGTTGTA AGACCGTTTA TATAGAGGTT TTCGACAGGA TAGATAAACC
                                 cpsFA
                   1901                                              1950
    Subtype III-2  ---------- ---------- ---------- ---------- ----------
      Serotype VI  ---------- ---------- ---------- ---------- ----------
      Serotype Ib  ---------- ---------- ---------- ---------- ----------
Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
     Serotype VII  ---------- ---------- ---------- ---------- ----------
Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
    Subtype III-1  ---------- ---------- ---------- ---------- ----------
      Serotype IV  ---------- ---------- ---------- ---------- ----------
       Serotype V  ---------- ---------- ---------- ---------- ----------
        Consensus  AACTTTGACA GGAAAATTAG TGTATCCTGT AACAGATAAA TTTATTGTTC

                   1951                                              2000
    Subtype III-2  ---------- ---------- ---------- ---------- ----------
      Serotype VI  ---------- ---------- a--------- ---------- ----------
```

```
       Serotype Ib  ---------- ---------- ---------- ---------- ----------
   Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
      Serotype VII  ---------- ---------- ---------- ---------- ----------
   Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
      Subtype III-1  ---------- ---------- ---------- ---------- ----------
       Serotype IV  ---------- ---------- ---------- ---------- ----------
        Serotype V  ---------- ---------- ---------- ---------- ----------
         Consensus  AGTGGGAAGA AATGAAAAAA GTTTATCCTA AGGCAATTAA TTTAGGAGGA

                    2001                                              2050
      Subtype III-2  ---------- ---------- ---------- ---------- ----------
       Serotype VI  ---------- ---------- ---------- ---------- ----------
       Serotype Ib  ---------- ---------- -----a---- ---------- ----------
   Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
      Serotype VII  ---------- ---------- ---------- ---------- ----------
   Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
      Subtype III-1  ---------- ---------- ---------- ---------- ----------
       Serotype IV  ---------- ---------- ---------- ---------- ----------
        Serotype V  ---------- ---------- ---------- ---------- ----------
         Consensus  ATTTTTTAAT GATTTTTGTC ACAGTGGGGA CACATGAACA GCAGTTCAAC
                       cpsF | cpsG
                    2051                                              2100
      Subtype III-2  ---------- ---------- ---------- ---------- ----------
       Serotype VI  ---------- ---------- ---------- ---------- ----------
       Serotype Ib  ---------- ---------- ---------- ---------- ----------
   Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
      Serotype VII  ---------- ---------- ---------- ---------- ----------
   Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
      Subtype III-1  ---------- ---------- ---------- ---------- ----------
       Serotype IV  ---------- ---------- ---------- -------a-- ----------
        Serotype V  ---------- ---------- ---------- ---------- ----------
         Consensus  CGTCTTATTA AAGAAGTTGA TAGATTAAAA GGGACAGGTG CTATTGATCA

                    2101                                              2150
      Subtype III-2  ---------- ---------- ---------- ---c------ ----------
       Serotype VI  ---------- ---------- ---------- ---------- ----------
       Serotype Ib  ---------- ---------- ---------- ---------- ----------
   Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
      Serotype VII  ---------- ---------- ---------- ---------- ----------
   Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
      Subtype III-1  ---------- ---------- ---------- ---c------ ----------
       Serotype IV  ---------- ---------- ---------- ---------- ----------
        Serotype V  ---------- ---------- ---------- ---------- ----------
         Consensus  AGAAGTGTTC ATTCAAACGG GTTACTCAGA CTTTGAACCT CAGAATTGTC

                    2151                                              2200
      Subtype III-2  ---------- ---------- ---------- ---------- ----------
       Serotype VI  ---------- ---------- ---------- ---------- ----------
       Serotype Ib  ---------- ---------- ---------- ---------- ----------
   Serotype II/III-4  ---------- ---------- ---------- ---------- ----------
      Serotype VII  ---------- ---------- ---------- ------g--- -----g----
   Serotype Ia/III-3  ---------- ---------- ---------- ---------- ----------
      Subtype III-1  ---------- ---------- ---------- ---------- ----------
       Serotype IV  ---------- ---------- ---------- ---------- ----------
        Serotype V  ---------- ---------- ---------- ---------- ----------
         Consensus  AGTGGTCAAA ATTTCTCTCA TATGATGATA TGAACTCTTA CATGAAAGAA
                                               cpsGA
                   2201                    2226
     Subtype III-2  ---------- ---------- ------
       Serotype VI  ---------- ---------- ------
       Serotype Ib  ---------- ---------- c-----
  Serotype II/III-4  ---------- ---------- ------
      Serotype VII  ---------- ---------- ------
  Serotype Ia/III-3  ---------- ---------- ------
     Subtype III-1  ---------- ---------- ------
       Serotype IV  ---------- ---------- ------
        Serotype V  ---------- ---------- ------
         Consensus  GCTGAGATTG TTATCACACA TGGCGG
                           cpsGA1       *
```

Notes.

- Underlining parts show the relevant PCR and sequencing primers.
- Lowercase letters show the sequence polymorphism sites.
- Boldface parts show the gene start and stop codons.

3.4.5. Interserotype heterogeneity.

There were 56 sites of heterogeneity between the eight MS. The most suitable sites, for use in PCR and sequencing for MS, were a group of 23 sites nearest to the 3'-end of the region (Figure 3.1.), which were contained in the regions corresponding to the amplicons of primer pair of cpsES3-cpsGA1. Firstly, they were consistent across two panels of reference strains and most clinical isolates (the only exceptions were the small number of subtypes III-3 and III-4 isolates, see below). Secondly, they were relatively concentrated within a 790 bp region (positions 1437-2226), which is a convenient length for sequencing in a single reaction. Thirdly, they contained enough heterogeneity sites to allow differentiation, with few exceptions, of MS Ia-VII. Based only on this 790 bp region, subtype III-3 cannot be distinguished from MS Ia, nor subtype III-4 from MS II. However, they can be identified by MS III-specific PCR (see below).

Serotype VIII does not form amplicons with primer pairs targeting the 790 bp region, but can be identified by exclusion after PCR identification of GBS. In this study, one MS VIII isolate was identified, for which none of the primer pairs that amplify the 2226 bp region (in addition to those that amplify the 790 bp region) produced amplicons.

3.4.5. Mixed serotype-specificities in single isolates.

Eleven isolates were identified as one MS on the basis of the MS-specific PCR and overall sequence (within the 2226 or 2217 bp segment) but their sequences differed at some sites from isolates of the same MS and shared site-specific characteristics of

another. They included five subtype III-3 isolates and three subtype III-4 (see above). One non-serotypable reference strain (Prague 25/60), which was identified as MS II, differed from other MS II isolates at five sites at the 5'-end of the region, and was identical with MS III at three of these sites. Prague 25/60 MS III-specific PCR was negative. One clinical isolate identified as CS II, and MS II on the basis of its overall sequence, had bases at nine sites at the 5'-end of the region, that were characteristic of serotype Ib; MS Ib-specific PCR was negative. Finally, one CS V reference strain (Prague 10/84) had the same sequencing result as the corresponding sequence in GenBank (AF349539), but both were different, at three sites at the 5'-end of the region, from sequences of the other MS V strains that we studied.

All of these mixed-serotype specificities, except for those associated with subtypes III-3 and III-4, occurred at the 5'-end region of the 2226 or 2217 bp fragment. This supported our selection of the 3'-end 790 bp as the sequencing target for MS. Using this target, all MS were correctly identified except for MS III belonging to subtypes III-3 and III-4, which can be identified by MS III-specific PCR (see below).

3.4.6. MS identification based on MS-specific PCR targeting the 3'-end of *cpsG-cpsH-cpsI/cpsM*.

Our sequence alignment results showed that there was significant sequence heterogeneity in the 3'-end of *cpsG-cpsH-cpsI/cpsM* (data not shown), which makes it appropriate for use in the design of specific primer pairs for differentiation of serotypes Ia, Ib, III, IV, V, and VI directly by PCR (Table 3.2.). Using two panels of reference strains and the specified conditions, all MS-specific primer pairs amplified DNA only from the corresponding serotypes. When clinical isolates were tested, a MS was assigned, by PCR only, to 179 of 206 (86.9%) clinical isolates as follows: MS Ia 40; MS Ib 35; MS III 58; MS IV 7; MS V 36; MS VI 3.

3.4.7. Comparison of serotype identification results between MS and CS.

After CS and MS had been completed, the results were compared. Initial results were discrepant for 15 isolates, all were resolved by retesting or correction of clerical errors.

Table 3.4. Comparison of the results of conventional serotyping (CS) and molecular serotype identification (MS)/subtyping of 206 clinical GBS isolates.

	MS/subtype										
CS	**Ia**	**Ib**	**II**	**III-1**[a]	**III-2**[a]	**III-3**[a]	**III-4**[a]	**IV**	**V**	**VI**	**VIII**
Ia	38										
Ib		30									
II			25								
III				27	20	4	3				
IV								7			
V									31		
VI										2	
VIII											1
NT[b]	2	5	1	3	1				5	1	
Total (206)[c]	40	35	26[c]	30	21[c]	4	3	7	36	3	1

Notes.

a. For details of MS III sequence subtypes see text.

b. NT=nontypable.

c. One mixed culture was included as two separate isolates (one serotype II, one subtype III-2).

The CS and MS and sequence subtyping results are shown in Table 3.4. A MS was assigned to all isolates by PCR and/or sequencing, compared with 188 of 206 (91.3%) by CS. Specific PCR has not yet been developed for MS II and VIII, so all MS II isolates were determined by sequencing only and one MS VIII isolate by exclusion (see above). For all other isolates, the results of PCR and sequencing were consistent, except for subtypes III-3 and III-4 and other minor sequence differences described above. CS results correlated well with PCR results.

Final CS and MS results were the same for all 188 isolates (100%) for which results for both methods were available. Eighteen clinical isolates that were non-serotypable by CS, were assigned MS as follows: Ia, two; Ib, five; II, one; subtype III-1, three; subtype III-2, one; V, five; and VI, one.

3.4.8. Mixed culture.

Four clinical isolates gave positive results with MS III-specific PCR, but were provisionally identified as MS II by sequencing. Three were CS III and one CS II, with a weak cross-reaction with serotype III antiserum. These isolates were studied further by subculturing 12 individual colonies of each. All subcultures were tested by MS III-specific PCR. All 12 colony subcultures of the three CS III isolates were positive by MS III-specific PCR and the isolates were therefore classified as subtype III-4 (see above). However, 11 of 12 colony subcultures of the fourth isolate were negative by MS III-specific PCR; and one was positive by MS III-specific PCR. It was therefore assumed that this was a mixed culture, predominantly of MS/CS II. The CS of these two isolates were confirmed by retesting each individually. The one MS III-specific PCR positive and CS III colony was subsequently identified as subtype III-2 and included as an additional clinical isolate (total 206 in all).

3.4.9. Algorithm for serotype assignment of GBS by PCR and sequencing (see Table 3.2. for primer sequences).

Table 3.5. Algorithm for GBS molecular serotype (MS) identification by PCR and sequencing.

Amplification primer pairs*	PCR product size (base pairs)	Interpretation
GBS identification primer pairs		
Sag59-Sag190	196	GBS (*S. agalactiae*)
DSF2-DSR1	276	GBS (*S. agalactiae*)
GBS MS identification by MS-specific PCR		
IacpsHS1-cpsIA	354	serotype Ia
IbcpsIS-IbcpsIA1	523	serotype Ib
IIIcpsHS-cpsIA	641	serotype III
IVcpsHS1-IVcpsMA	379	serotype IV
VcpsHS2-VcpsMA	374	serotype V
VIcpsHS1-VIcpsIA	360	serotype VI
GBS MS identification by sequencing		
cpsES3-cpsGA1	790	refer to Figure 3.1. to identify
cpsES3-cpsFA	450	serotypes according to the
cpsFS-cpsGA1	423	sequence heterogeneity

Note.

*See Table 3.2. for primer sequences.

In order to make GBS serotype identification using our PCR and sequencing method practicable, we designed an algorithm for clinical use (Table 3.5.). All the primers (except the inner sequencing primers) used were given high melting temperature (>70 ^{o}C), so rapid cycle PCR could be used.

3.5. DISCUSSION

Capsule production in GBS is controlled by capsular polysaccharide synthesis (*cps*) gene cluster (Chaffin *et al*., 2000), which had been sequenced for serotype Ia (Yamamoto *et al*., 1999) and serotype III (Chaffin *et al*., 2000) before we began our study. Differences between the two serotypes and serotype-specific definition regions of *cpsH* (capsular polysaccharide repeating unit polymerase gene, serotype definition gene) had been identified (Chaffin *et al*., 2000; Yamamoto *et al*., 1999). Corresponding sequences for serotype Ib (Miyake *et al.,* submitted into GenBank, 2001 [GenBank accession number: AB050723]), and for serotypes IV, V, and VI (McKinnon *et al.,* submitted into GenBank, 2001 [GenBank accession numbers: AF355776, AF349539, AF337958, respectively]) were released when the project was nearly finished but those for the other three serotypes (II, VII and VIII), the sequences of *cps* gene clusters, have not been published previously.

The published sequences of *cps* gene clusters for serotypes Ia and III showed considerable homology at the 3'-end of *cpsD*-*cpsE*-*cpsF*-and the 5'-end of *cpsG* (Chaffin *et al*., 2000; Yamamoto *et al*., 1999) and we hypothesised that these regions would also be relatively conserved in other serotypes. Our study and the recently released sequences for *cps* gene clusters of serotypes Ib, IV, V and VI have supported this hypothesis. We designed a series of primers to amplify a 2226 or 2217 bp segment in this region and found that amplicons were obtained from all serotypes except VIII (Chaffin *et al*., 2000; Yamamoto *et al*., 1999). This confirmed a previous suggestion that serotype VIII is significantly different from other serotypes in this region (Cieslewicz *et al*., 2001).

Using eight serotype (Ia to VII) reference strains, we showed more than 50 heterogeneity points between serotypes (Figure 3.1.). Using 63 selected clinical isolates that had been serotyped by conventional methods, we found that these inter-serotype differences were generally consistent and specific, especially the 23 sites clustered at the 3'-end of the regions (corresponding to the regions amplified by the primer pair cpsES3-cpsGA1). We used these differences to assign serotypes to the remaining clinical isolates collected in this study, without knowledge of the serotype obtained by conventional methods.

Sequence analysis of the 3'-end of *cpsG*-*cpsH*-*cpsI*/*cpsM* for serotypes Ia, III, Ib, IV, V and VI showed that this region is highly variable, and *cpsH* also contains serotype-specific definition sites (Chaffin *et al*., 2000), making this region a suitable target for direct serotype identification by PCR. We designed MS-specific primers for MS Ia, Ib, III, IV, V and VI and confirmed their specificity, initially, using two CS reference panels. When used to test 206 clinical isolates, MS-specific PCR alone correctly identified 86.9%. Using rapid-cycle MS-specific PCR, results are available within one working day. In future, it will be possible to extend this method to all MS, when *cps* gene cluster sequences in this region are available for serotypes II, VII and VIII.

MS II and VII can be identified by sequencing the 790 bp PCR amplicons of the 3'-end of *cpsE*-*cpsF*-the 5'-end of *cpsG*. A positive GBS-specific PCR and negative PCR results with all the primers that amplify the 790 bp, identified MS VIII by exclusion. In future, and in some laboratories currently, sequencing of the 790 bp PCR amplicons of the 3'-end of *cpsE*-*cpsF*-and the 5'-end of *cpsG* for all isolates may be more convenient, as only one method and fewer primers are needed. However, if sequencing is not available in-house, the turn-around time is longer and a small proportion of serotypes would be wrongly assigned (subtypes III-3 and III-4 as MS Ia and II, respectively). This could be avoided by screening with MS III-specific PCR first. Sequencing the 790 bp PCR amplicon, allows MS III to be subtyped on the basis of the sequence heterogeneity.

Previous studies have shown that serotypes Ia, Ib, II, III and V are those most frequently isolated from normally sterile sites, in the United States and several countries (Hickman *et al*., 1999; Kalliola *et al*., 1999). Serotypes VI and VIII are the predominant serotypes isolated from patients in Japan (Lachenauer *et al*., 1999), but are uncommon elsewhere. Although our isolates were selected, they were probably representative of those causing disease in Australasia; Ia, Ib, II, III and V were the commonest serotypes identified, although there were small numbers of serotypes IV, VI and VIII.

Up to 13 % of GBS isolates are non-serotypable (Tyrrell *et al*., 2000); in our study the proportion was 8.7% (18/206) using antisera against all nine serotypes. Failure to react with anisera may be due to decreased type-specific-antigen synthesis (Palacios *et al*., 1997); non-encapsulated phase variation; or insertion or mutation in genes of *cps* gene clusters (Cieslewicz *et al*., 2001; Sellin *et al*., 2000). One non-serotypable strain GBS in our study had a T base deletion in *cpsG* gene, which caused a change in the *cpsG* reading frame.

In summary, we have developed an alternative to conventional serotyping for GBS, which is accurate and reproducible, can be performed by any laboratory with access to PCR and sequencing and, importantly, does not require panels of serotype-specific antisera that are increasingly difficult to maintain. All isolates are serotypeable and sequencing of a relatively limited 790 bp region can provide additional subtyping information for MS III. In future we will combine this with further PCR typing to identify members of the family of variable surface proteins that are important virulence factors for GBS and known GBS mobile genetic elements (to be reported separately). These extended typing methods will provide comprehensive strain identification that will be useful for epidemiological studies that will be needed to monitor GBS isolates before and after introduction of GBS conjugate vaccines.

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°C for 1 second, 45-72°C (according to the primer *T*m values or as previously described) for 1 second and 74°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°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°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).

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

This chapter has been published in *Journal of Medical Microbiology* 2003; 52:337-344. (see Appendices section)

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* °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	ISSa*4*	78.5	AF165983	**326**CGT ATC TGT CAC TTA TTT CCC TGC GGG TGT CTC C**359**
ISSa4A1	ISSa*4*	75.2	AF165983	**639**GCC GAT GTC ACA ACA TAG TTC AGG ATA TAG CCA G**606**
ISSa4A2	ISSa*4*	74.5	AF165983	**780**CGT AAA GGA GTC CAA AGA TGA TAG CCT TTT TGA 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: **GG**G A**TC CGA TT** (AF064785) vs **CA**G A**-- -GG** T**A** (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.

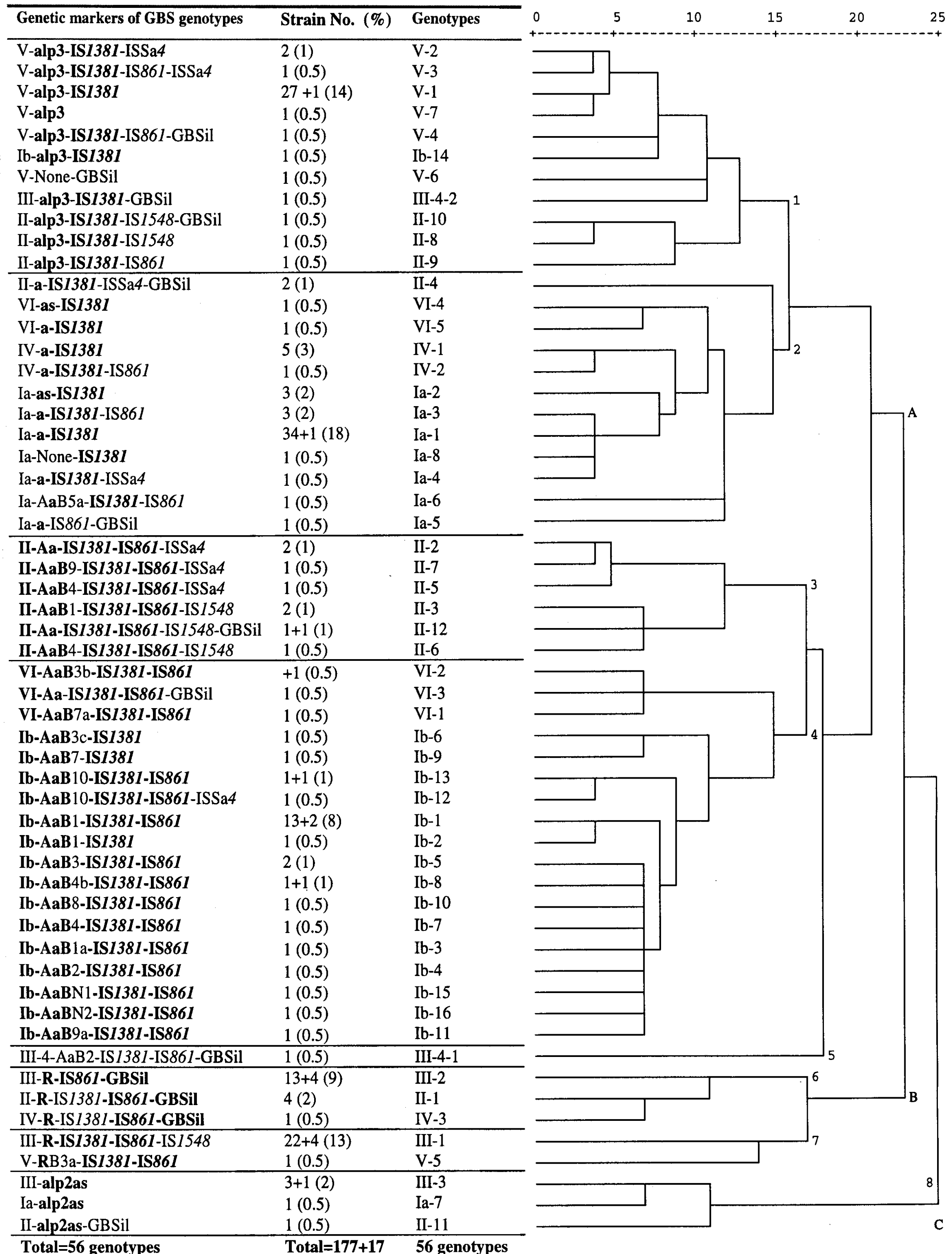
Theoretically there are 13 possible GBS MS/sst (eight MS - Ia, Ib, II, IV-VIII, four sst III 1-4 and *cps* gene cluster absent) and at least ten pgp (none, “Aa”, “AaB”, “a”, “as”, “R”, “RB”, “alp2as”, “alp3” or “alp4a”). If the 22 *bac* subgroups identified so far are included, there are up to 31 pgp. If the five mge were independently, randomly distributed and present or absent, there would be $13x31x2^5$=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-IS1381-ISSa4	2 (1)	V-2
V-alp3-IS1381-IS861-ISSa4	1 (0.5)	V-3
V-alp3-IS1381	27 +1 (14)	V-1
V-alp3	1 (0.5)	V-7
V-alp3-IS1381-IS861-GBSil	1 (0.5)	V-4
Ib-alp3-IS1381	1 (0.5)	Ib-14
V-None-GBSil	1 (0.5)	V-6
III-alp3-IS1381-GBSil	1 (0.5)	III-4-2
II-alp3-IS1381-IS1548-GBSil	1 (0.5)	II-10
II-alp3-IS1381-IS1548	1 (0.5)	II-8
II-alp3-IS1381-IS861	1 (0.5)	II-9
II-a-IS1381-ISSa4-GBSil	2 (1)	II-4
VI-as-IS1381	1 (0.5)	VI-4
VI-a-IS1381	1 (0.5)	VI-5
IV-a-IS1381	5 (3)	IV-1
IV-a-IS1381-IS861	1 (0.5)	IV-2
Ia-as-IS1381	3 (2)	Ia-2
Ia-a-IS1381-IS861	3 (2)	Ia-3
Ia-a-IS1381	34+1 (18)	Ia-1
Ia-None-IS1381	1 (0.5)	Ia-8
Ia-a-IS1381-ISSa4	1 (0.5)	Ia-4
Ia-AaB5a-IS1381-IS861	1 (0.5)	Ia-6
Ia-a-IS861-GBSil	1 (0.5)	Ia-5
II-Aa-IS1381-IS861-ISSa4	2 (1)	II-2
II-AaB9-IS1381-IS861-ISSa4	1 (0.5)	II-7
II-AaB4-IS1381-IS861-ISSa4	1 (0.5)	II-5
II-AaB1-IS1381-IS861-IS1548	2 (1)	II-3
II-Aa-IS1381-IS861-IS1548-GBSil	1+1 (1)	II-12
II-AaB4-IS1381-IS861-IS1548	1 (0.5)	II-6
VI-AaB3b-IS1381-IS861	+1 (0.5)	VI-2
VI-Aa-IS1381-IS861-GBSil	1 (0.5)	VI-3
VI-AaB7a-IS1381-IS861	1 (0.5)	VI-1
Ib-AaB3c-IS1381	1 (0.5)	Ib-6
Ib-AaB7-IS1381	1 (0.5)	Ib-9
Ib-AaB10-IS1381-IS861	1+1 (1)	Ib-13
Ib-AaB10-IS1381-IS861-ISSa4	1 (0.5)	Ib-12
Ib-AaB1-IS1381-IS861	13+2 (8)	Ib-1
Ib-AaB1-IS1381	1 (0.5)	Ib-2
Ib-AaB3-IS1381-IS861	2 (1)	Ib-5
Ib-AaB4b-IS1381-IS861	1+1 (1)	Ib-8
Ib-AaB8-IS1381-IS861	1 (0.5)	Ib-10
Ib-AaB4-IS1381-IS861	1 (0.5)	Ib-7
Ib-AaB1a-IS1381-IS861	1 (0.5)	Ib-3
Ib-AaB2-IS1381-IS861	1 (0.5)	Ib-4
Ib-AaBN1-IS1381-IS861	1 (0.5)	Ib-15
Ib-AaBN2-IS1381-IS861	1 (0.5)	Ib-16
Ib-AaB9a-IS1381-IS861	1 (0.5)	Ib-11
III-4-AaB2-IS1381-IS861-GBSil	1 (0.5)	III-4-1
III-R-IS861-GBSil	13+4 (9)	III-2
II-R-IS1381-IS861-GBSil	4 (2)	II-1
IV-R-IS1381-IS861-GBSil	1 (0.5)	IV-3
III-R-IS1381-IS861-IS1548	22+4 (13)	III-1
V-RB3a-IS1381-IS861	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.

CHAPTER 6

POSTGENOMIC TAXONOMY OF HUMAN UREAPLASMAS

Fanrong Kong, Gwendolyn L. Gilbert

Centre for Infectious Diseases and Microbiology (CIDM), Institute of Clinical Pathology and Medical Research (ICPMR), Westmead, New South Wales 2145, Australia

This chapter is in press in *International Journal of Systematic and Evolutionary Microbiology* .

Statement of Joint Authorship

Kong, F. (candidate)

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

Gilbert G. L. (supervisor)

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

6.1. SUMMARY

In 2000, the full genome sequence of *Ureaplasma parvum* (previously known as *Ureaplasma urealyticum*) serovar 3 was released. In 2002, after prolonged debate, it was agreed that the former *U. urealyticum,* should be divided into two species – *U. parvum* and *U. urealyticum.* To provide additional support for this decision and improve our understanding of the relationship between these two species, we studied four "core" genes or gene clusters in ATCC reference strains of all 14 serovars of *U. parvum* and *U. urealyticum.* These core regions were the rRNA gene clusters, the elongation factor Tu genes (*tuf*), urease gene complexes and multiple banded antigen genes (*mba*). The known *U. parvum* genome sequences (GenBank accession number: NC_002162) were used as reference. DNA insertions and deletions (indels) were found in all of the gene regions studied, except *tuf*, but they were found only between, not within the two species. An incidental finding was that there was inter-copy heterogeneity for rRNA gene complex sequences. Sequence analysis (sequence heterogeneity and especially indels) of all four selected targets consistently support the separation of human ureaplasmas into two species. Except for MBA, there was less heterogeneity in amino acid sequences of proteins, between species, than in the nucleic acid seqeuences of the corresponding genes. The level of heterogeneity in amino acid and base sequences at the 5'end of the species-specific regions of MBA were almost identical. Analysis of our results provided an interesting case study to help resolve some common problems in the use of sequence data to infer phylogenetic relationships and support taxonomic changes. We recommend that, to avoid confusion, the new nomenclature be used for human ureaplasmas in future publications.

6.2. INTRODUCTION

There have been two recent major developments that affect our understanding of human ureaplasmas. Firstly, the full genome sequence of *U. parvum* (previous

U. urealyticum) serovar 3 was released in 2000 (GenBank accession number: NC_002162) (Glass *et al.*, 2000). Secondly, the taxonomy of human ureaplasmas changed in 2002 (Robertson *et al.*, 2002), when the two former *U. urealyticum* biovars were given full species status, as *U. parvum* (previously biovar parvo or biovar 1) and *U. urealyticum* (previously biovar T-960 or biovar 2) (Robertson *et al.*, 2002).

It is now accepted that a decision to create a new species should be based on many independent phenotypic and genotypic characteristics – the theory of "polyphasic taxonomy" (Vandamme *et al.*, 1996). However, molecular methods and genome-based criteria have become more accessible and attractive (Gurtler & Mayall, 2001; Stackebrandt *et al.*, 2002). DNA-DNA hybridization showing <70% homology between whole genomes is accepted as the most definitive criterion or "gold standard" for separate prokaryote species (Murray & Schleifer, 1994). The traditional 70% DNA-DNA hybridization value used to delineate genomic species was found to correspond to genome mispairings in the range 13-13.6% or 0.097-0.104 nucleotide substitutions per site (Mougel *et al.*, 2002). Similarity of <97% in the 16S rRNA gene is the most widely used practical alternative (Murray & Schleifer, 1994), but these criteria may conflict and additional alternative targets are required (Pettersson *et al.*, 2000; Dellaglio, *et al.*, 2004).

The two ureaplasma species exhibit many distinct phenotypic and genotypic properties (including DNA-DNA hybridization showing <70% homology), which support the change in taxonomy and fulfil the requirements of the polyphasic theory (Chrisstiansen *et al.*, 1981; Harasawa *et al.*, 1991; Vandamme *et al.*, 1996). However, continued use of the old single-species nomenclature in some recent publications (Baier *et al.*, 2003; Daxboeck *et al.*, 2003) is potentially confusing. To strengthen the case for acceptance and exclusive use of the new ureaplasma taxonomy (Robertson *et al.*, 2002), we studied four "core" genes/gene clusters – the rRNA gene complex, elongation factor Tu gene (*tuf*), urease gene cluster and multiple banded antigen gene (*mba*) – of all 14 human ureasplasma serovars. These

four regions were chosen because previous studies have shown that they were promising targets for study of the phylogeny of ureaplasmas and mycoplasmas (Kamla, *et al.*, 1996; Kong, *et al.*, 1999b, 2000b). In addition, we used this as a case study to help resolve some common problems with the use of sequence data to infer phylogeny and to support the establishment of new taxonomy (Ludwig *et al.*, 1998).

6.3. MATERIALS AND METHODS

6.3.1. Bacterial strains.

Reference strains of four *U. parvum* and ten *U. urealyticum* serovars, obtained directly from the American Type Culture Collection (ATCC), were the same as those used in our previous studies (Kong *et al.*, 1999a) and are listed in Chapter 1.

6.3.2. Value of *U. parvum* serovar 3 genome in oligonucleotide primers design.

The full genome sequence of *U. parvum* serovar 3 (Glass *et al.*, 2000) greatly facilitated sequencing of the selected genes and gene clusters of the other three *U. parvum* and ten *U. urealyticum* serovars. For this study, we used the following steps: Firstly, to identify conserved regions: we compared known sequences of genes corresponding to our selected genes and gene clusters – rRNA gene complex, *tuf* and urease gene clusters – in other *Mycoplasma* spp. and ureaplasma serovars (Fraser *et al.*, 1995; Himmelreich *et al.*, 1996; Neyrolles *et al.*, 1996), in addition to *mba*, which we and others have studied in detail previously (Zheng, *et al.*, 1995; Kong *et al.*, 1999a). Based on the results, we designed primers and amplified target regions for sequencing. The target regions sequenced were:

a) the whole rRNA cluster, including a short region upstream of 16S rRNA gene (for *U. parvum* serovars only)-16S rRNA gene-16S/23 rRNA intergenic spacer-23S rRNA gene-23S/5S rRNA intergenic spacer-5S rRNA gene-and a short region downstream of 5S rRNA gene;

b) almost the full length of *tuf*;
c) the whole urease gene cluster, including short regions upstream of *ureA*-*ureA*-*ureB*-*ureC*-*ureE*-*ureF*-*ureG*-*ureD*-and downstream of *ureD*.

The amplification and sequencing primers used in the study are shown in Table 6.1. Most primers were used for both amplification and sequencing but some (as inner sequencing primers) were used only for sequencing.

6.3.3. DNA preparation and PCR.

DNA preparation and PCR were performed as previously described (Kong *et al.*, 1999b).

6.3.4. Sequencing and sequence analysis.

The PCR products were sequenced with Applied Biosystems (ABI) *Taq* DyeDexoy terminator cycle-sequencing kits according to standard protocols. All sites showing unexpected heterogeneity, such as those indicating rRNA gene intercopy sequence variation and the unique heterogeneity site in serovar 13 *tuf* (see below), were sequenced at least twice, to confirm the results. When necessary, different PCR amplicons and/or inner sequencing primers were used for sequencing.

The initial sequencing results were analysed with *Bestfit* program in *Comparison* program group and then joined together to determine sequences of whole genes/gene clusters. The multiple sequence alignments were performed with *Pileup* and *Pretty* programs from the *Multiple Sequence Analysis* program group. All of the programs/program groups are available in WebANGIS (http://www1.angis.org.au/WebANGIS/), ANGIS (Australian National Genomic Information Service).

Table 6.1. Primers used for sequencing three ureaplasma different genes/gene clusters.

Primer names[a]	Target genes/regions[a]	*Tm* °C[b]	GenBank numbers[c]	Primer sequences[d]
113S1	UU113	68.4	AE002111	**9814**GAA GAA CCC ACC AAA TAC GAG CAG**9837**
113S2[e]	UU113	62.7	AE002111	**9862**TTG TTG GTG AAC AAA AAT ACA TCA**9885**
303S1	UU303	64.9	AE002127	**5713**TTG ATG CAA AAA GAT CAG GTT GTA G**5737**
303S2[e]	UU303	61.7	AE002127	**5800**GAG AAA CAA GCT GAA CAT AAT GAT C**5824**
16S10A[e]	16S rRNA	66.4	AE002127	**6144**AAT CCT GAG CCA GGA TCA AAC TC**6122**
16S23A	16S rRNA	74.3	AE002127	**6156**GCC GCC AGC GTT AAT CCT GAG C**6135**
SP1623S1	16S-23S rRNA spacer	65.3	AE002127	**7830**CTT TCT AAT CAT TGA CAT TAA GTT GTC AGT G**7860**
SP1623S2[e]	16S-23S rRNA spacer	62.8	AE002127	**7843**GAC ATT AAG TTG TCA GTG AAC AGA AAC**7869**
23S32S	23S rRNA	63.6	AE002127	**7933**CTA AGA GCT TAT GGT GA/GA TGC CTT G**7957**
23S523S	23S rRNA	69.4	AE002127	**8424**GAA CGG TGA AAA GAA CCC AGA GAT G**8448**
23S503A[e]	23S rRNA	65.7	AE002127	**8449**CCA TCT CTG GGT TCT TTT CAC C**8428**
23S516A	23S rRNA	63.7	AE002127	**8464**GGT TCT ATT TCA CTC CCA TCT CTG**8441**
23S1104S	23S rRNA	68.7	AE002127	**9006**GCA AGG ATG TTG GCT TAG AAG CAG**9029**
23S1082A[e]	23S rRNA	68.7	AE002127	**9030**GCT GCT TCT AAG CCA ACA TCC TTG**9007**
23S1134S[e]	23S rRNA	67.5	AE002127	**9034**CGT TTA AAG AGT GCG TAA CAG CTC AC**9059**
23S1117A	23S rRNA	70.8	AE002127	**9067**CTC GAC AAG TGA GCT GTT ACG CAC TC**9042**
23S1721S	23S rRNA	68.9	AE002127	**9622**AAG GAA CTC TGC AAA TTA ACC CCG T**9646**

23S1698A[e]	23S rRNA	68.0	AE002127	**9647**TAC GGG GTT AAT TTG CAG AGT TCC T**9623**
23S 1729A	23S rRNA	68.5	AE002127	**9677**TTT TAC AGC GAG CAC CCC TTA TTG**9654**
23S2257S	23S rRNA	68.5	AE002127	**10158**GAC AGT GTT AGG TGG GCA GTT TGA C**10182**
23S2237A[e]	23S rRNA	71.9	AE002127	**10186**CCC AGT CAA ACT GCC CAC CTA ACA C**10162**
23S2251A	23S rRNA	78.6	AE002127	**10198**GGA GGC GAC CGC CCC AGT CAA AC**10176**
23S2578S	23S rRNA	74.6	AE002127	**10478**GGT TCG GCT GTT CGC CGA TTA AAG AG**10503**
23S2603A	23S rRNA	55.8	AE002127	**10551**AGA TAG GGA CCA ACC TGT CTC ACG**10528**
23S2772S	23S rRNA	65.2	AE002127	**10674**AAA CGC TGA AAG CAT CTA AGT GTG**10697**
5S41S	5S rRNA	67.7	AE002128	**86**GAA ATA CCT GTT CCC ATC CCG A**107**
5S60S[e]	5S rRNA	70.0	AE002128	**103**CCC GAA CAC AGA AGT CAA GCA CTC**126**
5S45A	5S rRNA	67.4	AE002128	**134**CGG CTC TAG AGT GCT TGA CTT CTG**111**
UU304A1[e]	UU304	63.0	AE002128	**330**TTC TAA TTG CAA TTC TTC AAG ACG**307**
UU304A2	UU304	71.8	AE002128	**400**CAC CTT GTT CGC GTG CAT CTT G**379**
UU114A1[e]	UU114	60.1	AE002112	**5246**A/GTT TAT TGT TTT TGG ATA TAC CAC C**5222**
UU114A2	UU114	66.2	AE002112	**5405**CGT CTT CTG GTG TTT GCA TAA TTG**5382**
TUF5S	*tuf*	61.3	AE002151	**1284**TTA ATT TTT AAG GAG ATT TAA AAT GGC**1258**
TUF32S[e]	*tuf*	62.4	AE002151	**1256**AAA GCT AAA TTT GAA AGA ACA AAA CC**1231**
TUF92S	*tuf*	60.5	AE002151	**1195**ATG GTA AAA CTA CTT TAA CAG CTG C**1171**
TUF163A	*tuf*	61.2	AE002151	**1076**TTG TAA TAC CAC GTT CTC TTT CTT C**1100**
TUF590S	*tuf*	67.9	AE002151	**697**TTG ATG AAT TAA TGG ACG CAG TTG A**673**

TUF646A	*tuf*	65.3	AE002151	**592**CGT CCT GAA ATT GTG AAT ACA TCT TC**617**
TUF874S[e]	*tuf*	66.9	AE002151	**414**AAA AGA AGA TGT TGA ACG TGG TCA AG**389**
TUF985A	*tuf*	61.1	AE002151	**253**TCT GTT GTT CTA AAA TAG AAT TGT GG**278**
TUF1132A[e]	*tuf*	65.0	AE002151	**107**GAC CTA CAG TTT TAC CAC CTT CAC G**131**
TUF1159A	*tuf*	59.0	AE002151	**77**ATT AAT TAC TTG TTT TAA TTA CGC TAC C**104**
UC424S	*ureC*	65.6	AE002140	**2676**CAG CTG GTG GTT TAG ATA CTC ACG**2653**
UC429S[e]	*ureC*	64.8	AE002140	**2672**TGG TGG TTT AGA TAC TCA CGT TCA C**2648**
UC910S	*ureC*	64.9	AE002140	**2181**TTT TAC CAG CTT CTA CAA ACC CAA C**2157**
UC947A[e]	*ureC*	60.6	AE002140	**2094**AAG TGG TGA CAT ACC ATT AAC ATA TC**2119**
UC959A	*ureC*	60.5	AE002140	**2083**CCT TAG GAT TTA AGT GGT GAC ATA C**2107**
UC1418S	*ureC*	68.0	AE002140	**1683**GG/TG ATC CAA AC/TG CTT CAA TTC CAA C**1659**
UC1450A	*ureC*	61.0	AE002140	**1601**A/GC/TT AGT TAA C/TA/GA ACG TCC ATA A/TGT TCC**1624**
UE93S	*ureE*	59.5	AE002140	**1147**GAA CAT TCA TTT AAC AAG CGA CGA C**1126**
UE119A	*ureE*	62.9	AE002140	**1073**CCA TAT TCA ACA TTT TGA TCT GAT G**1097**
UE143S[e]	*ureE*	62.9	AE002140	**1097**CAT CAG ATC AAA ATG TTG AAT ATG G**1073**
UF217S	*ureE*	59.5	AE002140	**632**GCG TTA CTT CTG TAT ATG AAT GAA C**608**
UF222A[e]	*ureF*	60.0	AE002140	**578**AAA TTG CC/TA ATA AAT CAC CAT GTA AC**603**
UF320A	*ureF*	63.1	AE002140	**484**CGA GTC TCT CTT GCT AAA CCT TG**506**
UF721S	*ureF*	64.7	AE002140	**129**C/TCT TGA AAT TGC ACA AAT GGA AC**107**
UG3A	*ureG*	62.4	AE002139	**11152**CCT ACA CCA ATA ATT AAT GGT CTT TTC**11178**
UG451A[e]	*ureG*	63.4	AE002139	**10706**CAC CAA CAT AAG GAG CTA AAT CAA C**10730**

UG475S	*ureG*	63.4	AE002139	**10730**GTT GAT TTA GCT CCT TAT GTT GGT G**10706**
UG499A	*ureG*	60.6	AE002139	**10655**CTT TAT TAC CAC GTG ATT TTA ATG TAT C**10682**
UD332S	*ureD*	59.0	AE002139	**10246**CA/TG AA/GC AAC AC/TA CAA ATA TC/TA CA/GT TAG G**10219**
UD379A	*ureD*	62.0	AE002139	**10147**TTA AAT TGG/T GCA/G AAC/T TTT CCA TCT TC**10172**
UD841A	*ureD*	62.3	AE002139	**9684**CTT C/TTA TGG TTT TCG TAA AAT TAA A/TGG**9710**
UU427A1[e]	UU427	60.4	AE002139	**9446**AAT AAA TTT TGC TAA AAA AGG CAT AC**9471**
UU427A2	UU427	56.5	AE002139	**9330**GT/CA/T GGT/C TTA AAA T/CTA ACA TCT ACA C**9354**
UU427A3	UU427	63.1	AE002139	**9175**CAT CAT CAA AAT CTT TAA TAC CAT CAT C**9199**

Notes.

a. Primers were named according to their target genes/regions, the 3'-end locations (the distance from the beginning of correspondent genes/regions), and directions of primers (sense or antisense).

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

c. All the GenBank sequences were from the related sections of *U. parvum* serovar three full genome sequences.

d. Numbers represent the numbered base positions at which primer sequences start and finish (numbering start point "1" refer to the start points "1" of correspondent gene GenBank accession numbers). Letters behind "/" indicate alternative nucleotides in different species/serovars, which were based on the comparison with the other related GenBank sequences.

e. Primers used only for sequencing; all the other primers were used for both PCR and sequencing.

6.3.5. Nucleotide sequence accession numbers.

The sequence data were deposited into the GenBank Nucleotide Sequence databases with the following accession numbers: rRNA gene complex: AF272599-AF272604, AF073446-AF073459, AF059322-AF059335 and AF272605-AF272630; *tuf*: AF270758-AF270770; urease gene clusters: AF085720-AF085733; *mba*: AF055358-AF055367 and AF056982-AF056984.

6.4. RESULTS AND DISCUSSION

6.4.1. Advantages of sequencing multiple strains.

In this study, as in our previous study of *mba* (Kong, *et al.*, 2000b), we sequenced the three target genes/gene clusters for 13 ureaplasma serovars (excluding serovar 3 sequences from the full genome, which were used as reference). The advantages of this approach are:

i) since the genes/gene clusters are relatively conserved between serovars within each species, the results for different serovars help to confirm the accuracy of sequencing results (Clayton *et al.*, 1995);

ii) the results help to differentiate interspecies from intraspecies heterogeneity (Mygind *et al.*, 1998).

A previous study showed that there is some intra-serovar, as well as inter-serovar/intra-species and inter-species heterogeneity in *mba* (Knox, *et al.*, 1998). However, there is limited intra-species heterogeneity in the other three gene regions studied and, even in *mba,* intraspecies heterogeneity is much less than between species (Kong, *et al.*, 1999b). Therefore, a single reference strain of each serovar provides enough examples of each species to demonstrate inter-species heterogenity, which was the focus of this study.

6.4.2. Key characteristics of the "core genes".

6.4.2.1. Inter-copy polymorphisms of the rRNA gene complex.

There was relatively little interspecies heterogeneity in 16S and 23S rRNA genes but considerably more in the two copies of the 5S rRNA genes and the corresponding intergenic spacer regions (Table 6.2.). In common with other targets studied, the intraspecies heterogeneities in these genes were minor and some were assumed to be due to intercopy differences between duplicate copies (Table 6.2.). Previous studies have shown sequence variation in duplicate copies of rRNA genes of other mollicutes (Pettersson *et al.*, 1996). Analysis of the two copies of the rRNA gene complex in *U. parvum* serovar 3 genome (GenBank accession number: AE002111+AE002112 and AE002127+AE002128) showed inter-copy heterogeneity between 16S rRNA genes (one site), 16S-23S rRNA intergenic spacer regions (two sites), 23S rRNA genes (four sites) and 5S rRNA genes (one site) but none in the 23S-5S rRNA intergenic spacer regions. In sequences of the corresponding genes of the other 13 serovars the result was "N" (i.e. unknown or unidentifiable nucleotide) rather than "A, T, C or G", at several sites, even after repeat sequencing or use of different amplication and sequencing primers. We assumed that most, if not all, of these were due to inter-copy polymorphisms (Table 6.2.) (Ueda *et al.*, 1999). If these inter-copy polymorphisms were ignored, the intraspecies heterogeneity in rRNA gene complexes between the two human ureaplasma species was very low. In future, the design of primers or probes or study the phylogenetic relationships should take account of polymorphisms between multi-copy rRNA gene complexes (Gurtler, 1999).

6.4.2.2. *U. urealyticum* serovar 13 EF-Tu gene (*tuf*).

Previous studies have shown that differences in *tuf* can distinguish species and may reflect some phenotypic relationships better than 16S rRNA gene (Kamla *et al.*, 1996). EF-TU gene (*tuf*) DNA sequences were the same in serovars within each

species, except for that of serovar 13 of *U. urealyticum*. It contains two base differences (but the same amino acid sequences) compared with the other nine *U. urealyticum* serovars. This difference is of interest in view of another reported difference between serovar 13, which gives an intermediate response in the Mn^{2+} (manganese)-inhibition test, and all other *U. urealyticum* serovars, which are fully inhibited (Robertson & Chen, 1984).

6.4.2.3. Intraspecies and interspecies heterogeneity of urease gene clusters.

There were 21, nine and one intraspecies heterogeneity sites in the urease gene cluster DNA sequences for *U. parvum* serovars 1, 6, and 14 (compared with serovar 3), respectively. There were seven heterogeneity sites (or 12 base pairs – one site has 6 bp difference) in *U. urealyticum* serovar 2, compared with all other *U. urealyticum* serovars. These results show greater heterogeneity between urease gene clusters of *U. parvum* serovars than between those of *U. urealyticum* serovars, as we found previously for MBA genes (Kong *et al.*, 1999a).

Interspecies heterogeneity between urease genes clusters of the two species was greater in the intergenic spacer regions (where it varies from 16.8-31.8%) than in the genes themselves (where the range of heterogeneity is 5.9-9.7%). Variation in amino acid sequences, between species, is less (range 0.97-6.7%) than in nucleic acid sequences of urease genes (Table 6.3.).

6.4.2.4. The molecular clock is different for different MBA gene regions.

As we described previously, different MBA regions apparently evolve at different rates i.e. according to different “molecular clocks” (Bromham & Penny, 2003). The upstream regions are more heterogeneous than MBA gene themselves (Kong *et al.*, 1999a) and the repetitive regions more heterogeneous than the N-terminal regions. This should be taken into account when using different regions to infer the phylogeny (Kong *et al.*, 1999a).

Table 6.2. Comparison of inter-species, intra-species and inter-copy heterogeneity in rRNA gene complexes of *U. parvum* and *U. urealyticum*.

Genes/regions	DNA length	Heterogeneity sites: N (%)		
		Interspecies[b]	*U. parvum* intraspecies	*U. urealyticum* intraspecies
16S rRNA gene	1513[a]	14 (0.93)	2 (1[c])	2 (1[c])
16S-23S rRNA gene spacer	302	13 (4.3)	2 (2[c])	0
23S rRNA gene	2903	26 (0.90)	8 (7[c])	10 (9[c])
23S-5S rRNA gene spacer	71	3 (4.2)	0	0
5S rRNA_1 gene	115	7 (6.1)	1 (1[c])	3 (3[c])
5S rRNA_2 gene	115	8 (7.0)		
5S rRNA_1 gene-UU114 spacer[d]	136	31 (3.8)	-	-
5S rRNA_2 gene-UU304 spacer[d]	105	20 (19.0)	-	-

Notes.

a. Length modification was based upon two other mollicutes 16S rRNA genome annotations (Fraser *et al.*, 1995; Himmelreich *et al.*, 1996), especially considering *M. pneumoniae* annotation (Himmelreich *et al.*, 1996).

b. *U. parvum* and *U. urealyticum* interspecies heterogeneity sites were determined independently of intraspecies heterogeneity sites.

c. Numbers in parenthesis were assumed numbers to contain inter-copy heterogeneity (see text for further explanation).

d. The rRNA gene complex (or 5S rRNA gene) downstream external spacer regions.

6.4.2.5. Indels.

Analysis of insertions and deletions (indels) is a very useful tool with which to study bacterial phylogeny (Britten *et al.*, 2003; Gupta & Griffiths, 2002). We compared the distribution of indels in the four core genes/gene clusters of four *U. parvum* serovars with those of ten *U. urealyticum*. All indels were consistent between serovars within each species, which strongly supports the separation of *Ureaplasma* spp., based on indels. The rRNA gene complex of *U. parvum* differed from that of *U. urealyticum* as follows:

a) a TGTG insertion in 16S rRNA gene;
b) an AT (for operon 1) or A and C (for operon 2) deletion and AT insertion (for operons 1 and 2) in the 16S-23S rDNA intergenic spacer region;
c) a G insertion in 23S-5S rDNA intergenic spacer region;
d) a TTAGG (for operon 1) or AAAAA (for operon 2) deletion in the 5S rRNA gene.

There were no indels in the EF-TU gene (*tuf*). In the urease gene clusters, there were:

a) a TCAAT deletion in the *ureA-ureB* spacer;
b) AAC, T and CTA insertions in *ureB-ureC* spacer;
c) a CA deletion in *ureC-ureE* spacer and
d) an ACATT insertion in the *ureF-ureG* spacer.

Despite these specific differences, the numbers of insertions or deletions, sites and total number of bases in these three genes were not significantly different between the two species. In *mba*, there were no indels in species-specific sites, but there was an AAATT insertion, an AA deletion, a 45 bp deletion and a TC deletion in *U. parvum* upstream of *mba* (Kong, *et al.*, 2000b).

6.4.3. Genes, intergenic spacers or gene clusters?

Many studies have shown that intergenic spacer regions are more heterogeneous

than the neighbouring genes (Garcia-Martinez *et al.*, 1999; Kong, *et al.*, 1999b). Our study confirmed this by showing greater heterogeneity in the intergenic spacers, especially the external spacer regions, of both the rRNA gene complex and the urease gene clusters compared with the corresponding genes (Tables 6.2. and 6.3.) (Jung *et al.*, 2003). In addition, for urease gene clusters and *mba*, indels only existed in the gene spacer regions. Because the gene cluster as a whole is a functional group, we suggest that there are many advantages in considering them as a unit in basic and applied research.

i) Whole gene clusters contain both conserved and variable sequences and phylogenetic data derived from them are stable and discriminatory (Gurtler, 1999), which is valuable in solving taxonomic problems (Harasawa, 1999).

ii) Species-specific primer pairs based on whole gene clusters are generally more specific and easier to be designed than primers based on any single component (Kong *et al.*, 2000a).

6.4.4. DNA or protein sequence? Which protein or gene region?

To fulfil polyphasic theory requirements (Vandamme *et al.*, 1996), DNA sequences and protein amino acid sequences should be considered together (Agosti *et al.*, 1996). However, DNA sequences often reflect the phylogeny more accurately and have greater (about double) discriminatory power (Simmons *et al.*, 2002). Our study showed that, the *mba* species-specific region (5'-end or N-terminal) DNA (67/430= 15.6%) and the corresponding amino acid sequences (24/147=16.3%) have nearly identical levels of heterogeneity (Kong *et al.*, 2000b). However, urease gene subunit (Table 6.3.) and EF-TU gene DNA sequences are more heterogeneous than their corresponding protein amino acid sequences. For example, for the ureaplasma EF-TU gene DNA sequence heterogeniety was 54 of 1185 (4.6%) bases compared with 2 of 394 (0.5%) differences in amino acids between the two species. Presumably, genetic changes that significantly alter the structure, and therefore the function, of proteins such as enzymes are incompatible with survival. On the other hand, genetic

Table 6.3. Comparison of interspecies heterogeneity of DNA and amino acid sequences of the urease gene clusters of *U. parvum* and *U. urealyticum*.

Genes/regions	DNA		Amino acid	
	Length (bases)	Heterogeneity: N (%)	Length (a.a.)	Heterogeneity: N (%)
Upstream of *ureA*[a]	149	25 (16.8)		
ureA	306	18 (5.9)	101	5 (5.0)
***ureA-ureB* spacer**	51	9 (17.6)		
ureB	375	31 (8.3)	124	6 (4.8)
***ureB-ureC* spacer**	45	11 (24.4)		
ureC	1797	175 (9.7)	598	26 (4.3)
***ureC-ureE* spacer**	66	21 (31.8)		
ureE	450	37 (8.2)	149	10 (6.7)
ureF	753	64 (8.5)	250	15 (6.0)
***ureF-ureG* spacer**	81	18 (22.2)		
ureG	621	41 (6.6)	206	2 (0.97)
***ureG-ureD* spacer**	10	3 (30)		
ureD	864	82 (9.5)	287	15 (5.2)
***ureD*-UU427spacer[b]**	122	35 (28.7)		

Notes.

a. The urease complex (*ureA*) up stream external spacer region.

b. The urease complex (*ureD*) down stream external spacer region.

variation that causes antigenic variation in MBA is not only consistent with survival but also an advantage if it helps the organism to evade the host immune immune response.

Many surface protein antigen genes are used to study the phylogeny of different microbes and to develop practical identification and typing schemes (Bush & Everett, 2001; Stackebrandt, *et al.*, 2002). Sometimes, the gene or even gene region selected can significantly affect the results (Bromham & Penny, 2003). For example, ureaplasma MBA genes contain both species-specific (5'-end or N-terminal) and serovar definition sites (repetitive regions or 3'-end). Thus, the 5'-end or N-terminal would be the appropriate region for studying species-level phylogeny, rather than the repetitive regions (Zheng, *et al.*, 1995). If different bacterial species share almost identical protein antigens as a result of lateral gene transfer (Lawrence, 2002) – for example the *Streptococcus agalactiae* Alp3 protein and *Streptococcus pyogenes* R28 protein (Stalhammar-Carlemalm *et al.*, 1999) – the corresponding genes lose their value for studying species-level taxonomy (Thornton, 2002).

6.4.5. Why "core" genes?

Of the two ureaplasma species **–** *U. parvum* and *U. urealyticum* (Robertson *et al.*, 2002) – a full genome sequence was available only for *U. parvum* (Glass *et al.*, 2000). In future, our understanding of human ureaplasmas would be significantly improved by availability of the full genome sequence of *U. urealyticum* also. In particular, it would help to elucidate the nature and significance of the >80 kbp size difference between the two human ureaplasma species (Robertson *et al.*, 1990;

Fraser, *et al.*, 2000) in reverse evolution (Rocha & Blanchard, 2002) and pathogenesis (Povlsen *et al.*, 2002). Meanwhile, alternative strategies such as analysis of selected "core" genes or gene clusters (as in this study) can be used to infer the phylogenetic relationship between species (Daubin *et al.*, 2002). The rationale for the choice of these four genes was that the rRNA gene cluster (Stackebrandt, *et al.*, 2002) and *tuf* (Kamla, *et al.*, 1996) have been widely accepted targets for phylogenetic/taxonomic studies and the urease gene cluster and *mba* are unique determinants (Stackebrandt, *et al.*, 2002) of ureaplasma metabolism (Neyrolles, *et al.*, 1996) and antigenicity (Zheng, *et al.*, 1995).

6.5. CONCLUSION

Analysis of four "core" genes/gene clusters further supported the establishment of two separate human ureaplasma species – *U. parvum* and *U. urealyticum*. Significant differences between genes/gene clusters in the degree of heterogeneity between and within species sheds further light on the relationships between them and makes a useful case study to help understand common problems in use of sequence data to infer phylogeny and support taxonomic change (Ludwig *et al.*, 1998).

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.

CHAPTER 8

INTEGRATED STUDY OF GBS AND HUMAN UREAPLASMAS – THEIR EVOLUTIONARY "WISDOM"

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

Kong, F. (candidate)

Did all the bioinformatic work, 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.

8.1. Integrated study GBS and human ureaplasmas together.

8.1.1. Microbiology – in an era of paradigm shifts.

In microbiology, we are in an era of paradigm shifts (Bull *et al.*, 2000), one of which is the popular use of bioinformatics. In bioinformatics, the research strategy is based upon data collection and mining (retrieval and integration) to generate knowledge (the understanding of what is important about a situation) from information (the sum of everything we know about that situation) (Bull *et al.*, 2000). In the process, integrated analysis of different types of data may provide solutions, which analysis of either separately could not provide (van Belkum *et al.*, 2001). For example, the combination of comparative genomic and evolutionary data – an integrated phylogenomic approach – is one successful case (Fraser *et al.*, 2000; Eisen & Fraser, 2003).

This study aimed to determine whether several major, seemingly unrelated, aspects of two clinically and phylogenetically related, urogenital tract and perinatal conditional pathogens – group B streptococcus (GBS, *S. agalactiae*) and human ureaplasmas (*U. parvum* and *U. urealyticum*) – could be accounted for in a simple and unified manner (Almogy *et al.*, 2002). In particular, we aimed to show whether an integrated study of the evolution of GBS and human ureaplasmas could reveal aspects of both that may not otherwise be apparent.

8.1.2. GBS and human ureaplasmas – many clinical similarities and differences.

GBS and ureaplasmas are two of the most common microorganisms associated with obstetric and perinatal infections (Razin *et al.*, 1998; Schuchat, 1998). Ureaplasma colonization rates in pregnant women are normally higher (up to 80%) than those of GBS (up to 40%), and higher than in non-pregnant women (~20%) because of the possible effects of oestrogens (Reid *et al.*, 1993) and higher urea levels in the vagina

during pregnancy (Kenny & Cartwright, 1977). GBS colonization rates are similar in pregnant and nonpregnant women (Hoshina *et al*., 1991; Manning, 2003).

The pathogenic potential of GBS and ureaplasmas does not guarantee that disease will occur with their presence (Larsen & Monif, 2001). In fact, they rarely cause disease – most colonized people are healthy. Although GBS is more virulent than ureaplasmas and is a leading cause of perinatal and maternal septicemia, the incidence of disease is grossly disproportional to that of colonization (Larsen & Monif, 2001). Studies of ureaplasma pathogenesis are more controversial (Razin *et al*., 1998). The high colonization rates and low morbidity and mortality rates of GBS and especially of ureaplasma infections suggest that they are not "true" pathogens but conditional pathogens (Goncalves *et al*., 2002; Razin *et al*., 1998; Schuchat, 1998).

8.2. GBS and human ureaplasmas are phylogenetically closely related.

Many studies, especially recent genome-based studies, suggest that *Streptococcus* spp. (in particular GBS) and human ureaplasmas are phylogentically closely related.

8.2.1. Low G+C Gram positive bacteria.

Analysis of full genome sequences confirmed the findings of previous conventional studies, that *U. parvum* (Glass *et al*., 2000) and GBS (Glaser *et al*., 2002; Tettelin *et al*., 2002) belong to the low G+C group of Gram positive bacteria – although ureaplasmas are not typically Gram positive, because they lack a cell wall (Razin *et al*., 1998). The G+C content of *U. parvum* (25.5%) is lower than that of closely related mollicutes, *M. genitalium* (32%) (Fraser *et al*., 1995) and *M. pneumoniae* (40%) (Himmelreich *et al*., 1996). Similarly, the G+C content of GBS (36%) is lower than those of *S. pyogenes* (38.5%) (Nakagawa *et al*., 2003) and *S. pneumoniae* (40%) (Hoskins *et al*., 2001; Glaser *et al*., 2002). The relatively low G+C content of both, within their corresponding families, supports the close relationship between

GBS and ureaplasmas. It is consistent with the hypothesis that ureaplasmas arose by reductive evolution (see below) from GBS and *M. pneumoniae* from *S. pneumoniae* or *S. pyogenes.* It is supported by the observation that human pathogenic mollicutes and *Streptococcus* spp. that affect the respiratory tract have higher G+C contents (*S. pneumoniae,* ~40% and *M. pneumoniae,* 40%) than those that affect the urogenital tract (*U. parvum,* 25.5%, *M. genitalium* 32% and GBS, ~36%). This suggests at least partially site-specific evolution and ecology (Chen & Zhang, 2003; Hurst & Merchant, 2001; Knight *et al*., 2001; Sandberg *et al*., 2003).

8.2.2. Genome size reorganization – evolution by reduction.

Phylogenomic studies show that, when bacterial lineages make the transition from free-living to obligate host-associated bacteria, they undergo a major loss of genes and DNA (Moran, 2002). Ureaplasmas (especially *U. parvum*) evolved from their low G+C content ancestor by a drastic reduction in genome size – from about 2.2 Mbp to about 0.8 Mbp – mainly through loss of many cell components (Rocha & Blanchard, 2002). Pre-genomic studies (mainly based on PFGE) showed that the GBS genome size vary from 2,030 to 2,290 kb (Dmitriev *et al*., 2002). The genome sizes of the four serovars of *U. parvum* are 760 kbp and of the ten serovars of *U. urealyticum,* 840 to 1,140 kb (Robertson *et al*., 1990). The full genome sequences of *U. parvum* (Glass *et al*., 2000) and GBS (Glaser *et al*., 2002; Tettelin *et al*., 2002) support the accuracy of previous PFGE genome size predictions (Robertson *et al*., 1990). In the process of "evolution by reduction" (Razin *et al*., 1998), *U. urealyticum* is less advanced, and 80-380 kbp longer, than *U. parvum*. We hypothesize that *U. urealyticum* has not discarded all of its “hostile” genes, which may explain clinical and pathological differences between *U. parvum* and *U. urealyticum* – the latter is a less common commensal, but more likely to be associated with some disease syndromes, such as urethritis (Povlsen *et al*., 2002; Deguchi *et al*., 2004).

Table 8.1. General features of GBS and ureaplasma genomes.

	GBS NEM316	**GBS 2603 V/R**	***U. parvum***
Length (bp)	2, 211, 485	2, 160, 267	751, 719
Genes (no.)	2, 118	2, 175	613
Gene density (gene no./kbp)	0.958	1.006	0.816
G+C (%)	35.6	35.7	25.5
tRNA (no.)	80	80	30
rRNA (no.)	7	7	2
rRNA gene order	16-23-5	16-23-5	16-23-5
rRNA located regions	455kbp	406kbp	203kbp
***cps* genes**	III-3	V	*epsG*
Major protein antigen	Alp2	Rib	MBA
Mobile genetic elements	IS*Sag2*	IS*1381*-IS*861*-IS*1548*-GBSil-IS*Sag1*-IS*Sag2*	None
Hemolysin genes	*cylE*	*cylE*	*hlyC, hlyA*
Urease genes	-	-	urease gene complex

Note.

epsG **–** exopolysaccharide biosynthesis, glycosyltransferase gene.

Other characteristics that support the concept of "evolution by reduction" are summarized in Table 8.2. Briefly, they are:

- all 7 GBS rRNA operons are within a fairly short region on the right replichore, whereas those of other *Streptococcus* spp. are on the both right and left replichores (including *S. pneumoniae*, in which rRNA operons are also confined to a short region 400 kbp) (Nakagawa *et al*., 2003; Tettelin *et al*., 2001). Both *U. parvum* rRNA operons are also on the right replichore (*M. pneumoniae* and *M. genitalium* have single rRNA operon on right replichore) (Fraser *et al*., 1995; Himmelreich *et al*., 1996) .
- ratios of 16S rRNA operon copy number to full genome size are higher for GBS and *U. parvum* than for other *Streptococcus* spp. and *Mycoplasma* spp., respectively.
- GBS has a significantly greater number of tRNA genes than the other *Streptococcus* spp. (Hoskins *et al*., 2001; Tettelin *et al*., 2001; Glaser *et al*., 2002; Tettelin *et al*., 2002; Nakagawa *et al*., 2003), most of the tRNA genes of GBS and *S. pyogenes* (but not *S. pneumoniae*) are on the positive strand. *U. parvum* has slightly fewer tRNA operons than *M. pneumoniae* and *M. genitalium*. Like GBS, the first tRNA operons are located on the positive strand, although they are not the majority. All of those located on the positive strand (the first 9 and one other) have corresponding tRNA operons among those of GBS that are on the positive strand. This suggests they may come from the same ancestor/source.

8.2.3. Indels.

Bacterial evolution is associated with continuous generation of novel genetic variants, resulting from point mutations, genetic rearrangements and horizontal gene transfer (Ziebuhr *et al*., 1999). Gene insertions and deletions or indels (Britten *et al*., 2003) are the major events underlying the emergence and evolution of bacterial pathogens and symbionts (Britten *et al*., 2003; Ochman & Moran, 2001).

Table 8.2. Some similar characteristics of GBS and *U. parvum* and their closely related species that support the concept of reductive evolution from *Streptococcus* to *Mollicutes*.

	Genome size (kbp)	16S rRNA		tRNA	
		Right or left replichores	**N/genome size**	**N=**	**N= positive strand**
GBS	2,030-2,290	7R (455 kbp)	0.0032	80	first 69
S. pyogenes	1,852-1,901	4R+1L or 4R+2L (750 kbp)	0.0026-0.0032	57-60	first 42-49
S. pneumoniae	2,039-2,161	3R+1L (400 kbp)	0.0018-0.0019	58	first 4-5
U. parvum	752	2R (203 kbp)	0.0026	30	first 9[a]
U. urealyticum	840-1,140	-	0.0018-0.0023	-	-
M. pneumoniae	816	1R	0.0012	35	first 4-5, negative strand
M. genitalium	580	1R	0.0017	32	-

Note.

a. Although a minority of *U. parvum* tRNA genes was in the positive strand, all 10 (including the first 9) were among the first 69 *S. agalactiae* tRNA genes.

The analysis of indels shows excellent correlation with phylogeny based on 16S rRNA for nearly all species (Gupta & Griffiths, 2002). Indel analysis clearly places *Streptococcus* spp. and *U. parvum* in the low G+C Gram-positive group (Gupta & Griffiths, 2002) (See Figure 8.1., which was kindly provided by Dr Gupta).

Different functional categories of genes evolve at significantly different rates. It is suggested that nonessential genes drive evolutionary diversification (Jordan *et al*., 2002). If so, the “backbone” of the genome (including essential genes) should be conserved. This means that, after *M. genitalium*, *U. parvum* would be more useful than GBS in defining the minimum genome (Fraser *et al*., 1995), because more of its genes are “essential” (Jordan *et al*., 2002).

Long-term processes leading to the development of new species or subspecies are termed macroevolution, and short-term developments, which occur during days or weeks, are considered to be microevolution (Morschhauser *et al*., 2000). Indels apparently contribute to both macroevolution and microevolution (Gupta & Griffiths, 2002). In particular, it appears that significant deletions or "evolution by reduction" (macroevolution) led to *U. parvum* and *U. urealyticum* evolving from a GBS/ureaplasma common ancestor. In addition both deletions and insertions (especially pathogenicity islands) (Glaser *et al*., 2002) would have contributed to microevolutionary diversity and possibly the development of virulent clones of GBS (Jones *et al*., 2003).

8.2.4. Supertree.

Although indels and lateral gene transfer (LGT) (Ziebuhr *et al*., 1999) contribute to significant intra-species differences (Morschhauser *et al*., 2000), analysis of the highly conserved core genes or backbone of a species is likely to be more useful to infer bacterial phylogeny, as shown in our GBS comparative genomic study (Daubin *et al*., 2002). The “supertree” based on core genes confirms the positions of

Figure 8.1. (according to Daubin *et al*., 2002) The branching order of the main groups within *Bacteria* based on conserved indels present in various proteins. The thick arrows above the line show the evolutionary stages where the identified indels are postulated to have been introduced. The phylogenetic placements of species from completed bacterial genomes based on these signatures are as shown. The model predicts that once a signature has been introduced in the main lineage, all groups to the right of it should contain the signature whereas all groups preceding it should lack it. The model is strongly supported by the sequence data for completed bacterial genomes with only 11 exceptions from the predicted pattern observed in 1,842 observations. Abbreviations in the protein names are: PRPP synthetase – phosphoribosyl pyrophosphate synthetase; PAC-transformylase –5-aminoimidazole-4-carboxamide formyltransferase.

Streptococcus spp. (in particular GBS – personal communication with Dr Daubin suggested that GBS and *S. pyogenes* is in the neighbour branch) and *Mollicute* spp. (including *U. parvum*) in two neighbouring clusters of the low G+C Gram-positive bacteria (Daubin *et al*., 2002).

8.3. Evolution – not only useful for genome mining?

In the postgenomic era, much has been gained from combined and integrated genomic and evolutionary studies of human pathogenic bacteria (Fraser *et al*., 2000; Eisen & Fraser, 2003). Could we go another step further? Analysis of the evolution of GBS and human ureaplasmas may help us to understand how and perhaps why similarities and differences – presence or absence of genes, DNA substitution in noncoding regions and global patterns of synteny (conserved gene order) – have arisen across species (Eisen & Fraser, 2003).

8.3.1. Can we attribute some "wisdom" to GBS and human ureaplasmas?

Improving our understanding of the biomedical significance of GBS and human ureaplasmas may require changing the way we look at them (Woolhouse *et al*., 2002). This may show how a wide range of major, seemingly unrelated issues in the study of GBS and human ureaplasmas may be accounted for in a simple and unified manner, from an evolutionary perspective (Almogy *et al*., 2002). Researchers also need to show "wisdom", because the evolutionary perspective is itself a theory or hypothesis rather than "fact" or "truth". If we postulate that organisms – specifically GBS and human ureaplasmas – have some "wisdom" (at least as imposed by us), the hypothesis may be much easier to understand. An analysis of evolutionary "wisdom" can be used to understand bacterial metabolism, pathogenicity, physiology and behavior (Eisen & Fraser, 2003). For example, although co-evolution between host and pathogen has been difficult to prove rigorously, in practice it can have a major influence on the interpretation of genetic variation in biomedically important traits (Woolhouse *et al*., 2002).

A

Borrelia burgdorferi
Treponema pallidum
Spirochaetes
Chlamydophila pneumoniae
Chlamydia muridarum
Chlamydia trachomatis
Chlamydiales
Aquifex aeolicus
Thermotoga maritima
Hyperthermophilic bacteria
Synechocystis sp.
Cyanobacteria
Ureaplasma parvum
Mycoplasma genitalium
Mycoplasma pneumoniae
Bacillus halodurans
Bacillus subtilis
Staphylococcus aureus
Lactococcus lactis
Streptomyces pyogenes
Low G+C Gram-positives
Deinococcus radiodurans
Deinococcus/Thermus
Streptomyces coelicolor
Mycobacterium tuberculosis
Mycobacterium leprae
High G+C Gram positives
Campylobacter jejuni
Helicobacter pylori
Neisseria meningitidis
Xylella fastidiosa
Pseudomonas aeruginosa
Buchnera sp.
Escherichia coli
Vibrio cholerae
Pasteurella multocida
Haemophilus influenzae
Caulobacter crescentus
Rickettsia prowazekii
Proteobacteria
Aeropyrum pernix
Sulfolobus solfataricus
Crenarchaeota
Pyrococcus horikoshii
Pyrococcus abyssi
Halobacterium sp.
Archaeoglobus fulgidus
Methanobacterium thermoautotrophicum
Methanococcus jannaschii
Thermoplasma acidophilum
Euryarchaeota
Arabidopsis thaliana
Caenorhabditis elegans
Drosophila melanogaster
Saccharomyces cerevisiae
Eukaryota

Figure 8.2. Supertrees of 45 species constructed with 730 trees. Supertree based on trees made by BIONJ and a gamma distribution estimation of evolutionary rate heterogeneity. Personal communication with Dr Daubin suggested that *Streptococcus agalactiae* and *Streptococcus pyogenes* is in the neighbour branch.

8.3.2. Understanding the "wisdom" of GBS and human ureaplasmas through their population genetic structures.

We hypothesize that the evolutionary potential of GBS and human ureaplasmas is reflected in their population genetic structure (McDonald & Linde, 2002). We could even consider using population genetic structure to evaluate the "wisdom" of their evolution. In relation to pregnant women, human ureaplasmas apparently are more successful than GBS, because they colonize a much greater proportion of the population – up to 80% compared with 40% or less for GBS. *U. parvum* is a more common colonizer than *U. urealyticum* – their relative ratio is about 4 to 1 (Kim *et al*., 2003; Schuchat, 1998) – indicating that *U. parvum* is more "successful" than *U. urealyticum* in coevolution (Woolhouse *et al*., 2002).

8.3.3. Population structure of pathogens: the role of immune selection.

It is interesting to consider how the interplay between a pathogen (Almogy *et al*., 2002) and host immune responses, particularly to conserved and variable antigens, shapes the pathogen's population structure (Gupta & Anderson, 1999). At one extreme, immune selection against polymorphic determinants can cause pathogen populations to self-organize spontaneously into discrete antigenic types that may either be maintained over long periods or undergo cyclical or chaotic fluctuations. At the other extreme, diversity may be drastically reduced by competition, induced by a strong immune response against a conserved determinant, because the many different "clones" can be eliminated in the mean and very short time (Gupta & Anderson, 1999).

GBS is more virulent, stimulates a stronger host immune response and has a relatively lower population size in pregnant women, than ureaplasmas. The parasitic life style and molecular mimicry adopted by ureaplasmas, to avoid stmulating the host immune attack, means that they colonise a higher proportion of the population (Baseman & Tully, 1997). The N-terminal ends of the major surface antigens of

GBS (C alpha and Rib) and human ureaplasmas (MBA) are more important than their variable regions. GBS hide their variable regions by deletion of tandem repeat elements (Madoff *et al*., 1996) and ureaplasmas hide theirs by phase variation (Monecke *et al*., 2003), ureaplasmas (at least U. *parvum* serovar 3) also mimic human host antigens as well as by varying the number of repeats (Baseman & Tully, 1997).

The repetitive units of the major surface antigens of both GBS and human ureaplasmas have multiple different formats (Zheng *et al*., 1995; Zheng *et al*., 1996; Lachenauer *et al*., 2000). Because they are located at surface sites and recognised by the host (Zheng *et al*., 1994; Lachenauer *et al*., 2000), changing their format repeatedly allows them to to evade immune attack (Lachenauer *et al*., 2000; Monecke *et al*., 2003).

8.3.4. Common themes but different strategies – two interesting case studies?

Analysis of genomes of microbial pathogens has provided common themes relating to virulence, host adaptation and evolution – including lateral gene transfer (LGT), genome decay and antigenic variation among pathogens (Wren, 2000).

"Genome decay" is the process by which it is hypothesized that GBS and human ureaplasmas developed from a common ancestor. Loss of about half of the genome size (about 2.2 Mbp) of a common ancestor resulted in *U. urealyticum*, and loss of another small portion (80-380 kbp) in *U. parvum* (Razin *et al*., 1998). This resulted in loss of ureaplasma virulence genes, which are not essential for a parasite-like organism (Ochman & Moran, 2001; Razin *et al*., 1998).

"Lateral gene transfer (LGT)" in GBS is shown by the presence of mge and pathogenic islands (PIs) and is the process by which it can increase its virulence (Ochman & Moran, 2001; Glaser *et al*., 2002; Tettelin *et al*., 2002). The availability of complete genome sequences for multiple strains of GBS hopefully will allow

insight into what determines that certain bacterial strains are more pathogenic than others. Already, for GBS, the availability of two compete genomes, and others near completion, allow us to begin intraspecies comparative genomics (Kruger & Baier, 1997; Fraser *et al*., 2000; Whittam & Bumbaugh, 2002).

8.4. Repetitive sequences.

In contrast to the smaller genomes of obligate host-associated bacteria, like human ureaplasmas (Glass *et al*., 2000; Rocha & Blanchard, 2002), those of free-living bacteria, like GBS, often carry phages and other repetitive sequences that mediate genomic rearrangements (Glaser *et al*., 2002; Tettelin *et al*., 2002).

8.4.1. Repetitive sequences in protein antigens.

Genomic studies indicate that tandem-repeat gene polymorphisms are more common than is generally believed (Andersson *et al*., 2002). A catalogue of putative polymorphic repeats within transcribed sequences comprises a large set of potentially phenotypic or disease-causing loci (Andersson *et al*., 2002; Rocha & Blanchard, 2002). The major protein antigens of both GBS (family defined by Bca and Rib) and ureaplasmas (MBA) contain tandem-repeat regions (see below). Repetitive sequences are also found in other GBS but not *U. parvum* surface located proteins (Glass *et al*., 2000; Glaser *et al*., 2002; Tettelin *et al*., 2002).

8.4.2. Other genomic repeats.

Comparative genomic studies have shown that, among *Streptococcus* spp., *S. pyogenes* (3 genomes) contains 4-6 phages, which occupy 7.0-12.4% of genome length (Banks *et al*., 2002; Nakagawa *et al*., 2003) and *S. pneumoniae* contains many BOX (about 25 copies) (Martin *et al*., 1992) and RUP repetitive sequences (108 copies in serotype 4 genome) (Oggioni & Claverys, 1999; Hoskins *et al*., 2001; Tettelin *et al*., 2001). By contrast, although *S. agalactiae* contains several genomic islands (GIs) including possible pathogenic islands (PIs) (Glaser *et al*., 2002;

Tettelin *et al*., 2002), but no typical phages or repetitive sequences. This suggests that *S. agalactiae* is relatively more stable (or less flexible) than *S. pyogenes* and *S. pneumoniae* (Glaser *et al*., 2002; Tettelin *et al*., 2002; Nakagawa *et al*., 2003).

Rocha and Blanchard established a bioinformatic strategy to detect the major recombination "hot-spots" in the four published mollicute genomes (Rocha & Blanchard, 2002). *U. parvum* has fewer genomic repeats than *M. pneumoniae* (Himmelreich *et al*., 1996) and *M. genitalium* (Fraser *et al*., 1995; Rocha & Blanchard, 2002) – which suggests that it is less flexible than the two human *Mycoplasma* spp. (Mira *et al*., 2002).

The relative lack of repetitive sequences and genome stability, compared with their respective close relatives, is another common feature of GBS and human ureaplasmas.

8.5. Genotypes and diseases.

As in many other bacterial infections, there are a variety of disease manifestations and clinical outcomes for GBS and ureaplasma infections, the basis of which is not well understood. One factor that can contribute to occurrence, and differences in severity, of disease is the variation in virulence that exists among strains of a bacterial population (Whittam & Bumbaugh, 2002). The intra-species variability in gene content, genomic organization, and gene expression may account for variation in the severity of disease and for the diverse clinical outcomes of infection (Finlay & Falkow, 1997; Whittam & Bumbaugh, 2002) .

8.5.1. Why genotyping study?

GBS serotypes are based on capsular polysaccharide and serosubtypes on major surface proteins; human ureaplasmas serotypes are based on surface proteins (see below). Both GBS and human ureaplasmas have multiple serotypes and, for both,

the traditional serotyping methods**,** based on panels of antisera, are not very practical for routine use. Therefore, we developed molecular serotype identification systems.

Previously, the bases of GBS protein antigen "serosubtyping**"** and ureaplasma serotyping targets were not very clear (Zheng *et al*., 1995; Zheng *et al*., 1996; Lachenauer *et al*., 1999; Lachenauer *et al*., 2000). To explain them more clearly, we also studied the GBS (Kong *et al*., 2002b) and ureaplasma protein antigen gene targets (Kong *et al*., 2000b).

The relationships between species (for ureaplasmas) and/or serotype (for both GBS and ureaplasma) and pathogenesis/virulence still have not been resolved, partly because of the difficulties associated with conventional serotyping (Abele-Horn *et al*., 1997; Heggie *et al*., 2001; Manning, 2003).

For GBS, a molecular serotyping method would be valuable and practical alternative for use in disease surveillance and future vaccine design (Jones *et al*., 2003).

8.5.2. Our genotyping study.

Our genotyping studies (Kong *et al*., 2002a; Kong *et al*., 2002b; Kong *et al*., 2003) showed that there were differences in the distribution of GBS serotypes based on geographic area and patient age. GBS serosubtype III-2, in our system, is the high virulence clone previously recognised by others (Musser *et al*., 1989; Takahashi *et al*., 1998; Kong *et al*., 2003). Differences in distribution of ureaplasma serotypes are mainly determined by species; for example, *U. parvum*, especially serovars 3 and 6, are the predominant species and serovar (Kong *et al*., 2000; Knox *et al*., 1997 & 2003).

8.5.3. Considerations for future genotyping.

Now that more specific and discriminatory genotyping methods are available, further clinical studies will be required to elucidate the relationships between

genotype and pathogenicity, for both human ureaplasmas and GBS (Heggie *et al*., 1994; Heggie *et al*., 2001; Schuchat, 1998). Although serotype III (III-2 in particular) has been shown to be the most virulent serotype, all serotypes can cause infection. More discriminatory tools are needed to identify more specific virulence markers, based on data mined from multiple GBS genomes to develop genechip microarrays (or equilvalent technology) (Glaser *et al*., 2002; Tettelin *et al*., 2002). Because human ureaplasmas are less virulent than GBS, it is more difficult to draw conclusions about the relationship between serotypes and virulence; less definite associations, compared with GBS, are to be expected. How to deal with this problem is still a challenge. Host factors are important in pathogenesis of infection due to both organisms, but probably more so in ureaplasma infections.

8.6. Comparative genomics – GBS and human ureaplasmas?

While analysis of a single genome provides tremendous biological insights into any given organism, comparative analyses of multiple genomes will provide substantially more information (Fraser *et al*., 2000). As comparative genomics shifts from inter-species to intra-species differences, could we shift to comparing clusters of strains causing different types of infection (perinatal, urogenital tract, respiratory tract, etc.)? The transition from free-living or facultatively parasitic life cycles to permanent associations with hosts involves a major loss of genes and DNA (Rocha & Blanchard, 2002). By mining the respective genomes of ureaplasmas and the bacteria from which they are believed to have evolved, could we relate the loss of genome content to their reduced virulence? In future, the availability of the *U. urealyticum* genome should identify the reason for the 80-380 kbp size difference between it and *U. parvum*, and may allow to explain other differences between them.

8.6.1. Major surface protein antigens.

The major surface protein antigens of human ureaplasmas (MBA) and GBS (Bca/Rib/Alp2-5) are structurally similar (Zheng *et al.*, 1995; Lachenauer *et al.*, 2000; Schuchat, 1998), but there are differences in their repetitive sequences. A sequence homology search showed that the repetitive units of *U. parvum* serovar 3 MBA repetitive units have far more similarities to mammalian collagen/collagen-like proteins than other human ureaplasma and GBS surface proteins. Up to 100 matched sequences were found in *U. parvum* serovar 3, compared with 0-12 in *U. urealyticum* serovar 10 MBA and GBS Bca/Rib/Alp2-3 repetitive units. A previous study in our laboratory showed that monoclonal antibody against *U. parvum* serovar 3 MBA repetitive units, but not against those of other ureasplasma serovars, cross-reacted with normal rabbit lung tissue (Kirsty Hannaford-Turner, 2002 PhD thesis, University of Sydney). This provided indirect evidence supporting the sequence search results (Hance *et al.*, 1976).

As previously discussed, many more pregnant women carry ureaplasmas than GBS. The ratio of *U. parvum* and *U. urealyticum* carriage is 4 to 1 and *U. parvum* serotype 3 is the predominant serotype; *U. urealyticum* serotypes 7, 11 and 9 are rare. We hypothesise that the similarities between surface protein repetitive units and human connective tissue antigens have developed during coevolution and have influenced ureaplasma genetic population structure. This hypothesis is supported by the greater similarity of the predominant *U. parvum* serovar 3, and the fact that *U. urealyticum* serotypes 7, 11 and 9, which lack repetitive sequences, have the least similarity with human proteins.

8.6.2. Capsule and capsular polysaccharide.

GBS capsular polysaccharide is its major virulence factor. Human ureaplasmas also have a capsule-like structure or exopolysaccharide, which contains glucosyl-like resides (Whitescarver *et al.*, 1975; Robertson & Smook, 1976). In ureaplasmas, *epsG* encodes glycosyltransferase, which is involved in exopolysaccharide biosynthesis. There is no *cps* gene cluster, similar to that in GBS, in *U. parvum* genome (Glass *et al.*, 2000), but 40-68% homology between *eps* and *cps* genes has

been reported (Stingele *et al*., 1996). Again, this supports the concept of "degenerate or reverse evolution" of human ureaplasmas from gram-positive/*Streptococcus* spp. (Razin *et al*., 1998). Unlike GBS, the capsule-like polysaccharide of ureaplasmas is not the serotype-specific antigen.

Table 8.3. Sequence comparison of human ureaplasma and GBS major surface protein antigens.

Protein name	GenBank No.	Whole length	Length of N-terminal	Length of repetitive unit
MBA-Up-3	L202329	409 aa	151 aa	6 aa x 42.5
MBA-Uu-10	U50459	487 aa	147 aa	8 aa x 42.3
Bca	M97256	1020 aa	172 aa	82 aa x 9
Rib	U58333	1231 aa	172 aa	79 aa x 12
Alp2	AF208158	786 aa	172 aa	76 aa x 3
Alp3	AF245663	865 aa	172 aa	79 aa x 5

8.6.3. Mobile genetic elements (mge).

Comparative genomics has shown that *S. agalactiae* contains a large number of mge – including insertion sequences (IS), group II introns (Kong *et al*., 2003) and prophage-like structures (Glaser *et al*., 2002; Tettelin *et al*., 2002). No mge were found in the *U. parvum* genome (Glass *et al*., 2000) but a Tn*1545*-like transposon (Tn*916*) was found in tetracycline resistant ureaplasmas that carry *tetM* gene (de Barbeyrac *et al*., 1996; Taraskina *et al*., 2002). It would be of interest to establish whether the extra 80-380 kbp of *U. urealyticum* contains mge (de Barbeyrac *et al*., 1996).

8.7. GBS and human ureaplasma – taxonomy case studies supporting reconsideration of species definition.

8.7.1. The case for separation of *U. parvum* and *U. urealyticum*.

DNA-DNA hybridization shows only 60% homology between *U. parvum* and *U. urealyticum* (Chrisstiansen *et al.*, 1981; Harasawa *et al.*, 1991) and they have also different genome sizes (Robertson *et al*., 1990). Divergent nucleotide sequences of several highly conserved genes, which are identifiable by biovars/species-specific PCR targeting 16S rRNA, the 16S-23S rRNA intergenic regions, the genus-defining urease or the serovar-defining multiple-banded antigen (MBA) genes, also attest to their phylogenetic distinctiveness. Further evidence for establishment of two separate species – *U. parvum* and *U. urealyticum* – is provided by:

- distinctive RFLP and randomly amplified polymorphic DNA [RAPD] typing patterns;
- phenotypic differences (including clustering of antigenic types, polypeptide patterns of whole-cell preparations, differential inhibition by manganese, and polymorphism among their ureases, pyrophosphatases, diaphorases) (Robertson *et al*., 2002).

8.7.2. The case for combination of *S. agalactiae* and *S. difficile*.

Genotypic studies, including our study of the partial *cps* gene cluster, serotype Ib-specific PCR positive result (Kong *et al*., 2002a), and phenotypic studies, including whole-cell protein electrophoretic analysis and serological studies, showed that *S. difficile* is actually an atypical *S. agalactiae* (GBS) type Ib (Vandamme *et al*., 1997).

These two examples highlight some controversies in definitions of bacteria species (Robertson *et al*., 2002; Vandamme *et al*., 1997) and suggest that some issues are still unresolved. The polyphasic theory is an important contribution and deserves wider acceptance (Vandamme *et al*., 1996).

8.8. Prevention – eradicate or conserve to prevent disease?

Overwhelming evidence has demonstrated that GBS can be harmful to humans and there is some evidence that ureaplasmas are also harmful in some circumstances, although less hostile than GBS (Razin *et al*., 1998; Schuchat, 1998). Because of their conditional pathogenicity, the wisdom of attempting to eradicate colonization has been questioned because:

1) Only a very small proportion of GBS and human ureaplasma colonized persons will develop significant disease.
2) GBS and human ureaplasma strains are highly diverse at a genetic level and vary in virulence.
3) The antiquity of GBS and human ureaplasma infection in humans and their co-evolution suggests that both, but especially ureaplasmas, are commensal to humans. Their eradication may replace benign or even beneficial strains with more harmful ones and may provoke other problems.

It remains for carefully designed prospective studies, rather than hypotheses, to determine whether the potential risks of eradication of commensals, which are also

potential pathogens, outweigh the benefits (Hunt *et al*., 2001). Careful review of the literature confirms that GBS infection is a serious pathogen, albeit in a minority of those colonized. Intrapartum antibiotic prophylaxis (IAP) will be replaced by vaccines to prevent neonatal GBS infection (Paoletti & Madoff, 2002). Ideally, vaccines (or antibiotic prophylaxis) would be targeted at the small proportion of GBS carriers whose infants are most at risk.

For human ureaplasmas the issue is more difficult. There is no suitable bactericidal antibiotic and vaccination is unlikely to be appropriate because of the cross reactivity or mimicry between MBA and human antigens (although a DNA vaccine could theoretically bypass this problem) (Mabanta *et al*., 2003).

8.9. Conclusions.

Human ureaplasmas (*U. parvum* and *U. urealyticum*) (Razin *et al*., 1998) and group B streptococcus (GBS, *S. agalactiae*) (Schuchat, 1998) are both common potential perinatal pathogens. Studying them together provides more insight into their evolution and the pathogenesis of infection than studying them separately (van Belkum *et al*., 2001; Almogy *et al*., 2002).

Evolutionary and population genetics suggest that human ureaplasmas are "cleverer" than GBS (Razin *et al*., 1998) considering their ratio of "investment" (e.g. genome size) to "outcome" (e.g. different prevalence and bacterial population levels in pregnant women). During their coevolution with human beings, ureaplasmas apparently thrived in their "parasite-like" lifestyle and have finally reached a successful co-existence, with high population levels for their limited genomic resources. This required reduction in virulence – by loss of most of their capsule and mimicry by MBA repetitive units of host antigens – and becoming more parsiminonious and "friendly" – by using a host waste product (urea) as their energy source. This parasitic or symbiotic relationship contributes to controversy associated with study of human ureaplasma infection (Razin *et al*., 1998).

GBS have adopted another strategy. They have increased their virulence by retaining and acquiring new virulence molecules: capsule; surface protein family and IgA binding protein (C beta); mge and a variety of other specific virulence factors (Glaser *et al*., 2002). But considering their genome size and their population, they are not as wise (efficient or successful) as human ureaplasmas.

Reference List

1. **Abele-Horn, M., C. Wolff, P. Dressel, F. Pfaff, and A. Zimmermann**. 1997. Association of *Ureaplasma urealyticum* biovars with clinical outcome for neonates, obstetric patients, and gynecological patients with pelvic inflammatory disease. J.Clin.Microbiol. **35**:1199-1202.

2. **Abele-Horn, M., O. Genzel-Boroviczeny, T. Uhlig, A. Zimmermann, J. Peters, and M. Scholz**. 1998. *Ureaplasma urealyticum* colonization and bronchopulmonary dysplasia: a comparative prospective multicentre study. Eur.J.Pediatr. **157**:1004-1011.

3. **Abele-Horn, M., M. Scholz, C. Wolff, and M. Kolben**. 2000. High-density vaginal *Ureaplasma urealyticum* colonization as a risk factor for chorioamnionitis and preterm delivery. Acta Obstet.Gynecol.Scand. **79**:973-978.

4. **Abu-Shakra, M., D. Buskila, and Y. Shoenfeld**. 1999. Molecular mimicry between host and pathogen: examples from parasites and implication. Immunol.Lett. **67**:147-152.

5. **Adderson, E. E., S. Takahashi, and J. F. Bohnsack**. 2000. Bacterial genetics and human immunity to group B streptococci. Mol.Genet.Metab. **71**:451-454.

6. **Agosti, D., D. Jacobs, and R. DeSalle**. 1996. On combining protein sequences and nucleic acid sequences in phylogenetic analysis: the homeobox protein case. Cladistics. **12**:65-82.

7. **Ahmet, Z., P. Stanier, D. Harvey, and D. Holt**. 1999. New PCR primers for the sensitive detection and specific identification of group B beta-hemolytic streptococci in cerebrospinal fluid. Mol.Cell.Probes **13**:349-357.

8. **Almogy, G., N. Cohen, S. Stocker, and L. Stone**. 2002. Immune response and virus population composition: HIV as a case study. Proc.R.Soc.Lond.B.Biol.Sci. **269**:809-815.

9. **Amaya, R. A., C. J. Baker, W. A. Keitel, and M. S. Edwards**. 2004. Healthy elderly people lack neutrophil-mediated functional activity to type V group B streptococcus. J.Am.Geriatr.Soc. **52**:46-50.

10. **Andersson, S. G., C. Alsmark, B. Canback, W. Davids, C. Frank, O. Karlberg, L. Klasson, B. Antoine-Legault, A. Mira, and I. Tamas**. 2002. Comparative genomics of microbial pathogens and symbionts. Bioinformatics. **18 Suppl 2**:S17.

11. **Ang, C. W., A. P. Tio-Gillen, J. Groen, P. Herbrink, B. C. Jacobs, R. Van Koningsveld, A. D. Osterhaus, F. G. Van der Meche, and P. A. van Doorn**. 2002. Cross-reactive anti-galactocerebroside antibodies and *Mycoplasma pneumoniae* infections in Guillain-Barre syndrome. J.Neuroimmunol. **130**:179-183.

12. **Arakere, G., A. E. Flores, P. Ferrieri, and C. E. Frasch**. 1999. Inhibition enzyme-linked immunosorbent assay for serotyping of group B streptococcal isolates. J.Clin.Microbiol. **37**:2564-2567.

13. **Baier, R. J., J. Loggins, and T. E. Kruger**. 2003. Failure of erythromycin to eliminate airway colonization with *Ureaplasma urealyticum* in very low birth weight infants. BMC.Pediatr. **3**:10.

14. **Baker, C. J. and D. L. Kasper**. 1976. Correlation of maternal antibody deficiency with susceptibility to neonatal group B streptococcal infection. N.Engl.J.Med. **294**:753-756.

15. **Banks, D. J., S. B. Beres, and J. M. Musser**. 2002. The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. Trends Microbiol. **10**:515-521.

16. **Baseman, J. B., S. P. Reddy, and S. F. Dallo**. 1996. Interplay between mycoplasma surface proteins, airway cells, and the protean manifestations of mycoplasma-mediated human infections. Am.J.Respir.Crit Care Med. **154**:S137-S144.

17. **Baseman, J. B. and J. G. Tully**. 1997. *Mycoplasmas*: sophisticated, reemerging, and burdened by their notoriety. Emerg.Infect.Dis. **3**:21-32.

18. **Basham, L. E., V. Pavliak, X. Li, A. Hawwari, K. L. Kotloff, R. Edelman, and A. Fattom**. 1996. A simple, quantitative, reproducible avidin-biotin ELISA for the evaluation of group B streptococcus type-specific antibodies in humans. Vaccine **14**:439-445.

19. **Beckert, S., B. Kreikemeyer, and A. Podbielski**. 2001. Group A streptococcal *rofA* gene is involved in the control of several virulence genes and eukaryotic cell attachment and internalization. Infect.Immun. **69**:534-537.

20. **Berner, R., A. Bender, C. Rensing, J. Forster, and M. Brandis**. 1999. Low prevalence of the immunoglobulin-A-binding beta antigen of the C protein among *Streptococcus agalactiae* isolates causing neonatal sepsis. Eur.J.Clin.Microbiol.Infect.Dis. **18**:545-550.

21. **Berner, R., J. Csorba, and M. Brandis**. 2001. Different cytokine expression in cord blood mononuclear cells after stimulation with neonatal sepsis or colonizing strains of *Streptococcus agalactiae*. Pediatr.Res. **49**:691-697.

22. **Berner, R., M. Ruess, S. Bereswill, and M. Brandis**. 2002. Polymorphisms in the cell wall-spanning domain of the C protein beta-antigen in clinical *Streptococcus agalactiae* isolates are caused by genetic instability of repeating DNA sequences. Pediatr.Res. **51**:106-111.

23. **Bert, F., B. Picard, N. Lambert-Zechovsky, and P. Goullet**. 1995. Identification and typing of pyogenic streptococci by enzyme electrophoretic polymorphism. J.Med.Microbiol. **42**:442-451.

24. **Bevanger, L., A. I. Kvam, and J. A. Maeland**. 1995. A *Streptococcus agalactiae* R protein analysed by polyclonal and monoclonal antibodies. APMIS **103**:731-736.

25. **Bidet, P., N. Brahimi, C. Chalas, Y. Aujard, and E. Bingen**. 2003. Molecular characterization of serotype III group B-streptococcus isolates causing neonatal meningitis. J.Infect.Dis. **188**:1132-1137.

26. **Blanchard, A. and M. F. Barile**. 1989. Cloning of *Ureaplasma urealyticum* DNA sequences showing genetic homology with urease genes from gram-negative bacteria. Research in Microbiology **140**:281-290.

27. **Blanchard, A.** 1990. *Ureaplasma urealyticum* urease genes; use of a UGA tryptophan codon. Mol.Microbiol. **4**:669-676.

28. **Blue, C. E., G. K. Paterson, A. R. Kerr, M. Berge, J. P. Claverys, and T. J. Mitchell**. 2003. ZmpB, a novel virulence factor of *Streptococcus pneumoniae* that induces tumor necrosis factor alpha production in the respiratory tract. Infect.Immun. **71**:4925-4935.

29. **Blumberg, H. M., D. S. Stephens, C. Licitra, N. Pigott, R. Facklam, B. Swaminathan, and I. K. Wachsmuth**. 1992. Molecular epidemiology of group B streptococcal infections: use of restriction endonuclease analysis of chromosomal DNA and DNA restriction fragment length polymorphisms of ribosomal RNA genes (ribotyping). J.Infect.Dis. **166**:574-579.

30. **Bohnsack, J. F., S. Takahashi, S. R. Detrick, L. R. Pelinka, L. L. Hammitt, A. A. Aly, A. A. Whiting, and E. E. Adderson**. 2001. Phylogenetic classification of serotype III group B streptococci on the basis of *hylB* gene analysis and DNA sequences specific to restriction digest pattern type III-3. J.Infect.Dis. **183**:1694-1697.

31. **Bohnsack, J. F., A. A. Whiting, R. D. Bradford, B. K. Van Frank, S. Takahashi, and E. E. Adderson**. 2002. Long-range mapping of the *Streptococcus agalactiae* phylogenetic lineage restriction digest pattern type III-3 reveals clustering of virulence genes. Infect.Immun. **70**:134-139.

32. **Brakstad, O. G., L. Bevanger, and A. I. Kvam**. 1997. *Streptococcus agalactiae* beta gene and gene product variations. J.Med.Microbiol. **46**:999-1005.

33. **Britten, R. J., L. Rowen, J. Williams, and R. A. Cameron**. 2003. Majority of divergence between closely related DNA samples is due to indels. Proc.Natl.Acad.Sci.U.S.A **100**:4661-4665.

34. **Brodeur, B. R., M. Boyer, I. Charlebois, J. Hamel, F. Couture, C. R. Rioux, and D. Martin**. 2000. Identification of group B streptococcal Sip protein, which elicits cross-protective immunity. Infect.Immun. **68**:5610-5618.

35. **Bromham, L. and D. Penny**. 2003. The modern molecular clock. Nat.Rev.Genet. **4**:216-224.

36. **Bull, A. T., A. C. Ward, and M. Goodfellow**. 2000. Search and discovery strategies for biotechnology: the paradigm shift. Microbiol.Mol.Biol.Rev. **64**:573-606.

37. **Bush, R. M. and K. D. Everett**. 2001. Molecular evolution of the *Chlamydiaceae*. Int.J.Syst.Evol.Microbiol. **51**:203-220.

38. **Campbell, J. R., S. L. Hillier, M. A. Krohn, P. Ferrieri, D. F. Zaleznik, and C. J. Baker**. 2000. Group B streptococcal colonization and serotype-specific immunity in pregnant women at delivery. Obstet.Gynecol. **96**:498-503.

39. **Cassell, G. H., D. T. Crouse, K. B. Waites, P. T. Rudd, and J. K. Davis**. 1988. Does *Ureaplasma urealyticum* cause respiratory disease in newborns?. Pediatr. Infect. Dis. J. **7**:535-541.

40. **Castro-Alcaraz, S., E. M. Greenberg, D. A. Bateman, and J. A. Regan**. 2002. Patterns of colonization with *Ureaplasma urealyticum* during neonatal intensive care unit hospitalizations of very low birth weight infants and the development of chronic lung disease. Pediatrics **110**:e45.

41. **Chaffin, D. O., S. B. Beres, H. H. Yim, and C. E. Rubens**. 2000. The serotype of type Ia and III group B streptococci is determined by the polymerase gene within the polycistronic capsule operon. J.Bacteriol. **182**:4466-4477.

42. **Chatellier, S., C. Ramanantsoa, P. Harriau, K. Rolland, A. Rosenau, and R. Quentin**. 1997. Characterization of *Streptococcus agalactiae* strains by randomly amplified polymorphic DNA analysis. J.Clin.Microbiol. **35**:2573-2579.

43. **Chen, L. L. and C. T. Zhang**. 2003. Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages. Biochem.Biophys.Res. Commun. **306**:310-317.

44. **Cheng, X., A. Naessens, and S. Lauwers**. 1994. Identification of serotype 1-, 3-, and 6-specific antigens of *Ureaplasma urealyticum* by using monoclonal antibodies. J.Clin.Microbiol. **32**:1060-1062.

45. **Chhatwal, G. S.** 2002. Anchorless adhesins and invasins of Gram-positive bacteria: a new class of virulence factors. Trends Microbiol. **10**:205-208.

46. **Christiansen, C., F. T. Black, and E. A. Freundt.** 1981. Hybridization experiments with deoxyribonucleic acid from *Ureaplasma urealyticum* serovars I to VIII. Int.J.Syst.Bacteriol. **31**:259-262.

47. **Cieslewicz, M. J., D. L. Kasper, Y. Wang, and M. R. Wessels**. 2001. Functional analysis in type Ia group B streptococcus of a cluster of genes involved in extracellular polysaccharide production by diverse species of streptococci. J.Biol.Chem. **276**:139-146.

48. **Clayton, R. A., G. Sutton, P. S. Hinkle, Jr., C. Bult, and C. Fields**. 1995. Intraspecific variation in small-subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa. Int.J.Syst.Bacteriol. **45**:595-599.

49. **Cropp, C. B., R. A. Zimmerman, J. Jelinkova, A. H. Auernheimer, R. A. Bolin, and B. C. Wyrick**. 1974. Serotyping of group B streptococci by slide agglutination fluorescence microscopy, and microimmunodiffusion. J.Lab Clin.Med. **84**:594-603.

50. **Cunliffe, N. A., S. Fergusson, F. Davidson, A. Lyon, and P. W. Ross**. 1996. Comparison of culture with the polymerase chain reaction for detection of *Ureaplasma urealyticum* in endotracheal aspirates of preterm infants. J. Med. Microbiol. **45**:27-30.

51. **Cunningham, C. K., C. A. Bonville, J. H. Hagen, J. L. Belkowitz, R. M. Kawatu, Higgins, AM, and L. B. Weiner**. 1996. Immunoblot analysis of anti-*Ureaplasma urealyticum* antibody in pregnant women and newborn infants. Clin.Diagn.Lab.Immunol. **3**:487-492.

52. **Dalma-Weiszhausz, D. D., M. E. Chicurel, and T. R. Gingeras**. 2002. Microarrays and genetic epidemiology: a multipurpose tool for a multifaceted field. Genet.Epidemiol. **23**:4-20.

53. **Daubin, V., M. Gouy, and G. Perriere**. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res. **12**:1080-1090.

54. **Davies, H. D., C. Adair, A. McGeer, D. Ma, S. Robertson, M. Mucenski, L. Kowalsky, G. Tyrell, and C. J. Baker**. 2001. Antibodies to capsular polysaccharides of group B streptococcus in pregnant Canadian women: relationship to colonization status and infection in the neonate. J.Infect.Dis. **184**:285-291.

55. **Daxboeck, F., E. Iro, K. Tamussino, R. Krause, O. Assadian, and C. Wenisch**. 2003. Bacteremia with *Mycoplasma hominis* and *Ureaplasma urealyticum* in patients undergoing hysterectomy. Eur.J.Clin.Microbiol.Infect.Dis.

56. **de Barbeyrac, B., M. Dupon, P. Rodriguez, H. Renaudin, and C. Bebear**. 1996. A Tn*1545*-like transposon carries the *tet(M)* gene in tetracycline resistant strains of *Bacteroides ureolyticus* as well as *Ureaplasma urealyticum* but not *Neisseria gonorrhoeae*. J. Antimicrob. Chemother. **37**:223-232.

57. **Deguchi, T., T. Yoshida, T. Miyazawa, M. Yasuda, M. Tamaki, H. Ishiko, and S. Maeda**. 2004. Association of *Ureaplasma urealyticum* (biovar 2) with nongonococcal urethritis. Sex Transm.Dis. **31**:192-195.

58. **Dellaglio, F., G. E. Felis, and J. E. Germond.** 2004. Should names reflect the evolution of bacterial species? Int.J.Syst.Evol.Microbiol. **54**:279-281.

59. **Dmitriev, A., Y. Y. Hu, A. D. Shen, A. Suvorov, and Y. H. Yang**. 2002. Chromosomal analysis of group B streptococcal clinical strains; *bac* gene-positive strains are genetically homogenous. FEMS Microbiol.Lett. **208**:93-98.

60. **Dmitriev, A., A. Suvorov, and A. Totolian**. 1998. Physical and genetic chromosomal maps of *Streptococcus agalactiae*, serotypes II and III; rRNA operon organization. FEMS Microbiol.Lett. **167**:33-39.

61. **Dobrindt, U. and J. Hacker**. 2001. Whole genome plasticity in pathogenic bacteria. Curr.Opin.Microbiol. **4**:550-557.

62. **Domingues, D., L. T. Tavira, A. Duarte, A. Sanca, E. Prieto, and F. Exposto**. 2002. *Ureaplasma urealyticum* biovar determination in women attending a family planning clinic in Guine-Bissau, using polymerase chain reaction of the multiple-banded antigen gene. J.Clin.Lab Anal. **16**:71-75.

63. **Echahidi, F., K. van Geel, S. Lauwers, and A. Naessens**. 2002. Comparison of two methods for serotyping *Ureaplasma urealyticum* clinical isolates. J.Microbiol.Methods **49**:157-161.

64. **Eisen, J. A. and C. M. Fraser**. 2003. Phylogenomics: intersection of evolution and genomics. Science. **300**:1706-1707.

65. **Enright, M. C. and B. G. Spratt**. 1999. Multilocus sequence typing. Trends Microbiol. **7**:482-487.

66. **Erdogan, S., P. K. Fagan, S. R. Talay, M. Rohde, P. Ferrieri, A. E. Flores, C. A. Guzman, M. J. Walker, and G. S. Chhatwal**. 2002. Molecular analysis of group B protective surface protein, a new cell surface protective antigen of group B streptococci. Infect.Immun. **70**:803-811.

67. **Farley, M. M.** 2001. Group B streptococcal disease in nonpregnant adults. Clin.Infect.Dis. **33**:556-561.

68. **Fasola, E. L., A. E. Flores, and P. Ferrieri**. 1996. Immune responses to the R4 protein antigen of group B streptococci and its relationship to other streptococcal R4 proteins. Clin.Diagn.Lab Immunol. **3**:321-325.

69. **Feikin, D. R., P. Thorsen, S. Zywicki, M. Arpi, J. G. Westergaard, and A. Schuchat**. 2001. Association between colonization with group B streptococci during pregnancy and preterm delivery among Danish women. Am.J.Obstet.Gynecol. **184**:427-433.

70. **Ferrieri, P.** 1988. Surface-localized protein antigens of group B streptococci. Rev.Infect. Dis. **10 Suppl 2**:S363-S366.

71. **Finlay, B. B. and S. Falkow**. 1997. Common themes in microbial pathogenicity revisited. Microbiol.Mol.Biol.Rev. **61**:136-169.

72. **Fitzgerald, J. R. and J. M. Musser**. 2001. Evolutionary genomics of pathogenic bacteria. Trends Microbiol. **9**:547-553.

73. **Foxman, B. and L. Riley**. 2001. Molecular epidemiology: focus on infection. Am.J.Epidemiol. **153**:1135-1141.

74. **Franken, C., G. Haase, C. Brandt, J. Weber-Heynemann, S. Martin, C. Lammler, A. Podbielski, R. Lutticken, and B. Spellerberg**. 2001. Horizontal gene transfer and host specificity of beta-haemolytic streptococci: the role of a putative composite transposon containing *scpB* and *lmb*. Mol.Microbiol. **41**:925-935.

75. **Fraser, C. M., J. Eisen, R. D. Fleischmann, K. A. Ketchum, and S. Peterson**. 2000. Comparative genomics and understanding of microbial biology. Emerg.Infect.Dis. **6**:505-512.

76. **Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, et al.** 1995. The minimal gene complement of *Mycoplasma genitalium*. Science **270**:397-403.

77. **Gabig, M. and G. Wegrzyn**. 2001. An introduction to DNA chips: principles, technology, applications and analysis. Acta Biochim.Pol. **48**:615-622.

78. **Galas, D. J. and S. J. McCormack**. 2003. An historical perspective on genomic technologies. Curr.Issues Mol.Biol. **5**:123-127.

79. **Garcia-Martinez, J., S. G. Acinas, A. I. Anton, and F. Rodriguez-Valera**. 1999. Use of the 16S--23S ribosomal genes spacer region in studies of prokaryotic diversity. J.Microbiol.Methods **36**:55-64.

80. **Gelfand, E. W.** 1993. Unique susceptibility of patients with antibody deficiency to mycoplasma infection. Clin.Infect.Dis. **17 Suppl 1**:S250-S253.

81. **Givner, L. B., L. Gray, and T. M. O'Shea**. 1995. Antibodies to tumor necrosis factor-alpha: use as adjunctive therapy in established group B streptococcal disease in newborn rats. Pediatr Res. **38**:551-554.

82. **Glaser, P., C. Rusniok, C. Buchrieser, F. Chevalier, L. Frangeul, T. Msadek, M. Zouine, E. Couve, L. Lalioui, C. Poyart, P. Trieu-Cuot, and F. Kunst**. 2002. Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. Mol.Microbiol. **45**:1499-1513.

83. **Glass, J. I., E. J. Lefkowitz, J. S. Glass, C. R. Heiner, E. Y. Chen, and G. H. Cassell**. 2000. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. Nature **407**:757-762.

84. **Goncalves, L. F., T. Chaiworapongsa, and R. Romero**. 2002. Intrauterine infection and prematurity. Ment.Retard.Dev.Disabil.Res.Rev. **8**:3-13.

85. **Granlund, M., F. Michel, and M. Norgren**. 2001. Mutually exclusive distribution of IS*1548* and GBS*i1*, an active group II intron identified in human isolates of group B streptococci. J.Bacteriol. **183**:2560-2569.

86. **Granlund, M., L. Oberg, M. Sellin, and M. Norgren**. 1998. Identification of a novel insertion element, IS*1548*, in group B streptococci, predominantly in strains causing endocarditis. J.Infect.Dis. **177**:967-976.

87. **Grattard, F., B. Pozzetto, B. B. de, H. Renaudin, M. Clerc, O. G. Gaudin, and C. Bebear**. 1995. Arbitrarily-primed PCR confirms the differentiation of strains of *Ureaplasma urealyticum* into two biovars. Mol.Cell.Probes **9**:383-389.

88. **Gravekamp, C., D. L. Kasper, J. L. Michel, D. E. Kling, V. Carey, and L. C. Madoff**. 1997. Immunogenicity and protective efficacy of the alpha C protein of group B streptococci are inversely related to the number of repeats. Infect.Immun. **65**:5216-5221.

89. **Gravekamp, C., B. Rosner, and L. C. Madoff**. 1998. Deletion of repeats in the alpha C protein enhances the pathogenicity of group B streptococci in immune mice. Infect.Immun. **66**:4347-4354.

90. **Gupta, S. and R. M. Anderson**. 1999. Population structure of pathogens: the role of immune selection. Parasitol.Today **15**:497-501.

91. **Gupta, R. S. and E. Griffiths**. 2002. Critical issues in bacterial phylogeny. Theor.Popul.Biol. **61**:423-434.

92. **Gurtler, V. and B. C. Mayall**. 2001. Genomic approaches to typing, taxonomy and evolution of bacterial isolates. Int.J.Syst.Evol.Microbiol. **51**:3-16.

93. **Hacker, J. and E. Carniel**. 2001. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep. **2**:376-381.

94. **Hakansson, S., L. G. Burman, J. Henrichsen, and S. E. Holm**. 1992. Novel coagglutination method for serotyping group B streptococci. J.Clin.Microbiol. **30**:3268-3269.

95. **Hamels, S., J. L. Gala, S. Dufour, P. Vannuffel, N. Zammatteo, and J. Remacle**. 2001. Consensus PCR and microarray for diagnosis of the genus *Staphylococcus*, species, and methicillin resistance. Biotechniques **31**:1364-2.

96. **Hance, A. J., K. Bradley, and R. G. Crystal**. 1976. Lung collagen heterogeneity. Synthesis of type I and type III collagen by rabbit and human lung cells in culture. J.Clin.Invest **57**:102-111.

97. **Hannaford, K., D. A. Todd, H. Jeffery, E. John, K. Blyth, and G. L. Gilbert**. 1999. Role of *Ureaplasma urealyticum* in lung disease of prematurity. Arch.Dis.Child Fetal Neonatal Ed **81**:F162-F167.

98. **Hansen, S. M., N. Uldbjerg, M. Kilian, and U. B. Sorensen**. 2004. Dynamics of *Streptococcus agalactiae* colonization in women during and after pregnancy and in their infants. J.Clin.Microbiol. **42**:83-89.

99. **Harasawa, R., K. Dybvig, H. I. Watson, and G. H. Cassell.** 1991. Two genomic clusters among 14 serovars of *Ureaplasma urealyticum*. Syst.Appl. Microbiol. **14**:393-396.

100. **Harasawa, R., Y. Imada, H. Kotani, K. Koshimizu, and M. F. Barile**. 1993. *Ureaplasma canigenitalium* sp. nov., isolated from dogs. Int.J.Syst.Bac. **43**:640-644.

101. **Harasawa, R. and Y. Kanamoto.** 1999. Differentiation of two biovars of *Ureaplasma urealyticum* based on the 16S-23S rRNA intergenic spacer region. J.Clin.Microbiol. 37:4135-4138.

102. **Harasawa, R.** 1999. Genetic relationships among mycoplasmas based on the 16S-23S rRNA spacer sequence. Microbiol.Immunol. **43**:127-132.

103. **Harris, T. O., D. W. Shelver, J. F. Bohnsack, and C. E. Rubens**. 2003. A novel streptococcal surface protease promotes virulence, resistance to opsonophagocytosis, and cleavage of human fibrinogen. J.Clin.Invest **111**:61-70.

104. **Harrison, H. R., W. T. Boyce, W. H. Haffner, B. Crowley, L. Weinstein, M. Lewis, and E. R. Alexander**. 1983. The prevalence of genital *Chlamydia trachomatis* and mycoplasmal infections during pregnancy in an American Indian population. Sex.Transm.Dis. **10**:184-186.

105. **Harrison, L. H., J. A. Elliott, D. M. Dwyer, J. P. Libonati, P. Ferrieri, L. Billmann, and A. Schuchat**. 1998. Serotype distribution of invasive group B streptococcal isolates in Maryland: implications for vaccine formulation. Maryland emerging infections program. J.Infect.Dis. **177**:998-1002.

106. **Hassan, A. A., A. Abdulmawjood, A. O. Yildirim, K. Fink, C. Lammler, and R. Schlenstedt**. 2000. Identification of streptococci isolated from various sources by determination of *cfb* gene and other CAMP-factor genes. Can.J.Microbiol. **46**:946-951.

107. **Hauge, M., C. Jespersgaard, K. Poulsen, and M. Kilian**. 1996. Population structure of *Streptococcus agalactiae* reveals an association between specific evolutionary lineages and putative virulence factors but not disease. Infect.Immun. **64**:919-925.

108. **Hedelin, H. H., P. A. Mardh, J. E. Brorson, M. Fall, B. R. Moller, and K. G. Pettersson**. 1983. *Mycoplasma hominis* and interstitial cystitis. Sex.Transm.Dis. **10**:327-330.

109. **Heden, L. O., E. Frithz, and G. Lindahl**. 1991. Molecular characterization of an IgA receptor from group B streptococci: sequence of the gene, identification of a proline-rich region with unique structure and isolation of N-terminal fragments with IgA-binding capacity. Eur.J.Immunol. **21**:1481-1490.

110. **Heggie, A. D., D. Bar-Shain, B. Boxerbaum, A. A. Fanaroff, M. A. O'Riordan, and J. A. Robertson**. 2001. Identification and quantification of ureaplasmas colonizing the respiratory tract and assessment of their role in the development of chronic lung disease in preterm infants. Pediatr.Infect.Dis.J. **20**:854-859.

111. **Heggie, A. D., M. R. Jacobs, V. T. Butler, J. E. Baley, and B. Boxerbaum**. 1994. Frequency and significance of isolation of *Ureaplasma urealyticum* and *Mycoplasma hominis* from cerebrospinal fluid and tracheal aspirate specimens from low birth weight infants. J.Pediatr. **124**:956-961.

112. **Hentschel, U. and J. Hacker**. 2001. Pathogenicity islands: the tip of the iceberg. Microbes.Infect. **3**:545-548.

113. **Hickman, M. E., M. A. Rench, P. Ferrieri, and C. J. Baker**. 1999. Changing epidemiology of group B streptococcal colonization. Pediatrics **104**:203-209.

114. **Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B. C. Li, and R. Herrmann**. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. **24**:4420-4449.

115. **Ho, M. Y., C. T. Wu, Y. T. Ku, F. Y. Huang, and C. C. Peng**. 1999. Group B streptococcal infection in neonates: an 11-year review. Acta Paediatr.Taiwan. **40**:83-86.

116. **Hofling, J. F., E. A. Rosa, M. J. Baptista, and D. M. Spolidorio**. 1997. New strategies on molecular biology applied to microbial systematics. Rev.Inst.Med.Trop.Sao Paulo **39**:345-352.

117. **Holm, S. E. and S. Hakansson**. 1988. A simple and sensitive enzyme immunoassay for determination of soluble type-specific polysaccharide from group B streptococci. J.Immunol.Methods **106**:89-94.

118. **Holmes, A. R., R. McNab, K. W. Millsap, M. Rohde, S. Hammerschmidt, J. L. Mawdsley, and H. F. Jenkinson**. 2001. The *pavA* gene of *Streptococcus pneumoniae* encodes a fibronectin-binding protein that is essential for virulence. Mol.Microbiol. **41**:1395-1408.

119. **Horner, P., B. Thomas, C. Gilroy, M. Egger, M. McClure, and D. Taylor-Robinson**. 2003. Antibodies to *Chlamydia trachomatis* heat-shock protein 60 kDa and detection of *Mycoplasma genitalium* and *Ureaplasma urealyticum* are associated independently with chronic nongonococcal urethritis. Sex.Transm.Dis. **30**:129-133.

120. **Horowitz, S., M. Mazor, J. Horowitz, A. Porath, and M. Glezerman**. 1995. Antibodies to *Ureaplasma urealyticum* in women with intraamniotic infection and adverse pregnancy outcome. Acta Obstet.Gynecol.Scand.**74**:132-136.

121. **Hoshina, K.** 1997. Minimization of the number of pregnant women to be treated with preventive procedure against GBS infection by means of antibody measurement. GBS infection group. Acta Paediatr.Jpn. **39**:546-549.

122. **Hoshina, K., Y. Suzuki, M. Kobayashi, S. Unno, M. Miyako, and K. Kaneko**. 1991. [Comparison between cultures of urine sediment and of the vagina for detection of carriers of group B Streptococcus (GBS) among pregnant women]. Kansenshogaku Zasshi **65**:992-995.

123. **Hoskins, J., W. E. Alborn, Jr., J. Arnold, L. C. Blaszczak, S. Burgett, B. S. DeHoff, S. T. Estrem, L. Fritz, D. J. Fu, W. Fuller, C. Geringer, R. Gilmour, J. S. Glass, H. Khoja, A. R. Kraft, R. E. Lagace, D. J. LeBlanc, L. N. Lee, E. J. Lefkowitz, J. Lu, P. Matsushima, S. M. McAhren, M. McHenney, K. McLeaster, C. W. Mundy, T. I. Nicas, F. H. Norris, M. O'Gara, R. B. Peery, G. T. Robertson, P. Rockey, P. M. Sun, M. E. Winkler, Y. Yang, M. Young-Bellido, G. Zhao, C. A. Zook, R. H. Baltz, S. R. Jaskunas, P. R. Rosteck, Jr., P. L. Skatrud, and J. I. Glass**. 2001. Genome of the bacterium *Streptococcus pneumoniae* strain R6. J.Bacteriol. **183**:5709-5717.

124. **Hughes, M. J., J. C. Moore, J. D. Lane, R. Wilson, P. K. Pribul, Z. N. Younes, R. J. Dobson, P. Everest, A. J. Reason, J. M. Redfern, F. M. Greer, T. Paxton, M. Panico, H.**

R. Morris, R. G. Feldman, and J. D. Santangelo. 2002. Identification of major outer surface proteins of *Streptococcus agalactiae*. Infect.Immun. **70**:1254-1259.

125. **Hunt, R. H., K. Sumanac, and J. Q. Huang**. 2001. Review article: should we kill or should we save *Helicobacter pylori*? Aliment.Pharmacol.Ther. **15 Suppl 1**:51-59.

126. **Hurst, L. D. and A. R. Merchant**. 2001. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc.R.Soc.Lond.B Biol.Sci. **268**:493-497.

127. **Hytonen, J., S. Haataja, and J. Finne**. 2003. *Streptococcus pyogenes* glycoprotein-binding strepadhesin activity is mediated by a surface-associated carbohydrate-degrading enzyme, pullulanase. Infect.Immun. **71**:784-793.

128. **Ilangovan, U., H. Ton-That, J. Iwahara, O. Schneewind, and R. T. Clubb**. 2001. Structure of sortase, the transpeptidase that anchors proteins to the cell wall of *Staphylococcus aureus*. Proc.Natl.Acad.Sci.U.S.A **98**:6056-6061.

129. **Israel, D. A., N. Salama, C. N. Arnold, S. F. Moss, T. Ando, H. P. Wirth, K. T. Tham, M. Camorlinga, M. J. Blaser, S. Falkow, and R. M. Peek, Jr.** 2001. *Helicobacter pylori* strain-specific differences in genetic content, identified by microarray, influence host inflammatory responses. J.Clin.Invest **107**:611-620.

130. **Itakura, A., O. Kurauchi, S. Morikawa, K. Matsuzawa, S. Mizutani, and Y. Tomoda**. 1996. A prospective study on the relationship between intrapartum maternal group-B streptococcal concentration and signs of infection in neonates. J.Obstet.Gynaecol.Res. **22**:101-105.

131. **Jacobs, E., A. Bartl, K. Oberle, and E. Schiltz**. 1995. Molecular mimicry by *Mycoplasma pneumoniae* to evade the induction of adherence inhibiting antibodies. J.Med.Microbiol. **43**:422-429.

132. **Jerlstrom, P. G., G. S. Chhatwal, and K. N. Timmis**. 1991. The IgA-binding beta antigen of the c protein complex of Group B streptococci: sequence determination of its gene and detection of two binding regions. Mol.Microbiol. **5**:843-849.

133. **Jiang, S. M., L. Wang, and P. R. Reeves**. 2001. Molecular characterization of *Streptococcus pneumoniae* type 4, 6B, 8, and 18C capsular polysaccharide gene clusters. Infect.Immun. **69**:1244-1255.

134. **Jones, N., J. F. Bohnsack, S. Takahashi, K. A. Oliver, M. S. Chan, F. Kunst, P. Glaser, C. Rusniok, D. W. Crook, R. M. Harding, N. Bisharat, and B. G. Spratt**. 2003. Multilocus sequence typing system for group B streptococcus. J.Clin.Microbiol. **41**:2530-2536.

135. **Jordan, I. K., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin**. 2002. Microevolutionary genomics of bacteria. Theor.Popul.Biol. **61**:435-447.

136. **Joyce, E. A., K. Chan, N. R. Salama, and S. Falkow**. 2002. Redefining bacterial populations: a post-genomic reformation. Nat.Rev.Genet. **3**:462-473.

137. **Jung, H. Y., S. Miyata, K. Oshima, S. Kakizawa, H. Nishigawa, W. Wei, S. Suzuki, M. Ugaki, T. Hibi, and S. Namba**. 2003. First complete nucleotide sequence and heterologous gene organization of the two rRNA operons in the phytoplasma genome. DNA Cell Biol. **22**:209-215.

138. **Jurgens, D., B. Sterzik, and F. J. Fehrenbach**. 1987. Unspecific binding of group B streptococcal cocytolysin (CAMP factor) to immunoglobulins and its possible role in pathogenicity. J.Exp.Med. **165**:720-732.

139. **Kalliola, S., J. Vuopio-Varkila, A. K. Takala, and J. Eskola**. 1999. Neonatal group B streptococcal disease in Finland: a ten-year nationwide study. Pediatr.Infect.Dis.J. **18**:806-810.

140. **Kamla, V., B. Henrich, and U. Hadding**. 1996. Phylogeny based on elongation factor Tu reflects the phenotypic features of mycoplasmas better than that based on 16S rRNA. Gene **171**:83-87.

141. **Katoh, M.** 2002. Paradigm shift in gene-finding method: From bench-top approach to desk-top approach (Review). Int.J.Mol.Med. **10**:677-682.

142. **Ke, D., C. Menard, F. J. Picard, M. Boissinot, M. Ouellette, P. H. Roy, and M. G. Bergeron**. 2000. Development of conventional and real-time PCR assays for the rapid detection of group B streptococci. Clin.Chem. **46**:324-331.

143. **Keefe, G. P.** 1997. *Streptococcus agalactiae* mastitis: a review. Can.Vet.J. **38**:429-437.

144. **Kenny, G. E. and F. D. Cartwright**. 1977. Effect of urea concentration on growth of *Ureaplasma urealyticum* (T-strain mycoplasma). J.Bacteriol. **132**:144-150.

145. **Kilian, M., J. Reinholdt, H. Lomholt, K. Poulsen, and E. V. Frandsen**. 1996. Biological significance of IgA1 proteases in bacterial colonization and pathogenesis: critical evaluation of experimental evidence. APMIS **104**:321-338.

146. **Kim, M., G. Kim, R. Romero, S. S. Shim, E. C. Kim, and B. H. Yoon**. 2003. Biovar diversity of *Ureaplasma urealyticum* in amniotic fluid: distribution, intrauterine inflammatory response and pregnancy outcomes. J.Perinat.Med. **31**:146-152.

147. **Knight, R. D., S. J. Freeland, and L. F. Landweber**. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. **2**:RESEARCH0010.

148. **Knox, C. L., D. G. Cave, D. J. Farrell, H. T. Eastment, and P. Timms**. 1997. The role of *Ureaplasma urealyticum* in adverse pregnancy outcome. Aust.N.Z.J.Obstet.Gynaecol. **37**:45-51.

149. **Knox, C. L., P. Giffard, and P. Timms.** 1998. The phylogeny of *Ureaplasma urealyticum* based on the MBA gene fragment. Int.J.Syst.Bacteriol. **48**:1323-1331.

150. **Knox, C. L., J. A. Allan, J. M. Allan, W. R. Edirisinghe, D. Stenzel, F. A. Lawrence, D. M. Purdie, and P. Timms**. 2003. *Ureaplasma parvum* and *Ureaplasma urealyticum* are detected in semen after washing before assisted reproductive technology procedures. Fertil.Steril. **80**:921-929.

151. **Kong, F., X. Zhu, W. Wang, X. Zhou, S. Gordon, and G. L. Gilbert**. 1999a. Comparative analysis and serovar-specific identification of multiple-banded antigen genes of *Ureaplasma urealyticum* biovar 1. J.Clin.Microbiol. **37**:538-543.

152. **Kong, F., G. James, Z. Ma, S. Gordon, W. Bin, and G. L. Gilbert**. 1999b. Phylogenetic analysis of *Ureaplasma urealyticum*--support for the establishment of a new species, *Ureaplasma parvum*. Int.J.Syst.Bacteriol. **49 Pt 4**:1879-1889.

153. **Kong, F., Z. Ma, G. James, S. Gordon, and G. L. Gilbert**. 2000a. Species identification and subtyping of *Ureaplasma parvum* and *Ureaplasma urealyticum* using PCR-based assays. J.Clin.Microbiol. **38**:1175-1179.

154. **Kong, F., Z. Ma, G. James, S. Gordon, and G. L. Gilbert**. 2000b. Molecular genotyping of human *Ureaplasma* species based on multiple-banded antigen (MBA) gene sequences. Int.J.Syst.Evol.Microbiol. **50 Pt 5**:1921-1929.

155. **Kong, F., S. Gordon, and G. L. Gilbert**. 2000c. Rapid-cycle PCR for detection and typing of *Mycoplasma pneumoniae* in clinical specimens. J.Clin.Microbiol. **38**:4256-4259.

156. **Kong, F., G. James, S. Gordon, A. Zelynski, and G. L. Gilbert**. 2001. Species-specific PCR for identification of common contaminant mollicutes in cell culture. Appl.Environ.Microbiol. **67**:3195-3200.

157. **Kong, F., S. Gowan, D. Martin, G. James, and G. L. Gilbert**. 2002a. Serotype identification of group B streptococci by PCR and sequencing. J.Clin.Microbiol. **40**:216-226.

158. **Kong, F., S. Gowan, D. Martin, G. James, and G. L. Gilbert**. 2002b. Molecular profiles of group B streptococcal surface protein antigen genes: relationship to molecular serotypes. J.Clin.Microbiol. **40**:620-626.

159. **Kong, F., D. Martin, G. James, and G. L. Gilbert**. 2003. Towards a genotyping system for *Streptococcus agalactiae* (group B streptococcus): use of mobile genetic elements in Australasian invasive isolates. J.Med.Microbiol. **52**:337-344.

160. **Kong, F. and G. L. Gilbert**. 2003. Using *cpsA-cpsB* sequence polymorphisms and serotype-/group-specific PCR to predict 51 *Streptococcus pneumoniae* capsular serotypes. J.Med.Microbiol. **52**:1047-1058.

161. **Koskiniemi, S., M. Sellin, and M. Norgren**. 1998. Identification of two genes, *cpsX* and *cpsY*, with putative regulatory function on capsule expression in group B streptococci. FEMS Immunol.Med.Microbiol. **21**:159-168.

162. **Kreikemeyer, B., K. S. McIver, and A. Podbielski**. 2003. Virulence factor regulation and regulatory networks in *Streptococcus pyogenes* and their impact on pathogen-host interactions. Trends Microbiol. **11**:224-232.

163. **Kruger, T. and J. Baier**. 1997. Induction of neutrophil chemoattractant cytokines by *Mycoplasma hominis* in alveolar type II cells. Infect.Immun. **65**:5131-5136.

164. **Kundsin, R. B., A. Leviton, E. N. Allred, and S. A. Poulin**. 1996. *Ureaplasma urealyticum* infection of the placenta in pregnancies that ended prematurely. Obstet.Gynaecol. **87**:122-127.

165. **Kvam, A. I., A. Efstratiou, L. Bevanger, B. D. Cookson, I. F. Marticorena, George, RC, and J. A. Maeland**. 1995. Distribution of serovariants of group B streptococci in isolates from England and Norway. J.Med.Microbiol. **42**:246-250.

166. **Kwon, H. Y., S. W. Kim, M. H. Choi, A. D. Ogunniyi, J. C. Paton, S. H. Park, S. N. Pyo, and D. K. Rhee**. 2003. Effect of heat shock and mutations in ClpL and ClpP on virulence gene expression in *Streptococcus pneumoniae*. Infect.Immun. **71**:3757-3765.

167. **Lachenauer, C. S., R. Creti, J. L. Michel, and L. C. Madoff**. 2000. Mosaicism in the alpha-like protein genes of group B streptococci. Proc.Natl.Acad.Sci.U.S.A **97**:9630-9635.

168. **Lachenauer, C. S., D. L. Kasper, J. Shimada, Y. Ichiman, H. Ohtsuka, M. Kaku, L. C. Paoletti, P. Ferrieri, and L. C. Madoff**. 1999. Serotypes VI and VIII predominate among group B streptococci isolated from pregnant Japanese women. J.Infect.Dis. **179**:1030-1033.

169. **Lang, S. and M. Palmer**. 2003. Characterization of *Streptococcus agalactiae* CAMP factor as a pore-forming toxin. J.Biol.Chem. **278:** 38167-38173.

170. **Larsen, B. and G. R. Monif**. 2001. Understanding the bacterial flora of the female genital tract. Clin.Infect.Dis. **32**:e69-e77.

171. **Larsson, C., M. Stalhammar-Carlemalm, and G. Lindahl**. 1996. Experimental vaccination against group B streptococcus, an encapsulated bacterium, with highly purified preparations of cell surface proteins Rib and alpha. Infect.Immun. **64**:3518-3523.

172. **Lawrence, J. G.** 2002. Gene transfer in bacteria: speciation without species? Theor.Popul.Biol. **61**:449-460.

173. **Levine, E. M., V. Ghai, J. J. Barton, and C. M. Strom**. 1999. Intrapartum antibiotic prophylaxis increases the incidence of gram-negative neonatal sepsis. Infect.Dis.Obstet.Gynecol. **7**:210-213.

174. **Limansky, A. S., M. C. Guardati, E. G. Sutich, I. E. Toresani, I. E. Bogado, G. M. Joannas, and A. M. Viale**. 1995. Characterization of clinical isolates of *Streptococcus agalactiae* by random amplified polymorphic DNA using degenerate oligonucleotides. Medicina **55**:681-684.

175. **Lin, F. Y., J. D. Clemens, P. H. Azimi, J. A. Regan, L. E. Weisman, J. B. Philips, III, G. G. Rhoads, P. Clark, R. A. Brenner, and P. Ferrieri**. 1998. Capsular polysaccharide types of group B streptococcal isolates from neonates with early-onset systemic infection. J.Infect.Dis. **177**:790-792.

176. **Lopardo, H. A., P. Vidal, P. Jeric, D. Centron, H. Paganini, R. R. Facklam, and J. Elliott**. 2003. Six-month multicenter study on invasive infections due to group B streptococci in Argentina. J.Clin.Microbiol. **41**:4688-4694.

177. **Ludwig, W., O. Strunk, S. Klugbauer, N. Klugbauer, M. Weizenegger, J. Neumaier, M. Bachleitner, and K. H. Schleifer**. 1998. Bacterial phylogeny based on comparative sequence analysis. Electrophoresis **19**:554-568.

178. **Luscombe, N. M., D. Greenbaum, and M. Gerstein**. 2001. What is bioinformatics? A proposed definition and overview of the field. Methods Inf.Med. **40**:346-358.

179. **Mabanta, C., G. Pryhuber, G. Weinberg, and D. Phelps**. 2003. Erythromycin for the prevention of chronic lung disease in intubated preterm infants at risk for, or colonized or infected with *Ureaplasma urealyticum*. Cochrane.Database.Syst.Rev. **4**:CD003744.

180. **Madoff, L. C., J. L. Michel, E. W. Gong, D. E. Kling, and D. L. Kasper**. 1996. Group B streptococci escape host immunity by deletion of tandem repeat elements of the alpha C protein. Proc.Natl.Acad.Sci.U.S.A **93**:4131-4136.

181. **Maeland, J. A., L. Bevanger, G. Iversen, and R. V. Lyng**. 1999. *bca*, beta gene, and gene product divergency in reference and prototype strains of *Streptococcus agalactiae*. Clin.Diagn.Lab Immunol. **6**:986-988.

182. **Maeland, J. A., O. G. Brakstad, L. Bevanger, and S. Krokstad**. 2000. Distribution and expression of *bca*, the gene encoding the c alpha protein, by *Streptococcus agalactiae*. J.Med.Microbiol. **49**:193-198.

183. **Magalhaes, M.** 1983. Persistent nongonococcal urethritis associated with a minocycline-resistant strain of *Ureaplasma urealyticum*: a case report. Sex.Transm.Dis. **10**:151-152.

184. **Magee, J. T., J. D. Fox, and S. L. Stubbs**. 2001. Cashing in your chips: speculation on the future of diagnostic laboratories in the era of DNA chips. J.Med.Microbiol. **50**:111-115.

185. **Mahillon, J. and M. Chandler**. 1998. Insertion sequences. Microbiol.Mol.Biol.Rev. **62**: 725-774.

186. **Mahillon, J., C. Leonard, and M. Chandler**. 1999. IS elements as constituents of bacterial genomes. Res.Microbiol. **150**:675-687.

187. **Manning, S. D.** 2003. Molecular epidemiology of *Streptococcus agalactiae* (group B streptococcus). Front Biosci. **8**:S1-18.

188. **Martin, B., O. Humbert, M. Camara, E. Guenzi, J. Walker, T. Mitchell, P. Andrew, M. Prudhomme, G. Alloing, R. Hakenbeck, and .** 1992. A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. Nucleic Acids Res. **20**:3479-3483.

189. **Martinez-Abarca, F. and N. Toro**. 2000. Group II introns in the bacterial world. Mol.Microbiol. **38**:917-926.

190. **Martinez, G., J. Harel, R. Higgins, S. Lacouture, D. Daignault, and M. Gottschalk**. 2000. Characterization of *Streptococcus agalactiae* isolates of bovine and human origin by randomly amplified polymorphic DNA analysis. J.Clin.Microbiol. **38**:71-78.

191. **Martinez, M. A., A. Ovalle, A. Santa-Cruz, B. Barrera, R. Vidal, and R. Aguirre**. 2001. Occurrence and antimicrobial susceptibility of *Ureaplasma parvum* (*Ureaplasma urealyticum* biovar 1) and *Ureaplasma urealyticum* (*Ureaplasma urealyticum* biovar 2) from patients with adverse pregnancy outcomes and normal pregnant women. Scand.J.Infect.Dis. **33**:604-610.

192. **Mawn, J. A., A. J. Simpson, and S. R. Heard**. 1993. Detection of the C protein gene among group B streptococci using PCR. J. Clin. Pathology **46**:633-636.

193. **McDonald, B. A. and C. Linde**. 2002. Pathogen population genetics, evolutionary potential, and durable resistance. Annu.Rev.Phytopathol. **40**:349-379.

194. **McDonald, H. M. and H. M. Chambers**. 2000. Intrauterine infection and spontaneous midgestation abortion: is the spectrum of microorganisms similar to that in preterm labor? Infect.Dis.Obstet.Gynecol. **8**:220-227.

195. **McKenzie, H., M. L. Donnet, P. W. Howie, N. B. Patel, and D. T. Benvie**. 1994. Risk of preterm delivery in pregnant women with group B streptococcal urinary infections or urinary antibodies to group B streptococcal and *E. coli* antigens. Br.J Obstet.Gynaecol. **101**:107-113.

196. **Mei, J. M., F. Nourbakhsh, C. W. Ford, and D. W. Holden**. 1997. Identification of *Staphylococcus aureus* virulence genes in a murine model of bacteraemia using signature-tagged mutagenesis. Mol.Microbiol. **26**:399-407.

197. **Michel, J. L., L. C. Madoff, K. Olson, D. E. Kling, D. L. Kasper, and F. M. Ausubel**. 1992. Large, identical, tandem repeating units in the C protein alpha antigen gene, *bca*, of group B streptococci. Proc.Natl.Acad.Sci. U.S.A. **89**:10060-10064.

198. **Mikamo, H., M. Higa, Y. Sato, Y. Hayasaki, Y. X. Hua, K. Yasuda-Kawazoe, and T. Tamaya**. 2000. Preventive procedures against GBS infection by means of antibody measurement. Rinsho Biseibutshu.Jinsoku.Shindan.Kenkyukai.Shi **11**:33-37.

199. **Mira, A., L. Klasson, and S. G. Andersson**. 2002. Microbial genome evolution: sources of variability. Curr.Opin.Microbiol. **5**:506-512.

200. **Monecke, S., J. H. Helbig, and E. Jacobs**. 2003. Phase variation of the multiple banded protein in *Ureaplasma urealyticum* and *Ureaplasma parvum*. Int.J.Med.Microbiol. **293**:203-211.

201. **Moran, N. A.** 2002. Microbial minimalism: genome reduction in bacterial pathogens. Cell **108**:583-586.

202. **Moreno, E.** 1997. In search of a bacterial species definition. Rev.Biol.Trop. **45**:753-771.

203. **Moore, M. R., S. J. Schrag, and A. Schuchat**. 2003. Effects of intrapartum antimicrobial prophylaxis for prevention of group-B-streptococcal disease on the incidence and ecology of early-onset neonatal sepsis. Lancet.Infect.Dis. **3**:201-213.

204. **Morschhauser, J., G. Kohler, W. Ziebuhr, G. Blum-Oehler, U. Dobrindt, and J. Hacker**. 2000. Evolution of microbial pathogens. Philos.Trans.R.Soc.Lond B Biol.Sci. **355**:695-704.

205. **Mougel, C., J. Thioulouse, G. Perriere, and X. Nesme**. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. Int.J.Syst.Evol.Microbiol. **52**:573-586.

206. **Mounsey, A., P. Bauer, and I. A. Hope**. 2002. Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. Genome Res. **12**:770-775.

207. **Moyo, S. R., J. A. Maeland, and R. V. Lyng**. 2001a. The putative R1 protein of *Streptococcus agalactiae* as serotype marker and target of protective antibodies. APMIS **109**:842-848.

208. **Moyo, S. R., J. A. Maeland, and J. Mudzori**. 2001b. Antibodies against *Streptococcus agalactia*e proteins c(alpha) and R4 in sera from pregnant women from Norway and Zimbabwe. Clin.Diagn.Lab Immunol. **8**:1110-1114.

209. **Moyo, S. R. and J. A. Maeland** . 2003. Antibodies raised in animals against the *Streptococcus agalactiae* proteins c alpha and R4 and normal human serum antibodies target distinct epitopes. J.Med.Microbiol. **52**:379-383.

210. **Munday, P. E., P. M. Furr, and D. Taylor-Robinson**. 1981. The prevalence of *Ureaplasma urealyticum* and *Mycoplasma hominis* in the cervix and anal canal of women. J. Infect. **3**:253-257.

211. **Murray, R. G. and K. H. Schleifer**. 1994. Taxonomic notes: a proposal for recording the properties of putative taxa of procaryotes. Int.J.Syst.Bacteriol. **44**:174-176.

212. **Musser, J. M.** 1996. Molecular population genetic analysis of emerged bacterial pathogens: selected insights. Emerg.Infect.Dis. **2**:1-17.

213. **Musser, J. M., S. J. Mattingly, R. Quentin, A. Goudeau, and R. K. Selander**. 1989. Identification of a high-virulence clone of type III *Streptococcus agalactiae* (group B streptococcus) causing invasive neonatal disease. Proc.Natl.Acad.Sci.U.S.A. **86**:4731-4735.

214. **Mygind, T., S. Birkelund, and G. Christiansen**. 1998. DNA sequencing reveals limited heterogeneity in the 16S rRNA gene from the *rrnB* operon among five *Mycoplasma hominis* isolates. Int.J.Syst.Bacteriol. **48**:1067-1071.

215. **Naessens, A., X. Cheng, S. Lauwers, and J. A. Robertson**. 1998. Development of a monoclonal antibody to a *Ureaplasma urealyticum* serotype 9 antigen. J.Clin.Microbiol. **36**:1125-1127.

216. **Naessens, A., W. Foulon, J. Breynaert, and S. Lauwers**. 1988. Serotypes of *Ureaplasma urealyticum* isolated from normal pregnant women and patients with pregnancy complications. J.Clin.Microbiol. **26**:319-322.

217. **Nagano, Y., N. Nagano, S. Takahashi, K. Murono, K. Fujita, F. Taguchi, and Y. Okuwaki**. 1991. Restriction endonuclease digest patterns of chromosomal DNA from group B beta-haemolytic streptococci. J.Med.Microbiol. **35**:297-303.

218. **Nair, S., C. Frehel, L. Nguyen, V. Escuyer, and P. Berche**. 1999. ClpE, a novel member of the HSP100 family, is involved in cell division and virulence of *Listeria monocytogenes*. Mol.Microbiol. **31**:185-196.

219. **Nair, S., E. Milohanic, and P. Berche**. 2000. ClpC ATPase is required for cell adhesion and invasion of *Listeria monocytogenes*. Infect.Immun. **68**:7061-7068.

220. **Nakagawa, I., K. Kurokawa, A. Yamashita, M. Nakata, Y. Tomiyasu, N. Okahashi, S. Kawabata, K. Yamazaki, T. Shiba, T. Yasunaga, H. Hayashi, M. Hattori, and S. Hamada**. 2003. Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. Genome Res. **13**:1042-1055.

221. **Neyrolles, O., S. Ferris, N. Behbahani, L. Montagnier, and A. Blanchard**. 1996. Organization of *Ureaplasma urealyticum* urease gene cluster and expression in a suppressor strain of *Escherichia coli*. J.Bacteriol. **178**:2725.

222. **Obaro, S. and R. Adegbola**. 2002. The pneumococcus: carriage, disease and conjugate vaccines. J.Med.Microbiol. **51**:98-104.

223. **Ochman, H. and N. A. Moran**. 2001. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. Science **292**:1096-1099.

224. **Oggioni, M. R. and J. P. Claverys**. 1999. Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae*. Microbiology **145**:2647-2653.

225. **Ollikainen, J., T. Heiskanen-Kosma, M. Korppi, M. L. Katila, and K. Heinonen**. 1998. Clinical relevance of *Ureaplasma urealyticum* colonization in preterm infants. Acta Paediatr. **87**:1075-1078.

226. **Osborn, A. M. and D. Boltner**. 2002. When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. Plasmid **48**:202-212.

227. **Palacios, G. C., E. K. Eskew, F. Solorzano, and S. J. Mattingly**. 1997. Decreased capacity for type-specific-antigen synthesis accounts for high prevalence of nontypeable strains of group B streptococci in Mexico. J.Clin.Microbiol. **35**:2923-2926.

228. **Pallen, M. J., A. C. Lam, M. Antonio, and K. Dunbar**. 2001. An embarrassment of sortases - a richness of substrates? Trends Microbiol. **9**:97-102.

229. **Paoletti, L. C. and L. C. Madoff**. 2002. Vaccines to prevent neonatal GBS infection. Semin.Neonatol. **7**:315-323.

230. **Paoletti, L. C., J. Pinel, K. D. Johnson, B. Reinap, R. A. Ross, and D. L. Kasper**. 1999. Synthesis and preclinical evaluation of glycoconjugate vaccines against group B streptococcus types VI and VIII. J.Infect.Dis. **180**:892-895.

231. **Papapetropoulou, M. and A. Pappas**. 1987. The acute urethral syndrome in routine practice. J.Infect. **14**:113-118.

232. **Petit, C. M., J. R. Brown, K. Ingraham, A. P. Bryant, and D. J. Holmes**. 2001. Lipid modification of prelipoproteins is dispensable for growth in vitro but essential for virulence in *Streptococcus pneumoniae*. FEMS Microbiol.Lett. **200**:229-233.

233. **Pettersson, B., T. Leitner, M. Ronaghi, G. Bolske, M. Uhlen, and K. E. Johansson**. 1996. Phylogeny of the *Mycoplasma mycoides* cluster as determined by sequence analysis of the 16S rRNA genes from the two rRNA operons. J.Bacteriol. **178**:4131-4142.

234. **Pettersson, B., J. G. Tully, G. Bolske, and K. E. Johansson**. 2000. Updated phylogenetic description of the *Mycoplasma hominis* cluster (Weisburg *et al*. 1989) based on 16S rDNA sequences. Int.J.Syst.Evol.Microbiol. **50**:291-301.

235. **Podbielski, A., O. Blankenstein, and R. Lutticken**. 1994. Molecular characterization of the *cfb* gene encoding group B streptococcal CAMP-factor. Med.Microbiol.Immunol.(Berl) **183**:239-256.

236. **Pollack, J. D.** 2001. *Ureaplasma urealyticum*: an opportunity for combinatorial genomics. Trends Microbiol. **9**:169-175.

237. **Pollack, J. D., M. A. Myers, T. Dandekar, and R. Herrmann**. 2002. Suspected utility of enzymes with multiple activities in the small genome *Mycoplasma* species: the replacement of the missing "household" nucleoside diphosphate kinase gene and activity by glycolytic kinases. OMICS. **6**:247-258.

238. **Povlsen, K., E. Bjornelius, P. Lidbrink, and I. Lind**. 2002. Relationship of *Ureaplasma urealyticum* biovar 2 to nongonococcal urethritis. Eur.J.Clin.Microbiol.Infect.Dis. **21**:97-101.

239. **Povlsen, K., J. S. Jensen, and I. Lind**. 1998. Detection of *Ureaplasma urealyticum* by PCR and biovar determination by liquid hybridization. J.Clin.Microbiol. **36**:3211-3216.

240. **Pritchard, D. G., B. Lin, T. R. Willingham, and J. R. Baker**. 1994. Characterization of the group B streptococcal hyaluronate lyase. Arch Biochem Biophys. **315**:431-437.

241. **Pritzlaff, C. A., J. C. Chang, S. P. Kuo, G. S. Tamura, C. E. Rubens, and V. Nizet**. 2001. Genetic basis for the beta-haemolytic/cytolytic activity of group B Streptococcus. Mol.Microbiol. **39**:236-247.

242. **Quentin, R., H. Huet, F. S. Wang, P. Geslin, A. Goudeau, and R. K. Selander**. 1995. Characterization of *Streptococcus agalactiae* strains by multilocus enzyme genotype and serotype: identification of multiple virulent clone families that cause invasive neonatal disease. J.Clin.Microbiol. **33**:2576-2581.

243. **Razin, S., D. Yogev, and Y. Naot**. 1998. Molecular biology and pathogenicity of mycoplasmas. Microbiol.Mol.Biol.Rev. **62**:1094-1156.

244. **Regan, J. A., M. A. Klebanoff, R. P. Nugent, D. A. Eschenbach, W. C. Blackwelder, Y. Lou, R. S. Gibbs, P. J. Rettig, D. H. Martin, and R. Edelman**. 1996. Colonization with group B streptococci in pregnancy and adverse outcome. VIP Study Group. Am.J.Obstet.Gynecol. **174**:1354-1360.

245. **Ren, Y. and X. Zhu**. 2003. Investigation on biovars and genotypes of *Ureaplasma urealyticum* in the cervix in a Chinese gynecologic check-up population and sex workers. Acta Derm.Venereol. **83**:175-178.

246. **Reid, I., M. A. Martinez, M. Biron, and R. Aceituno**. 1993. [Mycoplasma hominis and Ureaplasma urealyticum in pregnant and infertile women. Differences in tubal pathology]. Rev.Chil.Obstet.Ginecol. **58**:349-354.

247. **Robertson, J. and E. Smook**. 1976. Cytochemical evidence of extramembranous carbohydrates on *Ureaplasma urealyticum* (T-strain *Mycoplasma*). J.Bacteriol. **128**:658-660.

248. **Robertson, J. A. and M. H. Chen**. 1984. Effects of manganese on the growth and morphology of *Ureaplasma urealyticum*. J.Clin.Microbiol. **19**:857-864.

249. **Robertson, J. A., L. H. Honore, and G. W. Stemke**. 1986. Serotypes of *Ureaplasma urealyticum* in spontaneous abortion. Pediatr.Infect.Dis. **5**:S270-S272.

250. **Robertson, J. A., L. E. Pyle, G. W. Stemke, and L. R. Finch**. 1990. Human ureaplasmas show diverse genome sizes by pulsed-field electrophoresis. Nucleic Acids Res. **18**:1451-1455.

251. **Robertson, J. A. and G. W. Stemke**. 1982. Expanded serotyping scheme for *Ureaplasma urealyticum* strains isolated from humans. J.Clin.Microbiol. **15**:873-878.

252. **Robertson, J. A., G. W. Stemke, J. W. Davis, Jr., R. Harasawa, D. Thirkell, F. Kong, M. C. Shepard, and D. K. Ford**. 2002. Proposal of *Ureaplasma parvum* sp. nov. and emended description of *Ureaplasma urealyticum* (Shepard *et al*. 1974) Robertson et al. 2001. Int.J.Syst.Evol.Microbiol. **52**:587-597.

253. **Robertson, J. A., M. E. Stemler, and G. W. Stemke**. 1984. Immunoglobulin A protease activity of *Ureaplasma urealyticum*. J.Clin.Microbiol. **19**:255-258.

254. **Robertson, J. A., A. Vekris, C. Bebear, and G. W. Stemke**. 1993. Polymerase chain reaction using 16S rRNA gene sequences distinguishes the two biovars of Ureaplasma urealyticum. J.Clin.Microbiol. **31**:824-830.

255. **Rocha, E. P. and A. Blanchard**. 2002. Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution. Nucleic Acids Res. **30**:2031-2042.

256. **Rolland, K., C. Marois, V. Siquier, B. Cattier, and R. Quentin**. 1999. Genetic features of *Streptococcus agalactiae* strains causing severe neonatal infections, as revealed by pulsed-field gel electrophoresis and *hylB* gene analysis. J.Clin.Microbiol. **37**:1892-1898.

257. **Rolland, K., L. Mereghetti, S. Watt, S. Chatellier, and R. Quentin**. 2002. tRNA gene clusters at the 3' end of rRNA operons are specific to virulent subgroups of *Streptococcus agalactiae* strains, as demonstrated by molecular differential analysis at the population level. Microbiology **148**:1493-1499.

258. **Rubens, C. E., L. M. Heggen, and J. M. Kuypers**. 1989. IS*861*, a group B streptococcal insertion sequence related to IS*150* and IS*3* of *Escherichia coli.* J.Bacteriol. **171**:5531-5535.

259. **Russell, F. M. and J. Buttery**. 2003. Vaccine development for capsulate bacteria causing pneumonia. Curr.Opin.Pulm.Med. **9**:227-232.

260. **Saier, M. H., Jr.** 1996. Phylogenetic approaches to the identification and characterization of protein families and superfamilies. Microb.Comp Genomics **1**:129-150.

261. **Sanchez-Beato, A. R., E. Garcia, R. Lopez, and J. L. Garcia**. 1997. Identification and characterization of IS*1381*, a new insertion sequence in *Streptococcus pneumoniae*. J.Bacteriol. **179**:2459-2463.

262. **Sanchez, P. J. and J. A. Regan**. 1987. Vertical transmission of *Ureaplasma urealyticum* in full term infants. Pediatr. Infect.Dis.J. **6**:825-828.

263. **Sandberg, R., C. I. Branden, I. Ernberg, and J. Coster**. 2003. Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. Gene **311**:35-42.

264. **Sassetti, C. and E. J. Rubin**. 2002. Genomic analyses of microbial virulence. Curr.Opin.Microbiol. **5**:27-32.

265. **Schoolnik, G. K.** 2002. Functional and comparative genomics of pathogenic bacteria. Curr.Opin.Microbiol. **5**:20-26.

266. **Schrag, S., R. Gorwitz, K. Fultz-Butts, and A. Schuchat**. 2002. Prevention of perinatal group B streptococcal disease. Revised guidelines from CDC. MMWR Recomm.Rep. **51**:1-22.

267. **Schrag, S. J., S. Zywicki, M. M. Farley, A. L. Reingold, L. H. Harrison, L. B. Lefkowitz, J. L. Hadler, R. Danila, P. R. Cieslak, and A. Schuchat**. 2000. Group B streptococcal disease in the era of intrapartum antibiotic prophylaxis. N.Engl.J.Med. **342**:15-20.

268. **Schubert, A., K. Zakikhany, M. Schreiner, R. Frank, B. Spellerberg, B. J. Eikmanns, and D. J. Reinscheid**. 2002. A fibrinogen receptor from group B streptococcus interacts

with fibrinogen by repetitive units with novel ligand binding sites. Mol.Microbiol. **46**:557-569.

269. **Schuchat, A.** 1998. Epidemiology of group B streptococcal disease in the United States: shifting paradigms. Clin.Microbiol.Rev. **11**:497-513.

270. **Schuchat, A.** 1999. Group B streptococcus. Lancet **353**:51-56.

271. **Sellin, M., S. Hakansson, and M. Norgren**. 1995. Phase-shift of polysaccharide capsule expression in group B streptococci, type III. Microb.Pathog. **18**:401-415.

272. **Sellin, M., C. Olofsson, S. Hakansson, and M. Norgren**. 2000. Genotyping of the capsule gene cluster (*cps*) in nontypeable group B streptococci reveals two major *cps* allelic variants of serotypes III and VII. J.Clin.Microbiol. **38**:3420-3428.

273. **Shakhnovich, E. A., S. J. King, and J. N. Weiser**. 2002. Neuraminidase expressed by *Streptococcus pneumoniae* desialylates the lipopolysaccharide of *Neisseria meningitidis* and *Haemophilus influenzae*: a paradigm for interbacterial competition among pathogens of the human respiratory tract. Infect.Immun. **70**:7161-7164.

274. **Sharma, S., R. Brousseau, and S. Kasatiya**. 1998. Detection and confirmation of *Mycoplasma pneumoniae* in urogenital specimens by PCR. J.Clin.Microbiol. **36**:277-280.

275. **Simmons, M. P., H. Ochoterena, and J. V. Freudenstein**. 2002. Amino acid vs. nucleotide characters: challenging preconceived notions. Mol.Phylogenet.Evol. **24**:78-90.

276. **Smith, D. G., W. C. Russell, and D. Thirkell**. 1994. Adherence of *Ureaplasma urealyticum* to human epithelial cells. Microbiology **140**:2893-2898.

277. **Spellerberg, B.** 2000. Pathogenesis of neonatal *Streptococcus agalactiae* infections. Microbes.Infect. **2**:1733-1742.

278. **Spellerberg, B., S. Martin, C. Franken, R. Berner, and R. Lutticken**. 2000. Identification of a novel insertion sequence element in *Streptococcus agalactiae*. Gene **241**:51-56.

279. **Spellerberg, B., E. Rozdzinski, S. Martin, J. Weber-Heynemann, N. Schnitzler, R. Lutticken, and A. Podbielski** . 1999. Lmb, a protein with similarities to the LraI adhesin family, mediates attachment of *Streptococcus agalactiae* to human laminin. Infect.Immun. **67**:871-878.

280. **Spratt, B. G. and M. C. Maiden**. 1999. Bacterial population genetics, evolution and epidemiology. Philos.Trans.R.Soc.Lond B Biol.Sci. **354**:701-710.

281. **Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. Grimont, P. Kampfer, M. C. Maiden, X. Nesme, R. Rossello-Mora, J. Swings, H. G. Truper, L.Vauterin, A. C. Ward, and W. B. Whitman.** 2002. Report of the *ad hoc* committee for the re-evaluation of the species definition in bacteriology. Int.J.Syst.Evol.Microbiol. **52**:1043-1047.

282. **Stalhammar-Carlemalm, M., T. Areschoug, C. Larsson, and G. Lindahl**. 1999. The R28 protein of *Streptococcus pyogenes* is related to several group B streptococcal surface proteins, confers protective immunity and promotes binding to human epithelial cells. Mol.Microbiol. **33**:208-219.

283. **Stalhammar-Carlemalm, M., L. Stenberg, and G. Lindahl**. 1993. Protein Rib: a novel group B streptococcal cell surface protein that confers protective immunity and is expressed by most strains causing invasive infections. J.Exp.Med. **177**:1593-1603.

284. **Stemke, G. W. and J. A. Robertson**. 1985. Problems associated with serotyping strains of *Ureaplasma urealyticum*. Diagn.Microbiol.Infect.Dis. **3**:311-320.

285. **Stingele, F., J. R. Neeser, and B. Mollet**. 1996. Identification and characterization of the *eps* (Exopolysaccharide) gene cluster from *Streptococcus thermophilus* Sfi6. J.Bacteriol. **178**:1680-1690.

286. **Strauss, E. J. and S. Falkow**. 1997. Microbial pathogenesis: genomics and beyond. Science **276**:707-712.

287. **Subramanian, G., R. Mural, S. L. Hoffman, J. C. Venter, and S. Broder**. 2001. Microbial disease in humans: A genomic perspective. Mol.Diagn. **6**:243-252.

288. **Suvorov, A., A. Dmitriev, I. Ustinovitch, C. Schalen, and A. A. Totolian**. 1997. Molecular analysis of clinical group B streptococcal strains by use of alpha and beta gene probes. FEMS Immunol.Med.Microbiol. **17**:149-154.

289. **Takahashi, S., E. E. Adderson, Y. Nagano, N. Nagano, M. R. Briesacher, and J. F. Bohnsack**. 1998. Identification of a highly encapsulated, genetically related group of invasive type III group B streptococci. J.Infect.Dis. **177**:1116-1119.

290. **Takahashi, Y., K. Konishi, J. O. Cisar, and M. Yoshikawa**. 2002. Identification and characterization of *hsa*, the gene encoding the sialic acid-binding adhesin of *Streptococcus gordonii* DL1. Infect.Immun. **70**:1209-1218.

291. **Tamura, G. S., M. Herndon, J. Przekwas, C. E. Rubens, P. Ferrieri, and S. L. Hillier**. 2000. Analysis of restriction fragment length polymorphisms of the insertion sequence IS*1381* in group B Streptococci. J.Infect.Dis. **181**:364-368.

292. **Tamura, H., T. Kikuchi, R. Shirato, and H. Kato**. 2001. Cloning and DNA sequencing of the surface protein antigen I/II (PAa) of *Streptococcus cricetus*. FEMS Microbiol.Lett. **196**:251-256.

293. **Taraskina, A. E., A. M. Savicheva, T. A. Akopian, A. E. Soroka, K. T. Momynaliev, and V. M. Govorun**. 2002. Drift of *tetM* determinant in urogenital microbiocenosis containing mycoplasmas during treatment with a tetracycline antibiotic. Bull.Exp.Biol.Med. **134**:60-63.

294. **Tatusov, R. L., E. V. Koonin, and D. J. Lipman**. 1997. A genomic perspective on protein families. Science **278**:631-637.

295. **Taylor-Robinson, D.** 1986. The male reservoir of *Ureaplasma urealyticum*. Pediatric Infect. Dis. **5**:S234-S235.

296. **Taylor-Robinson, D. and P. M. Furr**. 1997. Genital mycoplasma infections. Wien.Klin.Wochenschr. **109**:578-583.

297. **Taylor-Robinson, D., P. M. Furr, and A. D. Webster**. 1985. *Ureaplasma urealyticum* causing persistent urethritis in a patient with hypogammaglobulinaemia. Genitourin.Med. **61**:404-408.

298. **Teng, L. J., S. W. Ho, H. N. Ho, S. J. Liaw, H. C. Lai, and K. T. Luh**. 1995. Rapid detection and biovar differentiation of *Ureaplasma urealyticum* in clinical specimens by PCR. J.Form.Med.Assoc. **94**:396-400.

299. **Teng, L. J., X. Zheng, J. I. Glass, H. L. Watson, J. Tsai, and G. H. Cassell**. 1994. *Ureaplasma urealyticum* biovar specificity and diversity are encoded in multiple-banded antigen gene. J.Clin.Microbiol. **32**:1464-1469.

300. **Tettelin, H., V. Masignani, M. J. Cieslewicz, J. A. Eisen, S. Peterson, M. R. Wessels, I. T. Paulsen, K. E. Nelson, I. Margarit, T. D. Read, L. C. Madoff, A. M. Wolf, M. J. Beanan, L. M. Brinkac, S. C. Daugherty, R. T. DeBoy, A. S. Durkin, J. F. Kolonay, R. Madupu, M. R. Lewis, D. Radune, N. B. Fedorova, D. Scanlan, H. Khouri, S. Mulligan, H. A. Carty, R. T. Cline, S. E. Van Aken, J. Gill, M. Scarselli, M. Mora, E. T. Iacobini, C. Brettoni, G. Galli, M. Mariani, F. Vegni, D. Maione, D. Rinaudo, R. Rappuoli, J. L. Telford, D. L. Kasper, G. Grandi, and C. M. Fraser**. 2002. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc.Natl.Acad.Sci.U.S.A **99:** 12391-12396.

301. **Tettelin, H., K. E. Nelson, I. T. Paulsen, J. A. Eisen, T. D. Read, S. Peterson, J. Heidelberg, R. T. DeBoy, D. H. Haft, R. J. Dodson, A. S. Durkin, M. Gwinn, J. F. Kolonay, W. C. Nelson, J. D. Peterson, L. A. Umayam, O. White, S. L. Salzberg, M. R. Lewis, D. Radune, E. Holtzapple, H. Khouri, A. M. Wolf, T. R. Utterback, C. L. Hansen, L. A. McDonald, T. V. Feldblyum, S. Angiuoli, T. Dickinson, E. K. Hickey, I. E. Holt, B. J. Loftus, F. Yang, H. O. Smith, J. C. Venter, B. A. Dougherty, D. A. Morrison, S. K. Hollingshead, and C. M. Fraser**. 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science **293**:498-506.

302. **Thornton, J.** 2002. Gene family phylogenetics: tracing protein evolution on trees. EXS **92:**191-207.

303. **Tibayrenc, M.** 1996. Towards a unified evolutionary genetics of microorganisms. Annu.Rev.Microbiol. **50**:401-429.

304. **Tong, H. H., L. E. Blue, M. A. James, and T. F. DeMaria**. 2000. Evaluation of the virulence of a *Streptococcus pneumoniae* neuraminidase-deficient mutant in nasopharyngeal colonization and development of otitis media in the chinchilla model. Infect.Immun. **68**:921-924.

305. **Tong, H. H., M. James, I. Grants, X. Liu, G. Shi, and T. F. DeMaria**. 2001. Comparison of structural changes of cell surface carbohydrates in the eustachian tube epithelium of chinchillas infected with a *Streptococcus pneumoniae* neuraminidase-deficient mutant or its isogenic parent strain. Microb.Pathog. **31**:309-317.

306. **Triscott, M. X. and G. H. Davis**. 1980. The distribution of group B streptococcal serotypes among isolates obtained in three Australian cities. Aust.J.Exp.Biol.Med.Sci. **58**:61-66.

307. **Tully, J. G.** 1993. Current status of the mollicute flora of humans. Clin.Infect.Dis.**17 Suppl 1**:S2-S9.

308. **Tyrrell, G. J., L. D. Senzilet, J. S. Spika, D. A. Kertesz, M. Alagaratnam, M. Lovgren, and J. A. Talbot**. 2000. Invasive disease due to group B streptococcal infection in adults: results from a Canadian, population-based, active laboratory surveillance study--1996. Sentinel Health Unit Surveillance System Site Coordinators. J.Infect.Dis. **182**:168-173.

309. **van Belkum, A.** 2003. High-throughput epidemiologic typing in clinical microbiology. Clin.Microbiol.Infect. **9**:86-100.

310. **van Belkum, A., M. Struelens, A. de Visser, H. Verbrugh, and M. Tibayrenc**. 2001. Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. Clin.Microbiol.Rev. **14**:547-560.

311. **Vandamme, P., B. Pot, M. Gillis, P. de Vos, K. Kersters, and J. Swings**. 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol.Rev. **60**: 407-438.

312. **Vandamme, P., L. A. Devriese, B. Pot, K. Kersters, and P. Melin**. 1997. *Streptococcus difficile* is a nonhemolytic group B, type Ib streptococcus. Int.J.Syst.Bacteriol. **47**:81-85.

313. **Wastfelt, M., M. Stalhammar-Carlemalm, A. M. Delisse, T. Cabezon, and G. Lindahl**. 1996. Identification of a family of streptococcal surface proteins with extremely repetitive structure. J.Biol.Chem. **271**:18892-18897.

314. **Watson, H. L., D. K. Blalock, and G. H. Cassell**. 1990. Variable antigens of *Ureaplasma urealyticum* containing both serovar-specific and serovar-cross-reactive epitopes. Infect.Immun. **58**:3679-3688.

315. **Whitescarver, J., F. Castillo, and G. Furness**. 1975. The preparation of membranes of some human T-mycoplasmas and the analysis of their carbohydrate content. Proc.Soc.Exp.Biol.Med. **150**:20-22.

316. **Whitescarver, J. and G. Furness**. 1975. T-mycoplasmas: a study of the morphology, ultrastructure and mode of division of some human strains. J.Med.Microbiol. **8**:349-355.

317. **Whittam, T. S. and A. C. Bumbaugh**. 2002. Inferences from whole-genome sequences of bacterial pathogens. Curr.Opin.Genet.Dev. **12**:719-725.

318. **Wiedmann, M.** 2003. ADSA Foundation Scholar Award--An integrated science-based approach to dairy food safety: *Listeria monocytogenes* as a model system. J.Dairy Sci. **86**:1865-1875.

319. **Wilkinson, H. W. and M. D. Moody**. 1969. Serological relationships of type I antigens of group B streptococci. J.Bacteriol. **97**:629-634.

320. **Woolhouse, M. E., J. P. Webster, E. Domingo, B. Charlesworth, and B. R. Levin**. 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat.Genet. **32**:569-577.

321. **Wren, B. W.** 2000. Microbial genome analysis: insights into virulence, host adaptation and evolution. Nat.Rev.Genet. **1**:30-39.

322. **Yamamoto, S., K. Miyake, Y. Koike, M. Watanabe, Y. Machida, M. Ohta, and S. Iijima**. 1999. Molecular characterization of type-specific capsular polysaccharide biosynthesis genes of *Streptococcus agalactiae* type Ia. J.Bacteriol. **181**:5176-5184.

323. **Zhao, J., F. Kong, R. Li, X. Wang, Z. Wan, and D. Wang**. 2001. Identification of *Aspergillus fumigatus* and related species by nested PCR targeting ribosomal DNA internal transcribed spacer regions. J.Clin.Microbiol. **39**:2261-2266.

324. **Zheng, X., K. Lau, M. Frazier, G. H. Cassell, and H. L. Watson**. 1996. Epitope mapping of the variable repetitive region with the MB antigen of *Ureaplasma urealyticum*. Clin.Diag.Lab. Immunol. **3**:774-778.

325. **Zheng, X., L. J. Teng, J. I. Glass, A. Blanchard, Z. Cao, M. C. Kempf, H. L. Watson, and G. H. Cassell**. 1994. Size variation of a major serotype-specific antigen of *Ureaplasma urealyticum*. Ann.N.Y.Acad.Sci.**730**:299-301.

326. **Zheng, X., L. J. Teng, H. L. Watson, J. I. Glass, A. Blanchard, and G. H. Cassell**. 1995. Small repeating units within the *Ureaplasma urealyticum* MB antigen gene encode serovar specificity and are associated with antigen size variation. Infect.Immun. **63**:891-898.

327. **Zheng, X., H. L. Watson, K. B. Waites, and G. H. Cassell**. 1992. Serotype diversity and antigen variation among invasive isolates of *Ureaplasma urealyticum* from neonates. Infect. Immun. **60**:3472-3474.

328. **Ziebuhr, W., K. Ohlsen, H. Karch, T. Korhonen, and J. Hacker**. 1999. Evolution of bacterial pathogenesis. Cell Mol.Life Sci. **56**:719-728.

329. **Zuerlein, T. J., B. Christensen, and R. T. Hall**. 1991. Latex agglutination detection of group-B streptococcal inoculum in urine. Diagn.Microbiol.Infect.Dis. **14**:191-194.

Appendices

A. Seven papers directly related to PhD thesis (PDF papers are not attached but can find easily from publisher websites):

1. **Kong, F., X. Zhu, W. Wang, X. Zhou, S. Gordon, and G. L. Gilbert**. 1999a. Comparative analysis and serovar-specific identification of multiple-banded antigen genes of *Ureaplasma urealyticum* biovar 1. J.Clin.Microbiol. **37**:538-543.

2. **Kong, F., G. James, Z. Ma, S. Gordon, W. Bin, and G. L. Gilbert**. 1999b. Phylogenetic analysis of *Ureaplasma urealyticum*--support for the establishment of a new species, *Ureaplasma parvum*. Int.J.Syst.Bacteriol. **49**:1879-1889.

3. **Kong, F., Z. Ma, G. James, S. Gordon, and G. L. Gilbert**. 2000a. Species identification and subtyping of *Ureaplasma parvum* and *Ureaplasma urealyticum* using PCR-based assays. J.Clin.Microbiol. **38**:1175-1179.

4. **Kong, F., Z. Ma, G. James, S. Gordon, and G. L. Gilbert**. 2000b. Molecular genotyping of human *Ureaplasma* species based on multiple-banded antigen (MBA) gene sequences. Int.J.Syst.Evol.Microbiol. **50**:1921-1929.

5. **Kong, F., S. Gowan, D. Martin, G. James, and G. L. Gilbert**. 2002a. Serotype identification of group B streptococci by PCR and sequencing. J.Clin.Microbiol. **40**:216-226.

6. **Kong, F., S. Gowan, D. Martin, G. James, and G. L. Gilbert**. 2002b. Molecular profiles of group B streptococcal surface protein antigen genes: relationship to molecular serotypes. J.Clin.Microbiol. **40**:620-626.

7. **Kong, F., D. Martin, G. James, and G. L. Gilbert**. 2003. Towards a genotyping system for *Streptococcus agalactiae* (group B streptococcus): use of mobile genetic elements in Australasian invasive isolates. J.Med.Microbiol. **52**:337-344.

These seven papers represent work performed during the past 6 years, in Australia, including several published before I formally started my PhD. In the first paper (Kong *et al*., 1999a), I began to think about developing a genotyping system for human ureaplasmas, in particular biovar 1 (later *U. parvum*). It showed some promising findings for later ureaplasma studies. Study of biovar 2 (later the "new" *U. urealyticum*) uncovered important clues to the study of human ureaplasma taxonomy and supported a need to upgrade the two biovars to species level as proposed by Professor Janet Robertson and others. The second paper (Kong *et al*., 1999b) further confirmed this and in the third (Kong *et al*., 2000a), and fourth

papers (Kong *et al*., 2000b), I decided the development of a genotyping system based on the proposed new human ureaplasma taxonomy. The paper of Professor Janet Robertson *et al.* (Robertson *et al*., 2002) finally endorsed the new taxonomy of human ureaplasmas.

Because of the challenges in understanding the pathogenesis of human ureaplasmas infection – I turned to GBS, which has a number of features in common with with ureaplasmas, to help solve the **puzzles**. The three GBS papers (Kong *et al*., 2002a; Kong *et al*., 2002b; Kong *et al*., 2003) are the "**by products**" of these enquires. The body text of the thesis (especially the third section) provides some answers to whether and how we can benefit from the integrated studies of GBS and human ureaplasmas.

B. Five papers indirectly related to PhD work (PDF papers are not attached but can find easily from publisher websites):

1. **Kong, F., S. Gordon, and G. L. Gilbert**. 2000. Rapid-cycle PCR for detection and typing of *Mycoplasma pneumoniae* in clinical specimens. J.Clin.Microbiol. **38**:4256-4259.

2. **Kong, F., G. James, S. Gordon, A. Zelynski, and G. L. Gilbert**. 2001. Species-specific PCR for identification of common contaminant mollicutes in cell culture. Appl.Environ.Microbiol. **67**:3195-3200.

3. **Zhao, J., F. Kong, R. Li, X. Wang, Z. Wan, and D. Wang**. 2001. Identification of *Aspergillus fumigatus* and related species by nested PCR targeting ribosomal DNA internal transcribed spacer regions. J.Clin.Microbiol. **39**:2261-2266.

4. **Robertson, J. A., G. W. Stemke, J. W. Davis, Jr., R. Harasawa, D. Thirkell, F. Kong, M. C. Shepard, and D. K. Ford**. 2002. Proposal of *Ureaplasma parvum* sp. nov. and emended description of *Ureaplasma urealyticum* (Shepard *et al*. 1974) Robertson *et al*. 2001. Int.J.Syst.Evol.Microbiol. **52**:587-597.

5. **Kong, F. and G. L. Gilbert**. 2003. Using *cpsA-cpsB* sequence polymorphisms and serotype-/group-specific PCR to predict 51 *Streptococcus pneumoniae* capsular serotypes. J.Med.Microbiol. **52**:1047-1058.