Adriana María Sanabria Moreno

An Attempt to Elucidate the Genes Encoding the Surface Exposed Proteins R3, Z1 and Z2 in Two Strains of *Streptococcus agalactiae* from Zimbabwe

Master's thesis in Molecular Medicine

Trondheim, June 2014

NTNU – Trondheim
Norwegian University of
Science and Technology

To my mother Cecilia Moreno and my daughter Maria Paz Gonzalez. You are the inspiration of all the things that I do in my life.

AKNOWLEDGEMENTS

This master thesis is a final project for receiving the degree Master of Science in Molecular Medicine. It has been carried out at the Norwegian University of Science and Technology (NTNU), Department of Laboratory Medicine, Children's and Women's Health (LBK) at the Faculty of Medicine, during the fall 2013 and the spring of 2014. My supervisors have been Professor Jan Egil Afset and Professor Finn Drabløs.

First, I would like to express my sincere gratitude to my main supervisor, **Jan Egil Afset** for his continuous support, his patience, motivation enthusiasm, and immense human qualities. His guidance helped me during the entire thesis elaboration. I could not have imagined having a better advisor.

I would also like to thank the following persons: my co-supervisor, **Finn Drabløs** for his support with the bioinformatics analysis. **Magnus Steigedal** directed the cloning experiment part, and at the same time was always kind and open to discuss all the aspects involved in the project. **Konika Chawla** taught me to manage the bioinformatics tools, we made a very good team, **Randhi Valsø Lyng** helped with the immunological testing, and **Janne Fossum** helped when I had problems with the DNA extractions, and **Andreas Radtake** for his interest in the project and his inputs.

Further, I would like to express my gratitude to **Johan Mæland**, who shared his strong knowledge on GBS with me and inspired me to continue. In addition, his help and feedback in the elaboration of this document were extremely valuable and I really appreciate them.

Marit Barstad you are the angel for the molecular medicine master students, thanks for your help, your advices, your friendship and support.

Finally, to **God** for putting my life into this path and giving me the tools needed to cross it. To **my family**, my mother Cecilia Moreno, my sisters Nelcy and Pilar, and my brother Leonardo, because my life is made of a piece from each of you, I love you so much. Finally, I am so thankful to my beloved husband Andres Gonzalez, because you are more than my love; you are my friend, my professor and my inspiration in the academic life. Without you, this process would have been much more difficult.

ABSTRACT

Surface exposed proteins of *Streptococcus agalactiae* (GBS) may be used in serotyping and may have a potential role as vaccine candidates. The proteins R3 and the recently discovered Z1 and Z2 were found to be important markers in GBS from Zimbabwe. However, their prevalence in most geographical areas, and the genes encoding these proteins have so far not been identified. Therefore, the aim of this work was to identify candidate genes (CGs) for the R3, Z1 and Z2 GBS surface exposed proteins in GBS.

Two GBS strains from Zimbabwe, GMFR293 and CMFR30, found to express R3, Z1 and Z2, and Z1, respectively, were genome sequenced. CMFR30 was sequenced on a Pacific Biosciences instrument and assembled to a complete genome. GMFR293 was sequenced by Roche 454 pyro sequencing, which was combined with optical mapping for assembly to a complete genome. RAST was used for *in silico* gene prediction and functional annotation for each genome, for comparison of predicted coding sequences (CDSs) and for comparison with four reference genomes of R3, Z1 and Z2 negative strains. The CDSs were analysed by various bioinformatics tools to identify candidate genes. CDSs were analysed to estimate the molecular weight (MW) of the encoded protein and to predict the potential surface exposition. Based on previous published characteristics of the R3, Z1 and Z2 proteins, CGs were chosen among CDSs encoding proteins of a MW higher than 50 kDa, which had a functional annotation as membrane or surface associated protein or as hypothetical protein (HP) predicted to be potentially surface exposed.

GBS strain GMFR293 comprised 2,037,090 bp and CMFR30 2,062,772 bp, respectively. A total of 2023 CDSs were predicted in GMFR293 and 2060 in CMFR30. Around 80% of all CDSs had a putative assigned function. Unique genes were identified when they were compared with the other GBS strains. 26% of the CDSs from both genomes were predicted as TM proteins. From these, 113 CDSs from strain GMFR293 had a MW >50 kDa: 21 harboured a signal peptide, eight and four had an LPxTG and/or YSIRK signal, respectively, and 14 were identified as lipoproteins. In comparison, of 70 CDSs predicted as TMs in CMFR30 that had a MW >50 kDa, nine harboured a signal peptide, seven and one had an LPxTG and/or YSIRK signal, respectively, and 6 were identified as lipoproteins. Finally, 51 CDSs were chosen as CGs for R3, Z1 and Z2 in the GMFR293 genome, and 32 CDSs were chosen as CGs for Z1 in the GMFR30 genome. Among them were CDSs annotated as hypothetical protein, with putative function and some with predicted function. The CGs identified by *in silico* analyses in this study need to be further tested in experimental analyses, before. This work demonstrates that identification of candidate genes for the surface exposed proteins R3, Z1 and Z2 can be done by comprehensive *in silico* characterization of selected reference genomes.

Among the CGs for R3 was a hypothetical protein of 105kDa which showed 97% similarity with the R5 (BPS) protein encoded by the *sar5* gene published in NCBI. To test the hypothesis whether R5 may be similar or identical to R5, the *sar5* gene was coned in *E. coli* LB21 expression of R3 protein and was thereafter tested by immunological methods. However, the observation that transformants were negative for expression of R3 by immunofluorescence testing may indicate that R3 and R5 are different proteins. However, there may be other possible explanations for these results, which need to be evaluated in further experiments.

In this study we have assembled two GBS strains to near complete genomes, and done a thorough *in silico* characterization of the two GBS genomes with prioritization of potential candidate genes for the surface associated proteins R3, Z1 and Z2. Final identification of the genes encoding these proteins depend on either that more information about the physical and phenotypic characteristics of these proteins becomes available in the future, or experimental analysis of expression of the proteins in overexpression or gene knockout experiments. This work describes the first attempt to identify CGs for these three GBS proteins.

Contents

List of Tables

List of Figures

Abbreviation

AMPs	Antimicrobial Peptides
CAMP	Cationic Antimicrobial Peptides
CDSs	Coding Sequences
CGs	Candidate Genes
CPS	Capsular polyssacharide
C-terminal	Carboxy-terminal
DNA	Deoxyribonucleic acid
EOD	Early onset disease
FbsA	Fibrinogen binding protein
GEIs	Genomic Islands
GBS	Group B Streptococcus
HGT	Horizontal Gene Transfer
HylB	Hyaluronate Lyase
kDa	Kilo Daltons
LOD	Late onset disease
LB	Luria Bertini Medium
Lmb	Lamining Binding Surface Protein
MW	Molecular Weight
N-terminal	Amino-terminal
ON	Over night
PacBio	Pacific Bioscience
PAs	Polyclonal antibody
RAST	Rapid Annotation using Subsystems Technology
RE	Relative error
ScpA	C5a peptidase A
ScpB	C5a peptidase B
Sip	Surface Immunogenic Protein
Srr-1	Serin-Rich Protein 1
Srr-2	Serin-Rich Protein 2
SrtA	Sortase A
SSRs	Small Sequence Repeats
TE	Tris-EDTA Buffer
TM	Transmembrane
TMH	Transmembrane Helix
TMHMM	Transmembrane Hidden Markov Model

1 INTRODUCTION

Streptococcus agalactiae (Group B streptococcus, GBS) is an important human and animal pathogen. In humans, it is the leading cause worldwide of diseases such as neonatal pneumonia, sepsis and meningitis. It is also a cause of morbidity among pregnant women, and was recently found to be pathogenic in immunocompromised adults[1; 2]. GBS strains are classified into ten different serotypes known so far (type Ia, Ib and II through IX) based on differences in capsular polysaccharide (CPS) [3-5]. In addition to the CPS, the proteins exposed on the bacterial cell surface are considered as important markers in typing of GBS. Also, several studies suggest that surface proteins play a major role in GBS binding during the invasion of human mucosal surfaces. Both the capsular antigen and the cell surface proteins are important targets of protective antibodies and as vaccine candidates[2; 4; 6].

GBS express several surface proteins. There are some highly conserved and others are highly associated with specific serotypes[7; 8]. The distribution of serotypes and surface protein vary with geographical region, ethnic origin and the virulence of clinical isolates[5; 9]. Therefore, effective vaccines based on strain variable surface antigens should preferably contain more than a single antigen in order to confer protection against predominant circulating serotypes[6].

These strain variable proteins include: the c proteins (α and β), the R proteins (R1 through R5) and the most recently described Z proteins (Z1 and Z2)[10; 11]. Many of the genes encoding surface proteins have been identified[9]. However, the genes coding R3 and Z proteins are unknown so far. The

identification of candidate genes for R3 and Z through *in silico* methods is the main aim of this study.

On the other hand, in a previous study on surface protein serotype markers in a GBS strain collection from Zimbabwe, it was observed that strains that expressed R3, almost always expressed R5 surface protein (97%)[5; 11]. Since the sequence of the gene encoding the R5 surface protein (*sar5*) has been published and is available in the NCBI data base, experiments using cloning and transformation of *sar5* could help to elucidate if this gene encodes R3 or not and thereby clarify if R3 and R5 are identical or distinct antigens.

To date, there are seven complete whole genome sequences and more than two hundred incomplete genome sequences of GBS strains, available as contigs, in the NCBI database (www.ncbi.nlm.nih.gov). That type of information available in the genomic databases, together with information from studies on serotype surface protein markers on GBS strains, bioinformatics software, recombinant DNA techniques and an accurate prioritization of candidate genes, constitutes keys steps in accelerating the discovery of gene functions of this important pathogen. This type of knowledge may also be of importance for the understanding of pathogenesis and for vaccines development.

1.1. The genus Streptococcus

The genus Streptococcus is a diverse group of Gram-positive bacteria with a considerable importance in medicine and in industry[12]. Various streptococci are important in several ecosystems, as part of the normal microbial flora of

animals and humans. However, they are also one of the most invasive groups of bacteria, being identified as causes of many infections in humans and animals. For instance, some species considered to be common cause of infections include: *S. pneumoniae*, *S. pyogenes*, *S. suis*, *S. dysgalactiae*, *S. agalactiae*, *S. mutans* and *S. viridans*[13; 14].

1.1.1 Classification and features

Taxonomically, the genus Streptococcus is classified as: Bacteria Kingdom, Phylum Firmicutes, Class Bacilli, Order Lactobacillales and Family *Streptococcaceae*. This Family includes the genera: *Enterococcus*, *Lactococcus* and *Streptococcus*. Phenotyphically, Streptococcus strains are Gram-positive cocci, less than 2 μm in diameter, and usually arranged in pairs or chains of varying lengths. They do not form spores, they are facultative anaerobic, catalase negative and have complex nutritional requirements[15; 16].

Streptococci are classified on the basis of colony morphology, hemolysis, biochemical reactions, and beta haemolytic streptococci mainly by serologic specificity. They are divided into three groups by the type of hemolysis on blood agar: β-hemolytic (clear, complete lysis of red cells), α hemolytic (incomplete, green hemolysis), and γ hemolytic (no hemolysis). The serologic grouping is based on ‘’Lancefield grouping’’, which is based on antigenic differences in cell wall carbohydrates (designed by a upper-case letter of the alphabet - groups A to V), in the cell wall pili-associated protein, and in the polysaccharide capsule in group B streptococci [13; 16].

Currently, there are more than 100 species within the Streptococcus genus [14]. Most of them are grouped in six ‘’species groups’’: Pyogenic, Mitis, Salivarius, Bovis, Anginosus and Mutans. However, some of the non-pyogenic streptococci (Mitis, Anginosus and Salivarius) often referred to as

viridans streptococci, have been resistant to satisfactory classification, which is reflected in frequently changing nomenclature and significant problems of identification by phenotypic analysis and by sequencing of 16S rRNA genes[15].

1.2. Group B Streptococccus (GBS): *Streptococcus agalactiae*

Streptococcus agalactiae belongs to the pyogenic group and constitutes the Lancefield's group B Streptococci (GBS)[16]. This Gram-positive encapsulated bacterium exhibits various types of haemolysis on blood agar, mostly β-hemolysis, but 1-3% do not cause any haemolysis[17].

The name GBS comes from the polysaccharide type anchored to their cell wall; the group B specific carbohydrate (GBC), and their serotype comes from their capsular polysaccharide antigen (CPA), which defines the ten different serotypes known today (Ia, Ib, II, III, IV, V, VI, VII, VIII and IX)[18].

1.2.1. Epidemiology and burden of GBS disease

GBS can be found as a commensal bacterium or as an opportunistic pathogen in humans and in animals (ruminants[19] , dogs, horses, guinea pigs[20], camel[21], cattle[22] and fish[23]). It is the leading cause of neonatal sepsis worldwide. In humans, the risk populations are: neonates, pregnant women and non-pregnant adults. In neonates GBS may cause pneumonia, sepsis or meningitis. GBS also causes morbidity among pregnant women, and it is also pathogenic in immunocompromised adults and in the elderly, where an increase in the number of cases have been reported from several countries[1; 2; 18]. The prevalence of GBS and serotype distribution has changed over time and between regions, both within and between countries[18; 24].

Neonatal GBS disease

New-borns are the population most affected by the impact of GBS disease in terms of severity and incidence. It takes place in the neonatal period up to the first 90 days of life. Neonatal GBS disease has been divided in two groups: early onset disease (EOD) and late onset disease (LOD)[25].

Early onset GBS disease (EOD) accounts to approximately 60-70 % of all neonatal GBS disease. It is defined as disease which starts within the first six days of life (0-6 days). EOD infection is usually caused by transmission of GBS from the mother either before or during birth. About 15% to 30% of pregnant women are colonized asymptomatically with GBS in the gastrointestinal and/or genital tracts[26; 27]. Infection takes place via vertical transmission, between the infant and a mother who is GBS carrier during the pregnancy. Around 50% of babies of colonized mothers become colonized, but only 0.5-2 per 1000 live births develop EOD due to GBS infection[27]. Maternal intrapartum GBS colonization is the primary risk factor for early-onset disease in infants. A classic prospective cohort study conducted during the 1980s revealed that pregnant women with GBS colonization were >29 times more likely than pregnant women with negative prenatal cultures to deliver infants with early-onset GBS disease[28]. In addition to maternal colonization, there are others factors associated with an increased risk of neonatal colonization, these include: male sex, black race, prolonged rupture of membranes, prematurity, low levels of maternal anti-GBS antibodies and intrapartum fever[25; 29]. The disease shows rapid progression, with signs like respiratory distress, apnea, or other signs of sepsis, which are often evident at birth or within the first 12 hours of life. It could present as pneumonia, sepsis or meningitis, or a combination of them[25].

Late onset GBS disease (LOD) is defined as infection occurring later in infancy from 7 to 90 days. It is caused predominantly by strains of serotype III**.** In this case, the infection can be acquired from the mother (perinatally) or from environmental sources (nosocomially or from community sources).

The two most common clinical manifestations of LOD are meningitis and bacteraemia. The mortality rate for the disease is significantly lower (2-6%) than the rate of EOD, but the morbidity is high[18].

The burden of GBS disease in new-borns

A review of the current burden of GBS disease was published by Edmonds *et al*. in 2012[29]. The study reported data collected after year 2000 from several countries around the world. In this study, the following were estimated: (a) the incidence of GBS invasive disease and case fatality in infants aged 0–89 days, (b) the incidence of EOD and LOD and (c) the distribution of GBS serotypes in invasive disease specimens.

There was substantial heterogeneity among the studies. Differences in incidence were observed both between and within geographic regions[24]. The overall incidence was of 0.53 cases per 1000 live births (range 0.44 - 0.62) in the European region, 0.67 (0.54 - 0.80) in the Americas and 0.15 (0.03 - 0.07) in Australasia. The mean case fatality rate was 9.6% (7.5 - 11.8). The incidence of EOD was 0.43 per 1000 live births (0.37-0.49) and the case fatality rate of EOD (6.2–18.3) were two-times higher than LOD[29].

The most prevalent serotype in all regions was CPS type III (48.9%) followed by types Ia (22.9%), V (9.1%), Ib (7.0%) and II (6.2%)[29]. The distribution of CPS types seems to be similar in Africa, western Pacific, Europe, the Americas, and the eastern Mediterranean regions, and it has not changed over the past 30 years[18].

Prevention

To prevent GBS diseases in neonates, screening based strategies and intrapartum antibiotic prophylaxis has been implemented in several

European countries and in the USA. Other strategy which is used by several countries including Norway is a risk based strategy where antibiotic treatment is given only in the presence of specific risk factors for GBS disease. These strategies have been shown to reduce the incidence of EOD, but not LOD, and had only a limited impact on the incidence of GBS disease in pregnant women. Therefore, a better method of protecting infants is required. Several different GBS carbohydrates and antigenic proteins have been considered candidates for potential vaccines. However, currently there is not a GBS vaccine available, although vaccination is an attractive preventative strategy. The current status of the GBS antigens that have been studied as potential vaccine candidates are summarized in the table 1-1[30].

Table 1-1. GBS antigens with potential as vaccine candidates[31; 32].

Antigen	Virulence factor	Preclinical studies	Clinical studies
Carbohydrates			
Group B antigen	No	Yes	No
Ia CPS	Yes	Yes	Phase 1 and 2
Ib CPS	Yes	Yes	Phase 1 and 2
II CPS	Yes	Yes	Phase 1 and 2
III CPS	Yes	Yes	Phase 1 and 2
V CPS	Yes	Yes	Phase 1
VI CPS	Yes	Yes	No
VIII CPS	Yes	Yes	No
Proteins			
C proteins			
Alpha	Yes	Yes	No
Betha	?	Yes	No
Epsilon	?	No	No
Rib	?	Yes	No
R proteins	?	No	No
C5a peptidase	?	Yes	No
Sip	?	Yes	No
LrrG	Yes	Yes	No
Pili	Yes	Yes	No

GBS disease in pregnant and post-partum woman

GBS has been reported as a pathogen in pregnant woman, who has a higher estimated relative risk for GBS disease (5.0, range 2.9 - 8.7) compared with non-pregnant women[33]. Maternal colonisation of GBS can vary depending on ethnicity and geographical distribution. The serotypes causing maternal GBS disease have been similar to those that cause EOD[18].

GBS cause different types of disease in mother and child. During pregnancy GBS infection can cause miscarriage, intra-amniotic and urinary tract infection. In the post-partum period a mother colonized with GBS could develop invasive disease, endometritis or chorioamnionitis (inflammation of the fetal membranes). Most pregnancy-associated disease of the mother occurs in the postpartum period[18; 33]. The recognition and identification of maternal GBS colonisation has been the key factor of preventive strategies of perinatal GBS disease.

GBS Disease in non-pregnant adults

GBS in non-pregnant adults cause diseases as: skin, and/or soft tissues infections, bacteraemia, pneumonia and less often problems as osteomyelitis, meningitis and endocarditis associated with considerable morbidity and mortality[18; 34]. The risk factors that have been shown to be related with disease in non- pregnant adults are: older age, diabetes mellitus, cardiovascular diseases, heart failure, history of cancer, alcoholism, obesity and liver and renal insufficiency.

The case fatality rate is markedly higher among adults than among new-borns. However, compared to neonatal disease, the epidemiology in non-pregnant adults has been less studied. The rate of invasive disease is approximately 7 cases per 100,000 non-pregnant adults. The risk of death is

lower among younger adults, and adults who do not have underlying medical conditions. The source of infection for adults is unknown[35].

1.2.2. Virulence factors of Group B streptococcus

The virulence of a microorganism is defined as the degree of pathogenicity or the relative capability of a microbe to cause host damage. GBS encodes a variety of virulence factors that facilitate its ability to invade the host, cause disease, and evade host defence mechanisms. Some of these virulence factors have been identified and characterized, and include: the cell wall carbohydrate antigen (group B antigen and capsular polysaccharides), toxins (β-hemolysin/cytolysin (β-H/C) and CAMP factor), pili and several surface proteins[36; 37].

Cell wall carbohydrates antigens: The two major factors by which this pathogen evades the host defence mechanism are the group B-specific antigen and the capsular polysaccharides.

Group B specific antigen is common to all GBS strains. It is composed of four different oligosaccharides: rhamnose, galactose, N-acetylglucosamine, and glucitol in a highly conserved structural arrangement[38].

Capsular polysaccharides confer serotype specificity and are considered as highly important GBS virulence factors. Currently, there are 10 different GBS serotypes (Ia, Ib, II to IX), each of them antigenically and structurally unique. They are complex carbohydrates composed of approximately 150 repeating oligosaccharide subunits and each subunit contains a mono-, di-, or disaccharide side chain terminating in an N-acetylneuraminic acid (sialic acid) residue. The ten serotypes are different by their arrangements of monosaccharides within the oligosaccharide repeat units[39].

The cell wall anchored polysaccharide capsule is recognized as virulence factor because it inhibits the deposition of alternative complement pathway factor C3b on the surface of the bacterium, causing decreased phagocytosis by macrophages and neutrophils in the absence of serotype-specific antibody[40].

Pore-forming toxins: GBS encodes at least two pore forming toxins: the β-hemolysin/cytolysin (β-H/C) and the CAMP factor. These promote the entry of the pathogen into the host cells, which facilitate their survival and dissemination[37].

β-hemolysin/cytolysin (β-H/C) is encoded by the *cylE* gene of GBS and its expression is associated with the production of an orange pigment. Invasive GBS infections are almost exclusively caused by β -hemolytic strains. The β-H/C is toxic for many eukaryotic cells and it has a strong influence on the intracellular survival of the bacteria inside the host. In addition, the orange pigmentation is related to the protection of GBS against the toxic effects of reactive oxygen species (ROS), generated by the oxidative mechanism of phagocytic killing by macrophages[41; 42].

CAMP factor is another secreted protein with pore-forming properties that has been observed to oligomerize and form discrete pores on susceptible target membranes. Experiments have shown an increased mortality when injection of purified CAMP factor is inoculated in rabbits and mice. However, its role in GBS pathogenesis remains controversial since some authors have observed that deleting the CAMP factor encoding gene (*cfb*) in a GBS strain does not result in attenuation of systemic virulence potential of this strain[43]. A suggested explanation for that observation is that the CAMP factor may be nonessential for GBS pathogenesis. Given their pore-forming abilities, it is also likely that β-H/C may play a compensatory role for the

absence of CAMP factor during infection. So, CAMP factor may only be essential for GBS pathogenesis in host niches where β-H/C activity is diminished[37].

Pili are small cell-surface exposed appendages that have been discovered as important virulence factors in GBS, as well as promising vaccine candidates. Pili mediate GBS resistance to antimicrobial peptides (AMPs), facilitate adherence and attachment of this pathogen to host cells, promote entry into the central nervous system and enhance biofilm formation and resistance to phagocyte killing.

In GBS there are three pathogenicity islands encoding pilin proteins: Pilus Island-1 (PI-1), Pilus Island-2a (PI-2a) and Pilus Island-2b (PI-2b). Pili are high molecular weight structures made of two subunits: the major backbone protein (BP) that is distributed along the pilus structure, and two ancillary proteins (AP), a major (AP1) and a minor (AP2) that are needed for pilus assembly. The pilus 2a backbone protein (BP-2a) is one of the most structurally and functionally characterized components of a potential vaccine formulation against GBS[37; 44].

Surface proteins consist of diverse groups of proteins that mediate bacteria-host receptor interactions. They act as adhesins and may also be involved in the evasion of the immune system. So far, 27 main surface proteins have been identified in GBS. Some of these are anchored to the bacterial membrane while others are just surface expressed proteins[14]. Some surface proteins are highly conserved and present in all GBS strains (see table 1-2) while others are highly associated with specific serotypes (see table 1-3)[7]. In addition, the proteins exhibit size variation between strains, depending on the number of nucleotide repeats in the corresponding genes[45].

Table 1-2. GBS Surface proteins expressed in most of the GBS strains

Protein	**Gene**	**Approx. MW (kDa)**	**Function and characteristics**	**Ref.**
Surface immunogenic protein (**Sip)**	*Sip*	45.5 kDa	Unknown function.	[46; 47]
C5a peptidase A (**ScpA**)	*ScpB*	120 kDa	Promotes resistance to phagocytosis Surface exposed protein.	[48]
C5a peptidase B **(ScpB**)	*ScpB*	140 kDa	Promotes resistance to phagocytosis Surface exposed protein.	[48]
Laminin binding surface protein (**Lmb**)	*Lmb*	34 kDa	Surface exposed lipoprotein. Role in colonization and invasion. Gene is located on a putative composite transposon.	[4]
Fibrinogen-binding protein (**FbsA**)	*fbsA*	110kDa	Binds to human fibrinogen and is involved in the adhesion of GBS to human cells.	[49]
Serine-rich protein (**Srr-1**)	*srr-1*	144 kDa	Promotes colonization by enhancing adhesion.	[50]
Serine-rich protein **(Srr-2**)	*srr-2*	132 kDa	Unknown function. Associated to CPS III. Highly virulent variants have been associated with the gene *srr-2.*	[51]
Cell surface associated protein **(CspA)**	*cspA*	7.3 kDa	Cleaves human fibrinogen and selected chemotaxins. Surface associated protein.	[52]
Hyaluronate lyase **(HylB)**	*hylB*	121.2 kDa	Associated with cell invasion.	[53]
Sortase A (**SrtA**)	*srtA*	27.1 kDa	Required for adhesion to epithelial cells.	[48]

Strain variable proteins are important GBS serotype markers. Among these strain variable proteins are included: the C protein (α and β subunits), the R proteins (R1 through R5), the alpha-like proteins and the most recently described Z proteins (Z1 and Z2)[10; 11]. These proteins are highly complex immunologically, and have sites with different antigenic specificities, and sites which seem to be immunologically identical[54]. Many of the genes encoding surface proteins have been identified[9]. However, the genes coding R3 and Z proteins have not been identified so far (see table 1-3). The characterization of their structures may advance the understanding of some details of the pathogenesis and the vaccines against GBS diseases.

Table 1-3. Strain variable surface proteins of GBS

Surface Protein	Gene	GenBank Number	Approx. MW (kDa)	CPS serotype association
c protein				
Cα	*bca*[55]	M97256	62.5 to 167 kDa	Ia, Ib, II, IX[2; 5; 6]
Cβ	*bac*[56]		130 kDa	Ia, Ib, II, IX[5; 6]
R proteins				
R1/ Alp2	*alp2*[57]	AF208158	74.7 kDa	Ia, III, V[8; 45]
R2/ Alp3	*alp3*[57]	AF245663	77.7 - 95.1 kDa	V, VII, VIII[45]
R3	unknown	-	140 kDa	Ia, II, III,V[5; 7; 10]
R4/ Rib	*rib*[46]	U583333	65-123 kDa	II, III, V, VIII[8; 9; 45]
R5	*sar5*[6]	AJ133114	105kDa	V [24]
Other alp-like proteins				
Alp1/Epsilon	*alp1*	U33554	23.98 - 43 kDa	Evenly distributed and prevalent in bovine strains[45]
Alp 4	*alp4*	AJ488912	38.63 kDa	*NT strains[45]
Z proteins				
Z1	unknown	-	>250 kDa	V [5]
Z2	unknown	-	135 kDa	V [11]

*****No typeable

Alpha-like protein family

Several of the major GBS surface proteins belong to a large protein family called the alp-like proteins (Alp). GBS strains usually contain at least one of the genes encoding Alp-like proteins[58]. Such genes are mosaic allelic structures generated by a recombination of modules at the same chromosomal locus, resulting in sharing of epitopes and immunological cross reactivity between different proteins belonging to this group[45]. Among the alp-like protein family and its encoding genes present in GBS are: Cα protein (*bca*), Alp1 (*epsilon/alp1*), Alp2 (*alp2*), Alp3 (*alp3*), Alp4 and R4/Rib (*rib*)[4].

Alp-like proteins are high molecular mass proteins. The biological function(s) of the Alp family of proteins remains unclear. However, it is known that deletion an Alp-like gene may cause attenuated virulence of the GBS strain (53). All Alp family proteins are constructed in a similar manner: 1) a signal peptide of ~50 amino acids (aa); 2) N terminus composed of ~180 aa; 3) C terminus with a variable number of identical and tandemly arranged repeats, each composed of ~80 aa; 4) C-terminal end of 40-50 aa and with a cell wall anchoring motif. Variable number of repeats results in variation in molecular mass of the proteins. Both the N terminus and the repeat region possess immunogenic domains of different immunological specificities. The level of sequence homology between the N and C termini of different Alps seems to determine the level of immunological cross-reactivity or uniqueness of these domains, for instance if domains are protein-specific.

C proteins

The C protein was the first surface protein which was identified in GBS. It is composed of α and β protein subunits. A GBS strain can express one of them or both[2]. The C alpha protein which is trypsin resistant has been found to be

present in many clinical GBS isolates, and has also been found in other Gram-positive organisms[4]. The calculated mass for the protein is 103 kDa. It consists of an a C-terminal domain (45 amino acids), containing an LPXTG peptidoglycan-anchoring motif, and an N-terminal domain (170 amino acids) followed by a variable number of tandem repeats (82 amino acids each)[46].The C Beta protein which is trypsin sensitive, is unrelated to the other component of the c antigen[4]. It is known to bind different components of the immune system, which suggest that Beta C protein plays a role in virulence. However, it is unknown if it is a virulence factor. The genes encoding the two components of the C protein are located in the same part of the GBS chromosome, but they are not closely linked[4].

R proteins

The R proteins of GBS are cell surface proteins that are resistant to certain proteases. They were described first in group A *Streptococcus*, but were later found to be present in several different B-haemolytic Streptococci (A, B, C, F, G, and L). However, they are not produced by all the strains[59]. Until now, five distinct species of R proteins have been identified in GBS, according to their immunoprecipitation reactions in agarose; R1, R2, R3, R4 and R5, However, some of the R proteins are alp-like proteins; for instance, R4 protein has been found to be identical to protein Rib.

In general, studies regarding serotype markers of GBS strain collections from different geographical locations have shown that the distribution of serotypes and surface protein change with geographical region and the ethnic origin[5; 9]. These proteins have been subject to scientific research with the aim of create vaccines against GBS. An effective vaccines based on strain variable surface antigens should preferably contain more than a single

antigen in order to confer protection against predominant circulating serotypes[6].

1.2.3. R3, Zs and R5 Surface proteins association

Among the less well studied GBS membrane proteins are R3, R5 and the most recently described Z proteins Z1 and Z2. These proteins were found to be present in a high proportion of GBS strains from pregnant women from Zimbabwe, but less common in clinical isolates from Norway. However, knowledge about the occurrence of these markers in other geographical locations is missing.

Among these proteins, R5 is the only one where the corresponding gene has been sequenced. Its relationship to the R3 and the Z proteins (which usually are present in the same strains) has been not established. The genes encoding the Z and R3 proteins have not yet been identified and sequenced, but their expression and some features has been determined by several antibody-based methods such as immunofluorescence[60], whole cell-based ELISA and by Western blotting[5].

R5 surface protein: Initially called BPS (group B protective surface protein), the R5 protein was described in 2002 as a new R-like protein. This protein was identified from the GBS strain Compton R (ATCC9828/ Compton 2560/Prague 2560) which was previously typed as R3 and R4 positive, using a polyclonal antiserum raised against the R protein fraction of this strain to screen a lambda Zap library. DNA sequence analysis showed that R5 belongs to a family of the GBS surface proteins with repetitive structures. It is formed by 979 amino acids and it contains two identical repeats of 76 amino acids separated by a 101 amino acids spacer in the C-terminal region. The protein has a signal

sequence and a membrane anchor region typical of a Gram positive surface protein. Its surface location was confirmed by immunogold electron microscopy using BPS specific antiserum, and it was identified as a unique protein separate from R3 and R4 by immunoprecipitation in agarose gels. R5 did not show cross-reaction with the R1 and R4 and appeared to be different from R3, the other surface proteins present in the Compton R strain[6].

Although R5 was found to be different from R3 in the initial study, later studies done on these proteins have indicated that they are highly related. A study on different serosubtype protein markers detected in GBS strains from Zimbabwe, showed that GBS isolates which were positive for R3 expression were almost always R5 gene positive (97%) as well[24]. In the same study, variable R3 antigen expression was found when some GBS strains were negative for the R3 protein expression in whole cell based ELISA but in a posterior absorption test, R3 expression was confirmed[24]. This results are agree with the previous knowledge about GBS genes may not always be expressed, or expressed in quantities insufficient for detection of the gene product[61].

An attempt to identify the R3 protein sequences from R3 positive GBS strains by mass spectrometry in 2010 resulted in a.a-sequences consistent with R5 protein sequences (unpublished results). This result, together with the inclination of R3 expression and R5 gene possession to occur together made it possible that the encoding genes and gene products, R3 and R5, could be identical. Elucidation of this possibility was one of the goals of the present study.

R3 surface protein: The R3 protein was described in 1972 as one of the members of the R proteins found in GBS[62]. Initially called P protein and then called R3 protein, it has been characterized by immunological methods. In

spite of its expression has been known for a long time, it has been not sequenced until now, perhaps because R3 has been considered of low prevalence. However, the R3 surface protein prevalence in GBS carrier strains has been found variable depending on the geographical site of the study. It was showed in a study comparing GBS strains collections from Zimbabwe and Norway that R3 expression occurred with a much higher frequency in Zimbabwe than in Norwegian isolates[24].

From the immunological experiments it is known that the R3 protein is a high molecular mass protein in the range of 130-140 kDa. It is a trypsin resistant protein that forms a ladder-like banding patterns in Western blot, suggestive of repetitive sequences, and is therefore known as a ladder forming protein (similar to Alp proteins). The R3 protein do not cross-reacted with any of the other GBS proteins identified until know[11].

In two recent studies of GBS from Zimbabwe, it was expressed by more than 20% of the strains, of which 75% belonged to serotype V[10]. There was a higher prevalence in GBS strains from Zimbabwe than in strains from Norway. The studies from Zimbabwe suggest that R3 may be more important in certain geographic areas[5; 10].

Z1 and Z2 Surface proteins: Currently there are two Z proteins, which has been identified and described recently[5; 11]. Initially, an unrecognized protein antigen called Z was detected because a supposedly R3 specific polyclonal antibody contained Z antibodies in contrast to the R3 monoclonal antibodies[5].

Z1 was found to be expressed by**: i)** a R3 reference strain (Praga 10/84, ATCC 49447) and **ii)** in 27.2% of GBS strains from Zimbabwe (usually in combination with R3 protein expression) and **iii)** in a lower number in GBS

strains from Norway, usually in combination with R3 protein expression. The new protein was shown to be similar physicochemically to R3, but immunologically distinct[5]. In a subsequent study, antiserum considered to be Z specific contained antibodies against two different antigens as well. They were identified from the pattern generated by immunoblotting with the strain 08-17 which resulted from its expression of the two proteins, later called Z1 and Z2. The original anti-R3 polyclonal antybody contained anti-Z2 antibodies due to the fact that for its preparation the antiserum had been cross-absorbed by a Z1-expressing strain but not by a Z2-expressing strain.

The genes encoding the two Z proteins have not been identified until now. However, immunologic methods such: ELISA, FAT, and Western blotting using the polyclonal antibodies to Z1, Z2, and R3 have been used to characterize and find associations between these proteins.

From the experiments using the methods previously mentioned it was possible to estimate the molecular mass of the proteins. Z1 is a high molecular mass protein of >250 kDa while Z2 is a lower molecular mass protein of ~135 kDa. The Z proteins generate multiple stained bands and have similar chromatographic features with respect to aggregate formation and charge; similar to the R3 protein as well.

Twenty eight GBS isolates of human and bovine origin from Zimbabwe and Norway were tested for expression of Z1, Z2 and R3 using antibody based methods. It was found that these GBS strains expressed one, some or none of these proteins. The association between the proteins varied. Twenty of the strains express any of the three proteins, four expressed all three antigens, two expressed Z2 and R3, one expressed Z1 or R3 only, and none expressed only Z2. In general the three proteins occurred with particularly high frequency (80%) in the CPS type V isolates[11].

The identification and characterization of the genes that encode R3 and the Z proteins and studies about the relationship between them and R5 will give a clearer and complete landscape of the genetic basis of such GBS surface associated proteins. This information will help to develop molecular methods for a more complete GBS serotyping and to study the potential of these proteins as vaccine candidate components.

1.3. Bacterial genome evolution

Bacteria retain most of their genetic information from generation to generation. However, they also need to develop strategies that allow them to acquire new genetic material in their genomes to adapt and survive in an environment that change continually. Genomes of more closely related bacteria are more conserved but the genome variability exists within different genera and among different isolates of a single bacterial species.

In the bacterial pan-genome, the ‘’core genome’’ is the conserved stable regions with relatively low mutational capacity containing the genes present in all strains. The "dispensable genome" is composed by genes that are present in more than one but not in all the strains, while the "unique genes" are specific to a single strain. The variable genome represents the total amount of foreign DNAs available for recipient cells. Free living bacteria genomes often carry phages and repetitive sequences mediating genetic rearrangements. Their genetic stability is associated with the genomic content of repeated sequences, mobile genetic elements, and influenced by the bacterial lifestyle. All this takes part in the bacterial genome evolution. The mains mechanisms that contribute to the plasticity of bacterial genome

are: the acquisition of DNA (gene gain), and the loss of genetic information (gene loss)[63]. The molecular and genetic mechanisms leading to these changes are summarized in the table 1-4.

Table 1-4. Mechanisms contributing to bacterial genome plasticity [63].

Genetic element or mechanism	Consequences
Gain of properties	
Point mutation	Alteration of gene expression
Homologous recombination	DNA rearrangements: inversion, duplication, deletion of DNA. Integration of horizontally acquired DNA
Transformation	Gain of additional genetic information
IS elements, composite transposons	Insertion, deletion, inversion of DNA, alteration of gene expression.
Integrons	Transfer of genes, DNA rearrangements
Conjugative transposons, plasmids	Conjugation Horizontal gene transfer Mobilization of other plasmids
Bacteriophages	Generalized or specialized transduction Horizontal gene transfer
*GEIs or PAIs, pathogenicity islets	Horizontal gene transfer. Integration and deletion of large DNA regions.
Loss of properties	
Point mutation	Alteration of gene expression, loss of function
Homologous recombination	DNA rearrangements, deletion of DNA, integration of horizontally acquired DNA
Transposition	Alteration of gene expression, loss of function

*GEI, genomic island; IS, insertion; PAI, pathogenicity island.

1.3.1. General features and genetic evolution of the GBS Genome

Sequencing of the GBS genome has provided valuable information to the understanding of this pathogen and how it cause disease in humans. To date, eight complete sequences and 292 draft GBS genomes have been deposited in the National Centre for Biotechnology Information database (NCBI), and a database called Strepto-DB for comparative genome analysis of group A (GAS) and group B (GBS) streptococci (http://oger.tu-bs.de/strepto_db) [64]. Among the complete GBS genomes some strains belong to the major disease causing GBS serotypes in humans and some isolates are from animal sources (See table 1-5). The GBS genomes are in the range of 1,800 to 2,160 Kb in size with approx. 1,710 to 2,055 predicted protein coding genes and a G+C content about 35%.

Table 1-5. GBS complete genome sequences in the NCBI database.

GBS Strain	Source	Genome Size (Mb)	GC%	Genes	Proteins
2603V/R	Human isolate	2.16	35.6	2,279	2,127
09mas018883	dairy cattle	2.14	35.5	2,190	2,089
A909	Human isolate	2.13	35.6	2,136	1,996
GD201008-001	Tilapia	2.06	35.6	2,088	1,964
ILRI005	dairy cattle	2.11	35.4	2,256	2,155
ILRI112	milk of camel	2.03	35.3	2,173	2,073
SA20-06	Tilapia	1.82	35.6	1,872	1,710
138P	-	1.84	35.5	1831	1539

The forces that drive the genome evolution of GBS have been studied by combining experimental and *in silico* approaches. Further analysis of the complete genome sequences using comparative genomics studies from eight sequenced strains from human and animal sources (2603V/R, NEM 316, A909, CJB111, H36B, 18RS21, COH1 and 515) has defined the composite

organization of GBS genomes. It was estimated that approximately 80% of genes belong to the core genome (minimum 1,806 genes) and around 20% to the dispensable genome. The number of shared genes in each genome varied because of gene duplications and paralogs. The number of new genes was decaying exponentially when a new sequence was added to the analysis. The number of new genes when comparing two genome sequences was in average 161, and this number decrease to 33 new strain specific genes after the eight genomes were added. The number of genes found in a single strain were 358 genes conformed by a varied number depending of the GBS strain (2603V/R (47), NEM316 (137), A909 (13), CJB111 (14), H36B (61), 18RS21 (13), COH1 (31) and 515 (20)). In other words, the number of genes classified as core genome, accessory genome and strain specific genes depended to high degree on the number of compared strains and, the more strains compared the lower number of core genes, higher number of accessory genes, and higher pan genome. All these aspects contribute to GBS genetic diversity[65].

In addition, genes classified as strain specific genes tended to cluster in genomic islands. These are highly variable between the different strains and for instance, the analysis of the NEM316 genome revealed 14 putative chromosomal pathogenicity islands containing surface proteins[66]. These data could suggest that horizontal transfer (HGT) is an important evolutionary force within GBS[67].

HGT is the processes that permits the exchange of DNA among organisms both within and between species[68]. The horizontal gene transfer can occur by one of three main mechanisms: transformation, transduction, or conjugation. **Transformation** refers to the process when a cell takes up isolated DNA from the environment and has the potential to transfer DNA between

distantly related organisms. A second mechanism is **conjugation**, which is defined as the direct transmission of DNA from one cell to another and the last one, **transduction**, which is phage mediated transfer of genetic materials. In the past few years, there has been growing evidence that HGT may play a vital role in the evolution of bacterial genomes[69].

The available GSB genome sequences have been reported to contain strong evidence of HGT events leading to virulence acquisition and genetic diversity. For instance, it has been suggested that the genes encoding the virulence factors capsular polysaccharides and surface membrane proteins were acquired by HGT[70; 71]. Also, it has been demonstrated that large conjugal exchanges have contributed significantly to the genome dynamics of GBS, strengthening the understanding of the role of integrative conjugative elements in the dynamics of bacterial chromosomes[71].

Repetitive sequences are often found in the genome of GBS strains, for instance the genes encoding the alpha-like protein group which has a region with a variable number of identical, tandem repeats[72]. Other data suggests that small repeats (SSRs) contribute to genome plasticity in GBS. Comparative genomic analysis of eight bacterial genomes showed evidence of genotypic variation in GBS caused by slipped strand mispairing in the SSR regions. A total of 2,233 SSRs were identified in the GBS reference genome 2603V/R. When these loci were examined in seven other GBS genomes, a total of 56 SSR loci were found to exhibit variation, where gain or loss of repeat units was observed in at least one other genome, resulting in aberrant genotypes. Changes by such a mechanism also lead to antigenic variation that could be used to escape selective pressure of specific antibodies[73].

Studies on genetic diversity in streptococcal species showed that GBS clusters together with *S. dysgalactiae* subsp. *dysgalactiae* and *S. dysgalactiae* subsp. *equisimilis*[74]. The presence of almost identical genes, mosaic genes and mobile genetic elements between 55 different streptococcal species are signals of genetic recombination events. This is thought to be the main cause of genetic change in several streptococcal species. On the other hand, genetic diversity in GBS populations has been studied by different methods, like multi-locus variable number of tandem repeats (MLVA)[75] and multilocus sequence typing (MLST)[76; 77]. The results obtained allow a better knowledge of the population structure, the genetic lineage and/or long-term evolutionary development of the GBS species. The results of MLST analyses led to the classification of GBS in different clonal complexes.

1.3.2. Bacterial genome sequencing and analyses

Bacterial genome sequencing and analysis is an important field of biological sciences. This approach was developed by a diverse group of scientists interested in a variety of topics related to genetics and the evolution. The mains steps that cover this field are: sequencing, assembly, ordering of contigs, annotation, genome comparison and extraction of common typing information[78].

Genome sequencing

DNA sequencing is the process to determine the nucleotide order of a given DNA sample. Genome sequence analysis allows to get information for the study of organisms, such as constitutive features (predicted encoding regions, ribosomal RNA operons, IS elements, repeat regions, G-C content, origins of replication, operon structure and so on, and assignment of gene name and functional role(s).

There are different sequencing techniques. The oldest and still used has been the Sanger DNA sequencing method. This technique uses sequence-specific termination of a DNA synthesis reaction using modified nucleotide substrates and it was used in the Human Genome Project (1990). It is considered as a ‘’first generation’’ technology and since its beginning in 1977 the method has been improved. Currently, this allows sequencing of up to 384 DNA fragments of up to 1.000 bp in length, with an accuracy value higher than 99.99%[79].

The genome sequencing technologies has continued progressing and improving over time. Newer methods have been developed and are referred to as next generation DNA sequencing (NGS). NGS technology combined with advances in bioinformatics, have resulted in what is called the new era of genomic science[80; 81].

Nowadays, genomes from humans and other model organisms have been sequenced. At the time of writing (04/2014), there were around 18,915 genome projects publicly available. In total, 3 041 complete genomes were finished while 15,874 were available as drafts. 362 belong to studies in archaeas, 906 in eukaryotes and 17,647 in bacteria (http://www.genomesonline.org). This reflects the considerable developments in sequenced genomes over the past decade.

The first bacterial genome sequenced was *Haemophilus influenza*[82] followed by *Mycoplasma genitalium*[83] in the same year (1995). They have been considered a milestone in microbial and genome sequencing studies[84]. Currently, there are more than 17.000 of microbial genome sequences (finished and unfinished) available in the data bases and bacterial genome sequencing technologies have been progressing and improving over time with developed instruments and platforms that allow facing the DNA

revolution; however, the functional analysis of encoded genes is still a challenge[85].

Next Generation Sequencing (NGS)

There are several NGS technologies commercially available today, including Roche/454 FLX, the Illumina/Solexa, life/APG, Helicos Biosciences and the most recently launched platform Pacific Bioscience[84]. Among the most important aspects that distinguish one technology from another are the combination of specific protocols, the type and quality of data produced, biological applications and their cost. The steps involved in Next Generation Sequencing includes: template preparation, sequencing and imaging, and genome alignment and assembly. Due to differences in methodology and technology between the NGS platforms each platform has advantages and disadvantages that should be taken into account when choosing the technology to use in specific sequencing projects and for analysing sequence data, both own and publicly available data[80].

Genome assembly

Most of the NGS technologies produce many data, but short sequence fragments (SRSs). These SRSs have to be assembled into continuous sequences referred to as ‘’contigs’’, which then need to be ordered and oriented to get a full genome sequence[86]. For assembly of reads into contigs, several annotation systems have been developed, for instance the Roche 454 FLX Titanium platform or the Newbler assembler from Roche.

Newbler is a software developed for *de-novo* genome assembly projects based on the Roche 454 sequencing platform[87]. This assembler was developed especially for working with the reads from the Roche/454 Life Science sequencing technology. It has been used for many large and small

genome assemblies (many bacteria). However, this assembler is not open source software which limits its uses.

During the assembly process, the program identifies pairwise overlaps between reads, constructing multiple alignments of overlapping reads. Then it introduces breaks into the multiple alignments in the regions where consistent differences are found between different sets of reads, giving as result a preliminary set of contigs that represent the assembled reads. Then a consensus base call is generated by using quality and flow signal information for each nucleotide. The Output consists of the contigs based on consensus sequences and the corresponding quality scores.

An additional approach to ensure correct assembly and contigs order of the genome is the use of physical maps constructed by restriction of the genome with enzyme digestion. This approach helps to improve the final genome assembly and also to verify the finished sequence data. Optical mapping is an approach to create ordered restriction maps from assemblies of single molecules[88].

After the sequencing process and the genome assembly, describing the status of such genome projects is important. The picture is further complicated by the lack of a community-accepted nomenclature that clearly defines levels of sequence completeness. Two, are the most common standards for purposes of sequence analysis: *finished genome sequence*, which represents a complete genome sequence, where the order and accuracy of every base pair have been verified. In contrast, a *draft genome sequence* represents a collection of contigs of various sizes with unknown order and orientation, that contains sequencing errors and possible misassembles. Finished data of the highest quality is the most desirable state for a genome sequence. However, this requires a relatively rigorous quality check and verification

with the aid of manual laboratory and computational processes[89]. Then, with the advent of the latest sequencing technologies, the terms "draft" and "finished" are no longer sufficient to describe the varying levels of genome sequence quality being produced, and terms such as "complete draft" or "essentially complete", "standard draft", "high quality draft", "improved high quality draft" and "Noncontiguous finished" have appeared to describe different standards [90].

Bacterial optical mapping

Optical mapping is a method for whole genome analysis that was introduced in 1995[91]. It may be used for genome assembly after sequencing[92]. The process comprises the creation of a genome restriction enzyme map of an organism, from very small quantities of high molecular weight DNA.

The technique includes running the DNA sample through nanochannels, which later are fixed in place, stained, digested and visualised using an optical microscope[88]. The individual fragments within the molecules of DNA are then measured and the molecules are assembled together according to matching patterns of cleavage, thus creating a *de-novo* restriction enzyme map[93]. Optical mapping provides a graphical representation of the location of restriction sites in the whole genome of the organisms under study. The maps are then analysed by computer-assisted interpretation software such as MapSolver™ developed by the company OpGen (http://www.opgen.com/). This tool allows the alignment and comparison of the contiguous optical map with the *in silico* restriction map, determined for the partially complete whole-genome assembly.

In microbiology, several studies have been done using the applications of the optical mapping approach. Several complete bacterial genomes have been assembled by integrating data from Roche 454 NGS with optical mapping

assembly. For instance: *Providencia stuartii*[94], *Xenorhabdus nemathopila* (ATCC 19061) and *X. bovienii*[95] *Yersinia pestis* KIM[96], *E. coli* and *S. cerevisiae* among others.

1.3.3. Gene prediction and annotation

Assembly is followed by gene prediction and annotation. The gene prediction process is the first step in genome analysis. This is a computational process in which regions of the DNA containing coding genes are identified. Annotation of a genome involves prediction of the limits of the genes (start codons and stop codons in all the open reading frames (ORFs)) and other genomic elements as well as the prediction of the function of the gene products. Today, the annotation of a gene involves integration of information from genome sequencing, bioinformatics analyses and experimental validation.

Rapid Annotation using Subsystems Technology (RAST) is an automatic database for rapid and accurate annotation of bacteria and archea genomes, which has been used by many researchers for prediction of gene function and discovery of new pathways. It was introduced in 1997 and so far, over 12,000 users worldwide have annotated more than 60 000 genomes using RAST[93].

The program identifies protein encoding genes, assigns gene function, predicts which subsystems are represented in the genome and use them to construct the metabolic network. In addition, RAST supports detailed comparison against existing genomes, determination of genes that the genome has in common with specific sets of genomes (or, genes that distinguish the genome from those in a set of existing genomes).

The RAST server implements two classes of asserted genes: **(1) subsystem based assertions** which are based on recognition of functional variants of subsystems (abstract functional roles), and **(2) non-subsystem based assertions** using integrated common approaches from a number of tools.

RAST is composed by a set of proteins (a protein families collection called ''*FIGfams*''), a family function and a decision procedure. When a new genome is submitted to RAST, genes are called and their annotations are made by comparison to the FIGfam collection. The program takes a protein sequence as input and decides if a protein could be added to the family by looking if it is globally similar to the members and shares a common function. They can be placed in the same family if they were located in the same subsystem (same functional role), the similarity region shared by the two sequences are above 70%, and if they come from closely related genomes. With these parameters, the program is able to recognize well over 90% of the genes in a newly sequenced strain[97].

Basically, the steps used by RAST to get the genome annotations are: (1) call the tRNA and rRNA genes, (2) make initial protein-encoding genes calling using GLIMMER3 to get *putative genes*, (3) establish a phylogenetic context by using a small set of representative protein sequences (universal in prokaryotes) to find the closest phylogenetic neighbours. For each detected gene the starting position is adjusted and moved from putative to determined genes, (4) a targeted search based on FIGfams that occurs in closely related genomes because they are likely to be found in the new genome, (5) recall protein-encoding genes using the previous training set, (6) processing the remaining putative genes against the entire FIGfam collection, (7) clean up remaining gene calls to remove overlaps and adjust starting positions using blast to determine similarity based evidence, (8) process the remaining,

unannotated protein encoding genes, and finally (9) construct a metabolic reconstruction (a collection of the active variants of subsystems that have been identified) connecting genes in the new genome. The metabolic network is assembled using biochemical reaction information associated with functional roles in the subsystems[93] .

1.3.4. GBS genome comparisons

Comparative genomics is the analyses between multiple genomes of closely related bacteria. It has allowed a better comprehension on many genomic variations, answering biological questions related to bacterial evolution, physiology and pathogenicity. In addition, comparative genomics analyses have led to an improvement of the process of genome annotations[98].

In the special case of GBS, the availability of genome sequences has allowed a better understanding of the evolutionary path followed by this species that belongs to a genus that encompasses many harmful pathogenic species. Comparative genomic studies in GBS have been done by Tettelin *et al.*[99] using multiple genomes of *Streptococcus agalactiae* strains and other species of pathogenic Streptococci (*S. pneumoniae* and *S. pyogenes) to* elucidate the molecular basis for GBS virulence. These studies revealed that the GBS genome has a substantial similarity with those of the related human pathogens *S. pyogenes* and *S. pneumoniae.* However, GBS was shown to be different from the other streptococci in several metabolic pathways and related membrane transport systems that probably relate to adaptation to distinct niches in its human and animal hosts[99]. On the other hand, the study also revealed that there was extensive genomic intra-species diversity.

Tettelin et al.[65] in a later study explored gene variability within the GBS species using the complete genome sequence of eight GBS representing the five major serotypes (human isolates and one of bovine origin). The results

suggested the composition of the GBS genome which can be described by its "pan-genome" formed by a core genome and a dispensable genome that consists of genes shared by all the strains studied and probably encodes functions related to the basic biology and phenotypes of the species.

1.3.5. Candidate gene prioritization

The candidate gene approach has been a pioneer in many fields of genetic, studies including epidemiology to find casual gene variants for candidate or genome wide association studies. *in silico* tools gives fast, efficient and reliable results, in addition to be an alternative to costly collections of experimental data[100]. Accurate prioritisation of candidate genes, constitutes a key step in accelerating the discovery of gene functions[101].

In silico candidate gene prioritisation ranks genes based on the features associated with the genes and the function of interest. Studies suggest that phylogenetic profiles provide a valuable tool for predicting gene-function linkage. It is because the phylogenetic profile of a gene is a reflection of its evolutionary history and can be defined as the differential presence or absence of a gene in a set of reference genomes[101]. For example, in GBS phylogenetic profiles of all GBS genes across 467 bacterial reference genomes were determined by candidate against all BLAST searches, which were then used to identify candidate virulence genes[101].

2 AIMS OF THE STUDY

2.1. Main objective

The main aim of this study was to identify candidate genes for the R3 and Z surface-exposed proteins in two GBS strain isolated from pregnant women in Zimbabwe.

2.2. Specific objectives

- First, to get as complete as possible genome sequences from two Zimbabwean GBS strains GMFR293, known to express the R3, Z1 and Z2 surface exposed proteins and CMFR30 that expresses only Z1.
- To use *in silico* methods to identify candidate genes for the R3, Z1 and Z2 proteins based on analysis of the sequence functional features, assisted by genome comparison approaches.
- To clarify if the R5 and R3 surface proteins are identical trough cloning experiments.

3 MATERIAL AND METHODS

3.1. GBS strains

The reference and prototype GBS strains used in this study are listed in table 3-1, including their capsular polysaccharide type and their serosubtype proteins markers.

Table 3-1 GBS strains used in this project.

GBS strain	CPS	GBS surface proteins	Procedure
GMFR293	V	**R3**, R4**, Z1, Z2**	Genome sequencing. Genome comparison.
CMFR30	Ib	Cβ, **Z1**	R5 PCR
2603 V/R	V	R4	
NEM316 (ATCC12403)	III	Alp2	
A909 (NCTC 11078)	Ia	α, β	Genome comparison
515	Ia	Alp1, **Z1**	
04-534	IX	Cα, Cβ, R3, Z1, Z2	
2603V/R	V	R4/ Rib	
08-17	V	R3, Z1, Z2	R5 PCR
161757	V	alp3	
ComptonR (NCTC9828/Prage2560)	NT	R3, R4, R5[a]	

[a] R5 was tested by PCR for the gene encoding *sar5*, not by antibody based methods.

The isolates were two strains from Zimbabwe which were chosen for sequencing, based on the presence of proteins markers reported in previous studies[11; 24]. The GMFR293 and CMFR30 strains were found to express the surface proteins of interest; R3, Z1 and Z2, and Z1, respectively. The rest of the strains listed were GBS reference strains of different serotypes used in different steps through this project. Most of them have been previously sequenced and are published as complete or draft genomes in the NCBI

database. All the strains were available at the GBS strain collection of the Department of Medical Microbiology, St. Olavs Hospital, Trondheim, Norway.

3.2. Genome sequence, assembly, annotation and Candidate genes prioritization

The procedure made to select the candidate genes coding for the R3 and Z surface proteins is summarized in the following flow chart, and described in detail in the next sections.

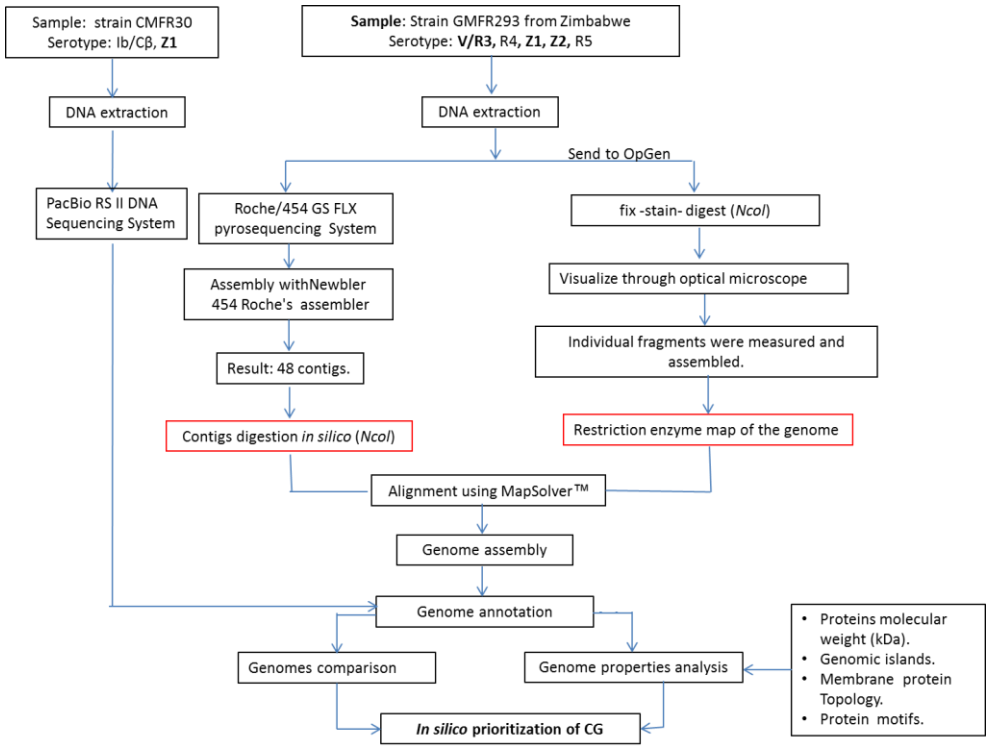


Figure 3-1. Flow chart explaining the methodology used to obtain candidate genes for R3, Z1 and Z2.

3.2.1. Genome sequencing and assembly

GMFR293 genome sequencing and assembly

The procedure to get the complete genome sequence of the GBS strain GMFR293 considers data produced through 454 Roche pyrosequencing and optical mapping.

Genome sequencing of strain GBS GMFR293 by the 454 GS-FLX sequencer resulted in a total of 159,529 reads of 59,021,738 bp in size. The reads were assembled using the Newbler GS *de novo* assembler software (www.454.com) using default assembly parameters.

GMFR293 optical map

In parallel to the 454 sequencing process, the same strain was sent to the OpGen Company (www.opgen.com) for optical restriction mapping of the bacterial genome. An optical map is an assembly of a number of partial restriction fragment maps into a single complete genome restriction map. In brief, the method consists of running the DNA through nanochannels (Figure **a**), fixing in place, staining, digestion with the restriction enzyme (Figure **b**), and visualization of fragments using an optical microscope interfaced with a digital camera. The individual fragments within the molecules of DNA are then stained, measured (Figure **c**) and assembled together according to matching patterns of cleavage (Figure **d**), thus creating a de novo restriction enzyme map (Figure **e**).

The optical map was based on the restriction of the GBS strain GMFR293 genome with the enzyme *Ncol*.

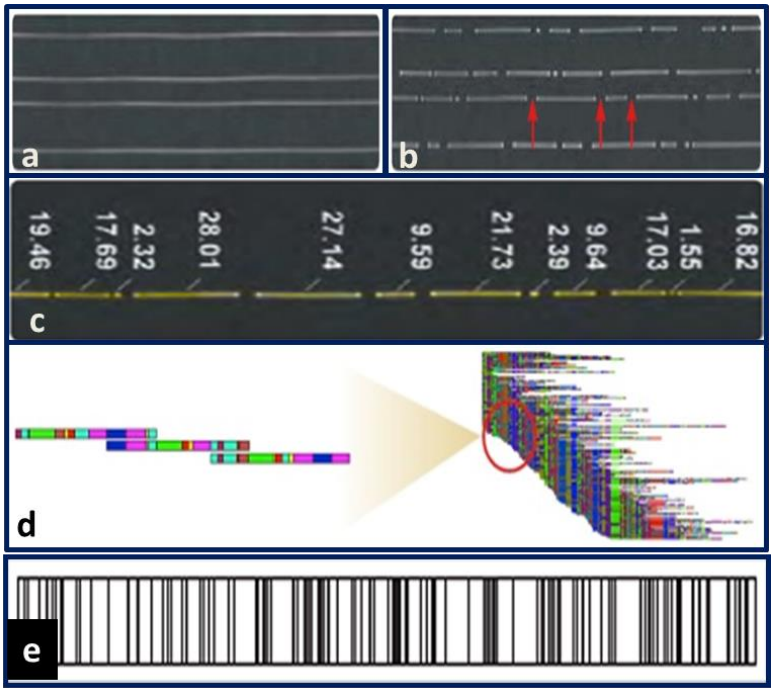


Figure 3-2. Steps in the creation of an optical map (http://www.opgen.com/)

Genome assembly

Using the MapSolver™ software from OpGen, the contigs from the assembly of reads from the 454 sequencing process were digested *in silico* with the same enzyme (*Ncol*) to create another optical map of the GMFR293 contigs.

Contigs restriction maps were aligned to the optical map of the GMFR293 genome. Thereby, many of the contigs could be ordered and oriented. In cases where misassembled contigs were identified, they were broken/joined and realigned. Thereby some of the gaps between contigs were closed. The remaining gaps were identified and their sequences were found by using Blast alignment of all contigs on the closest reference genome to find their

sequences. The closest reference genome was GBS 2603V/R. It was chosen by creating similarity clusters between GMFR293 and several GBS genomes with the tool clustering of the MapSolver™ software.

Assessment of the alignment of the *in silico* map of contigs and the optical map of the GMFR293 genome

The assembled genome produced an *in silico* map. This was subjected to verification through identification of uncertain regions, which were identified searching for differences in the restriction patterns between the optical and the *in silico* maps (DRP1). The parameters evaluated to identify them were: missing fragments and false/missing cuts between the optical map and the assembled (*in silico*) genome. After that, the relationship between fragment size and relative error (RE) was calculated (see Equation 1) in the optical map fragments, and plotted against the *in silico* map fragments.

$$RE = \frac{(in\ silico\ map\ fragment\ size) - (optical\ map\ fragment\ size)}{(in\ silico\ map\ fragment\ size)} \quad \textbf{(1)}$$

In parallel, FASTA sequences from the GBS reference genomes A909, 2603V/R and NEM316 (available in NCBI) were converted to *in silico* restriction maps using the MapViewer software (OpGen technologies, Inc), for direct comparison between the three GBS reference genomes and the GMFR293 optical maps. This comparison was performed to calculate fragments size variation, to identify restriction pattern differences (DRP2) and to use these data to identify which locations in the assembled genome

that would need to be verified through experimental work to validate the finished sequence.

GBS CMFR30 genome sequencing and assembly

The CMFR30 genome was sequenced using the PacBio RS II DNA Sequencing System. The sequencing process resulted in 98,249 reads with an average read length of 3,407 bp, and a total number of bases of 399.9 Mb. The obtained genome sequence was used for the comparative genomic analysis.

Reads obtained from the PacBio sequencing process were assembled using HGAP v2 (Pacific Biosciences). The process resulted in one single contig of 2,062,772 bp with 146,86 times average coverage. Further local assembly efforts were therefore not needed.

3.2.2. Genome annotation

The assembly of the GMFR293 genome and the sequencing of GBS CMFR30 was followed by gene prediction/annotation in which DNA regions containing coding sequences (CDSs) were identified. Annotation and analysis were performed using RAST[93] (Rapid annotation using subsystem Technology, http://rast.nmpdr.org) which uses by default the software GLIMMER3 to perform gene prediction. In order to enrich the annotation process, functional annotations were done in addition by using the web server webMGA[102] (http://weizhong-lab.ucsd.edu/metagenomic-analysis/server/cog/), which performs function annotation by using the RPS-Blast program at the Cluster Orthologous Groups (COG) database (prokaryotic proteins).

Physicochemical parameters of proteins

The software ProPAS (Protein Properties Analyses Software)[103] was used to calculate several physicochemical parameters of the proteins, including the isoelectric point (pI), hydrophobicity (Hy) and molecular weight (MW). CDSs coding for high MW proteins of more than 50 kDa was one of the parameters used to prioritize possible CGs for the R3, Z1 and Z2 proteins.

Prediction of Genomic Islands (GEIs)

GEIs are discrete DNA segments, which may be mobile or not, or no longer mobile, which differ among closely related strains[104]. In GMFR293 and CMFR30 genomics islands were predicted by using the IslandViewer software tool (http://www.pathogenomics.sfu.ca/islandviewer) which integrates the two sequence composition GI prediction methods SIGI-HMM and IslandPath-DIMOB, and a single comparative GI prediction method IslandPick[105]. In this process, default parameters were used.

Proteins topology

The methodology used to predict the potential location of the encoded proteins in the GMFR293 and CMFR30 genomes was based on prediction of transmembrane helix (TMH) and of retention of signal sequences that govern the transport and localisation of a protein in a cell. This was done to identify CDSs encoding potential surface exposed proteins, which could be membrane or secreted protein.

Transmembrane helix prediction (TMH)

Transmembrane helices are characteristic for membrane proteins. In this study we used TMHMM (a hidden Markov model (HMM)) for predicting the number of transmembrane helices, their location, and in/out orientation to

all the CDSs in the GMFR293 and CMFR30 genomes. Proteins predicted as transmembrane were considered potential candidate genes for the R3, Z1 and Z2 proteins.

Identification of motifs or domains

Pfam (http://pfam.sanger.ac.uk/) and HMMER (http://www.cbs.dtu.dk/services/TMHMM/) were used to search motifs described as cell wall anchoring or binding domains in Gram positive bacteria included Streptococci. Motifs or domains detected were considered significant if they obtained a score higher than 10 and the per-domain E-value was lower than 0.1[106]. These were used as query profiles in the analysis of the CDSs from the GMFR293 and CMFR30 genomes. ScanProsite (http://prosite.expasy.org/) was used for pattern recognition of lipoprotein, LPxTG and YSIRK signals, and SIGNALP (http://www.cbs.dtu.dk/services/SignalP/)[107] was used to identify signal peptides. Candidate lipoprotein signal peptides were flagged by matches with the pattern {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [99].YSIRK signal through the pattern [WYF][ST][IL][RK][KR]xxxGxxSV and LPxTG signal by matches with the pattern [LIF]PXT[GSN].

3.2.3. In silico genome comparison

Protein coding genes of GMFR293 and CMFR30 were compared against each other and also against genomes of four reference strains: A909, NEM316, 515 and 2603 V/R. The comparison was done by RAST comparison tool at the protein sequence level using BLASTP. Genome comparison was used to assist the selection of the CGs and to identify novel surface proteins.

3.2.4. In silico prioritization of candidate genes

The strategy followed for finding candidate genes was based on two complementary approaches. One was to compare the GMFR293 and CMFR30 genomes to related reference genomes published in the NCBI database, while the other approach was to test protein-encoding regions in the genome for properties associated with the proteins of interest, as MW and potential for being surface exposed proteins. The potential candidate genes presented the following attributes:

- CDSs encoding proteins with a MW higher than 50 kDa. This criterion was based on the assumption that R3, Z1 and Z2 are high molecular weight proteins.
- CDSs with **predicted functional annotations** as membrane associated, surface associated or hypothetical proteins.
- CDSs encoding proteins **predicted as potential surface located or secreted**. This criterion was based on the knowledge of surface exposition of R3, Z1 and Z2 proteins. Proteins predicted to have TMH are potential TM, proteins retaining LPxTG or YSIRK signals are predicted to be covalently or transiently linked to the cell wall and proteins carrying signal peptides are features of secreted proteins or lipoproteins.

3.3. Analysis of the *sar5* gene in relation to the expression of the R3 surface protein

In order to clarify if R3 and R5 are identical proteins, two procedures were used. First, a variety of R3 positive reference and prototype GBS strains

were tested using the *sar5* PCR and secondly, we cloned the gene encoding the R5 surface protein (*sar5*) behind an inducible promotor on plasmid pET15. The resulting plasmid (pET15*sar5*) were introduced into *E.coli* BL21cell and the strains containing the plasmids were then tested for R3 expression by immunofluorescence.

3.3.1. Bacterial strains, growth and media

GBS strains used for the experiments are listed in table 3-1. Additionally to the GBS strains, *E. coli* DH5α cells (plasmidic DNA production cells (pDNA)) and *E.coli* BL21 cells (recombinant protein production cells) from Life technologies were used for cloning experiments.

GBS strains stored at -80°C were grown over night (ON) on blood agar plates. *E.coli* cells stored at -80°C were grown ON in Luria-Bertani (LB) broth, unless otherwise specified. *E. coli* bacteria were grown onto LB agar with the presence of 100 μg of ampicillin/ml or on LB agar plates containing IPTG (inducer) when this was needed. Incubations were performed at 37°C, ON.

3.3.2. Chromosomal DNA extraction from GBS strains

For nucleic acid extraction, one colony was picked from subculture on a blood agar plates and added to 300 µl of a lysis solution containing 273 µl of Tris-EDTA (TE) buffer, 15 µl of lysozyme (20 mg/ml), 6 µl of proteinase K (20mg/ml) and mutanolysin (10.000 U/ml). The mixture was incubated at 37°C and 65°C for 15 minutes each. DNA was purified using the Qiagen column from the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) and eluted in a volume of 50 µl.

3.3.3. Oligonucleotide primers and PCR amplifications

The primers set used in this work and their sequences are listed in Table 3-2. The primers were designed based on the published sequence of the *sar5* gene of GBS Compton R (EMBL accession number **AJ133114.1**). The first primer set (reported previously[24]) was used to detect the *sar5* gene in the prototype and reference GBS genomes. Primer sets two and three were designed using the program Clone Manager 9 (Sci-Ed Software, http://www.scied.com/pr_cmbas.htm), to amplify the full-length *sar5* gene by PCR. These primers included restriction endonuclease recognition sites to enable subsequent cloning into a modified expression vector.

Table 3-2. Primers sets used through the experiments

Primer set	Primer name	Primer sequence 5'-3'
1	*Sar5* Forward	CGTAAATTTTCGGTTGGAATAGC
	Sar5 Reverse	GACGAACCACCGTTGTTTCAG
2	R5 F *XhoI*	GTCAACTCGAGATGTTTCGTAAATATAATTTTG
	R5 R *BamHI*	GAGCTGGATCCATCTATGATGTGATTATTAAC
3	R5 trunc F *XhoI*	GTCAACTCGAGACTCCAACAGGTG
	R5 R *BamHI*	GAGCTGGATCCATCTATGATGTGATTATTAAC

Amplification was carried out in a final volume of 25 µl containing the Tag Polymerase Promega® buffer 1X (10 mM Tris-HCl, pH 8.3; 50 mM KCl; 0.1% Triton® X-100); 1.5 mM de $MgCl_2$; 200 µM from each dinucleotide (dATP, dCTP, dGTP, and dTTP (Promega®), 0.4 µM of each primer; 1.5 units from the Taq Polymerase Promega® enzyme, and 1 µl from the DNA sample. The amplification conditions used are listed in table 3-3.

Table 3-3. PCR cycling conditions used through the experiments

Primer Set	Amplification phases					No of cycles
	Initial denaturation	Denaturation	Annealing	Extension	Final extension	
1	96°C/5min	95°C/1min	58°/45sec	72°C/10min	10°C/∞	36
2	96°C/5min	65°C/1min	50°C/45sec	72°C/10min	10°C/∞	36
3	96°C/5min	65°C/1min	53°C/45sec	72°C/3min	10°C/∞	36

The amplification products were visualized through electrophoresis in 1.0% agarose gels stained with ethidium bromide. To estimate the size of the amplified product, two molecular weight patterns were used: 1 kb DNA Ladder with a reading range between 10,000 and 250 bp, and a molecular weight pattern 2-Log DNA Ladder with fragments ranging from 100 bp to 10 kb, both from New England BioLabs®inc.

3.3.4. Identification and cloning of the Sar5 gene

Amplified fragments were cloned behind an inducible promoter on plasmid pET15b (Novagen (EMD Millipore)) and introduced into the pDNA production cells *E.coli* DH5α. It was done by ligating the *Ncol*/*BamHI* fragment of *Sar5* gene into *Ncol*/*BamHI* pET-15b and transforming the *E.coli* DH5α competent cells. Then, the plasmid carrying the *sar5* gene (pET15*sar5*) were introduced into *E.coli* BL21 cells and the strain containing the plasmids were streaked onto agar plated containing IPTG (inducer). Description is presented as follows:

Following PCR amplification (using primer set 2 and 3), the full-length products were digested with the restriction enzymes *Xhol* and *BamHI*. Digested products were purified using the QIAquick PCR Purification Kit of

QIAGEN and cloned into the vector pET15b which carries an N-terminal His•Tag® sequence followed by a thrombin site and multiple cloning sites. Plasmid DNA was prepared using the PureYield™ Miniprep System from Promega®. The vector was previously digested with the same restriction enzymes used to digest the PCR products, allowing insertion of the *sar5* gene into the vector. The resulting recombinant plasmid (pET15*sar5*) were used to transform the *E. coli* DH5α competent cells by the heat shock transformation method. Briefly, 10µl of the PCR product was mixed with 2µl of (10x) T4 ligase buffer, 4 µl of pET15b vector (40ng/µl), 2 µl of T4 DNA ligase and 3µl of deionised water. The mixtures were incubated ON to 16°C. Transformations of *E. coli* DH5α cells were made by mixing 5µl of the ligation reaction mixture with 50µl of competent cells on ice (20 min), heat shocking the cells at 42°C (30 sec) and cooling on ice (2 min). Then, LB medium (1ml) was added and the mixture was incubated at 37°C for two hours. Transformed cell cultures were plated on LB agar plates containing ampicillin (100µg/ml) and incubated at 37 °C ON.

To confirm that the pDNA producers contained the *sar5* gene, colony growth on LB agar plates containing ampicillin (100µg/ml) was grown in 2ml of LB ON. Plasmids were purified by the PureYield™ Miniprep System from Promega® and digested with the same restriction enzymes. The restrictions were checked to fragments of correct molecular weight through electrophoresis in 1.0% agarose gels stained with ethidium bromide. The transformation was also confirmed trough PCR using the plasmid DNA and the primers reverse *sar5* and R5 trunc F *Xhol*. Untransformed *E.coli* DH5α-cells were tested as control.

Then, the resulting plasmids produced by the *E.coli* DH5α-cells (pET15*sar5*) were introduced into *E.coli* BL21 (recombinant protein production cells) and

the strain containing the plasmid was streaked onto agar plated containing IPTG (inducer).

3.3.5. Test of sar5 transformants for R3 expression

E. coli LB21 *sar5* transformants were tested for R3 surface protein expression by immunofluorescence using rabbit polyclonal antibodies (PAs) raised against the R3 reference strain GBS Prague 25/60 (ATCC9828) previously shown to contain antibodies against R3. Slides for immunofluorescence testing were prepared from *E. coli* LB21 culture on LB medium and the testing was performed essentially as described in [60]. The antiserum was used diluted 1:50 and 1:200, respectively, and R3 expression was tested by using fluorescent anti-rabbit IgG antibodies.

4 RESULTS

4.1. GBS strain GMFR293 genome sequencing and assembly

The data obtained from 454 pyrosequencing and optical mapping allowed assembly of GMFR293 into a complete genome. First, sequencing of GMFR293 by 454 pyrosequencing resulted in a total of 159,529 reads of 59,021,738 Kb in size, with about 28 fold coverage of the genome. In total, 48 contigs with an average size of 55,582 bp and a median contig size (N50 value) of 133,175 were produced when the reads were assembled, using the assembly software Newbler.

By optical mapping of genomic DNA of strain GMFR293 restriction cut by *Ncol*, 196 fragments and 195 restriction cuts were identified. By this method the total size of the genome was estimated to 2,029,591 bp, with fragments in the range from 1,723 bp to 79,393 bp.

By aligning an *in silico* restriction map of the contigs from the assembly of sequencing reads using restriction cut sites similar to that of *Ncol* to the optical map, 78 % of the genome sequence assembly (11 contigs) was covered while 37 of the contigs did not align with the optical map. All contigs were then aligned with the most similar reference genome of strain GBS2603 V/R (Figure **4.1**). This allowed closure of the gaps and completion

of the genome. The final GMFR293 *in silico* map composed of 235 restriction fragments had a size of 2,033,090 bp.

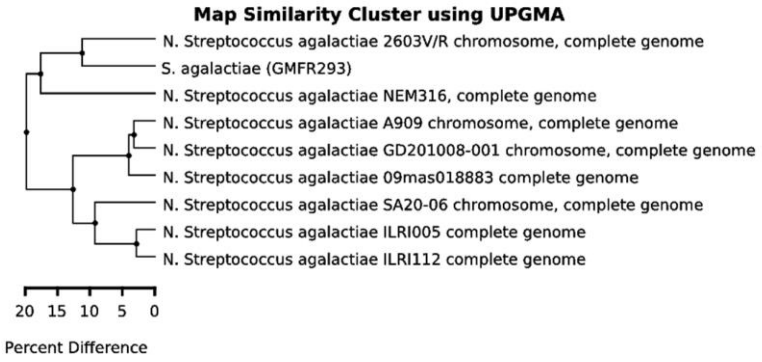


Figure 4-1. Phylogenetic tree showing similarity at genome level between GBS strain GMFR293 and other complete GBS genomes, including the most similar genome of reference strain 2603 V/R.

4.1.1. Assessment of GMFR293 genome assembly

To assess accuracy of the GMFR293 genome assembly, the optical restriction map and the generated *in silico* restriction map of assembled contigs were compared. A total of 67 fragments were classified as uncertain regions due to differences between the optical restriction map and the *in silico* restriction pattern (Appendix A contains the full list of these 67 uncertain regions). Among these were 27 fragments which were present in the *in silico* map but not in the optical map, and 40 fragments which were shared between the maps, but where there were differences in the fragment size. Relative sizing error was calculated (Figure 4-2), and for nine fragments the error was higher than 10% (Figure 4-3).

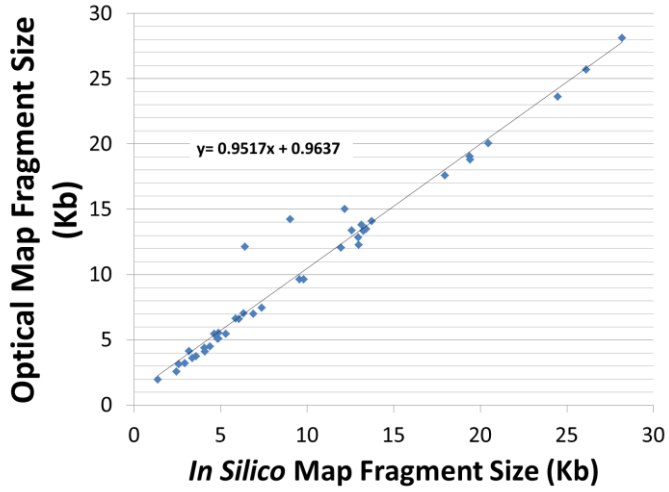


Figure 4-2. Plots of optical map fragment sizes versus *in silico* restriction map fragment sizes of 40 uncertain regions.

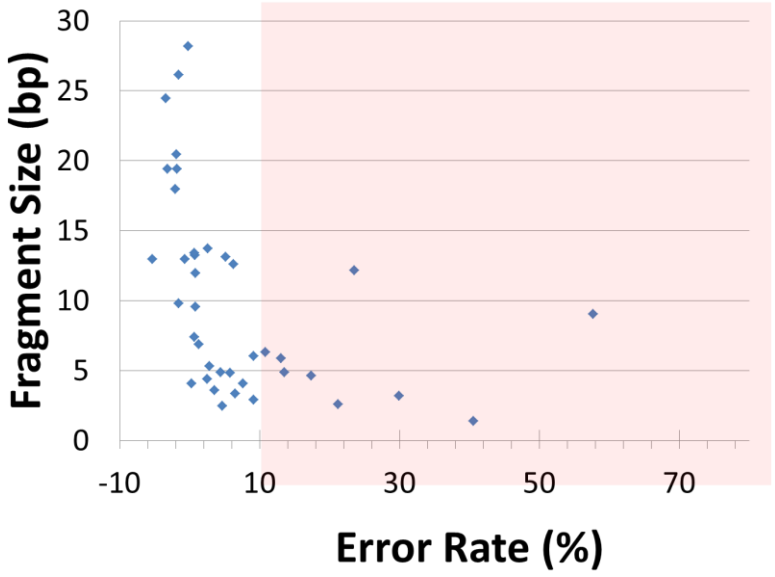


Figure 4-3. Relative fragment size error rate versus *in silico* restriction map sizes of 40 fragments from identified uncertain regions.

When the uncertain regions of the GMFR293 *in silico* map were compared with the *in silico* restriction maps of the three GBS references genomes of strain 2603V/R, A909 and NEM316, 16 fragments were found to have

different restriction patterns and 51 had identical restriction pattern with at least one of the reference genomes. Five of the 16 fragments had a different size, and 11 were unique for the GMFR293 *in silico* map.

Finally, after analysis of all the parameters evaluated, 29 fragments in the assembled genome, corresponding to 8.7% of the GMFR293 total genome, with fragment sizes between one and 28,186 bp (Figure 4-4) were still considered uncertain which should therefore preferably be subjected to experimental verification (Appendix B contains the full list of uncertain regions selected to verification), in order to confirm the accuracy of the finished sequence.

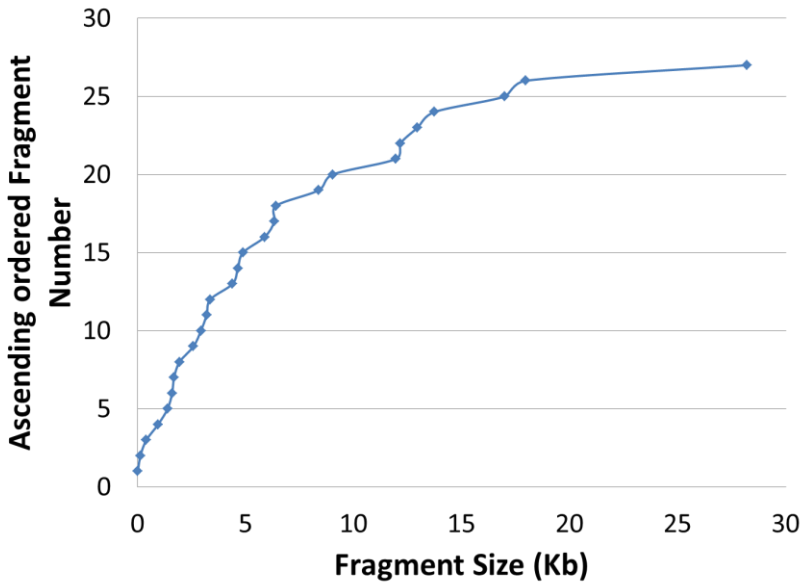


Figure 4-4 Ascending ordered fragment number versus fragment size in Kb of places in the genome assembly that must be verified experimentally to obtain a finished genome.

4.2. General features of the GMFR 293 and CMFR30 genomes

The complete GMFR293 genome without verification of the remaining uncertain regions mentioned above, consisted of a single circular chromosome of 2,037,090 bp, with a G+C content of 35.5%, containing 2,023 coding sequences (CDSs) with putative predicted protein encoded genes. The genome contained 95 RNAs composed by 74 tRNA, 14 rRNAs, and 7 sRNAs (see figure 4-5).

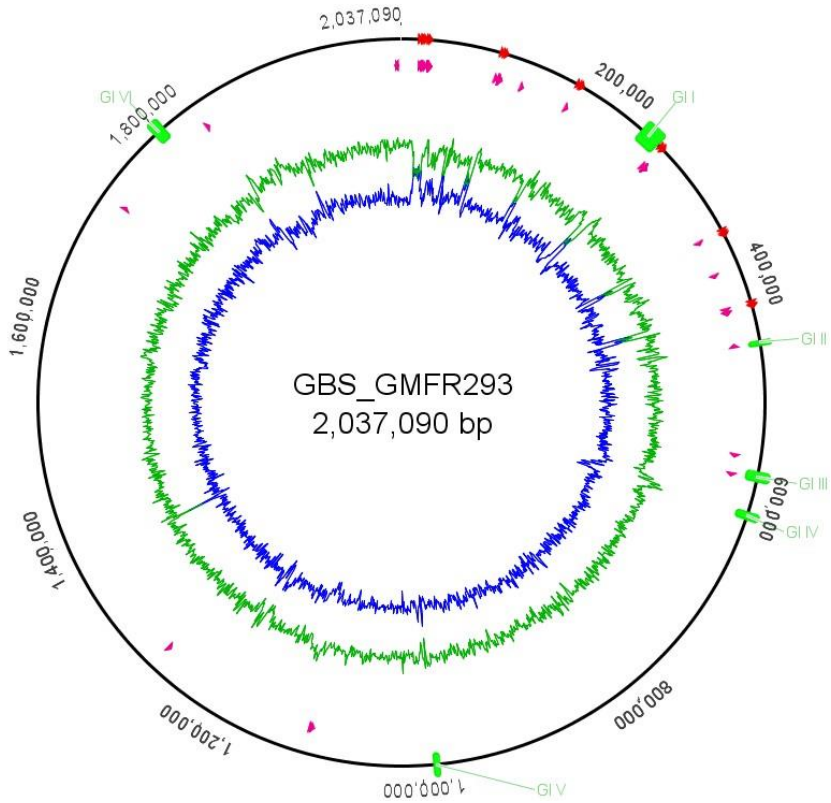


Figure 4-5 Circular representation of the genome of GBS strain GMFR293, analysed by *Geneious version 7.1*.[108] Arrows: Pink, tRNAs; Red, rRNAs; Green, Genomic islands. Inner AT graph (in green) and GC content (in blue).

The complete CMFR30 genome consisted of a single circular chromosome of 2,062,772 bp, with a G+C content of 35.4%. There were 2,060 coding sequences (CDs) with putative predicted protein encoding genes. The genome contained 88 RNAs composed by: 70 tRNA, 12 rRNAs, and 6 sRNAs (see figure 4-6). The general features of both the sequenced GBS genomes are presented in table 4-1.

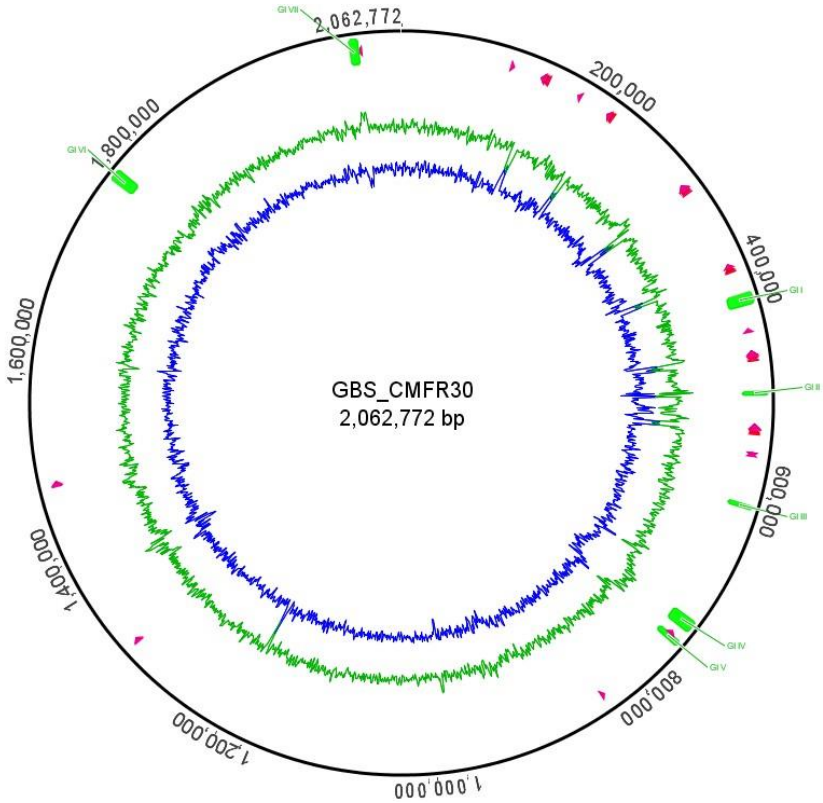


Figure 4-6 Circular representation of the CMFR30 genome, analysed by *Geneious version 7.1* [108]. Arrows: Pink, tRNAs; Red, rRNAs; Green, Genetic islands. Inner AT graph (in green) and GC content (in blue).

Table 4-1. General features of the GMFR293 and CMFR30 genomes.

Strain	Replicon	Size bp	GC%	CDs	tRNA	rRNA
GMFR293	Chrom	2.037.090	35.5	2023	74	14
CMFR30	Chrom	2.062.772	35.4	2060	70	12

In order to obtain more complete information about the gene coding sequences in the genomes, functional annotations were grouped into COG functional categories and FIGfams-subsystems (RAST).

The gene distribution of the two GBS genomes according to their COG functional categories is presented in table 4-2, and the statistics from the annotation process through COG functional categories is presented in table 4-3.

Table 4-2. Number of genes associated with the general COG functional categories in strain GMFR293 and CMFR30.

Code	Description	GMFR293		CMFR30	
		Value	%	Value	%
C	Energy production and conversion	56	2.76	55	2.71
D	Cell cycle control, cell division, chromosome partitioning	25	1.23	24	1.18
E	Amino acid transport and metabolism	152	7.51	148	7.31
F	Nucleotide transport and metabolism	83	4.10	84	4.15
G	Carbohydrate transport and metabolism	167	8.25	186	9.19
H	Coenzyme transport and metabolism	55	2.71	56	2.76
I	Lipid transport and metabolism	52	2.57	50	2.47
J	Translation, ribosomal structure and biogenesis	149	7.36	152	7.51
K	Transcription	134	6.62	137	6.77
L	Replication, recombination and repair	114	5.63	123	6.08
M	Cell wall/membrane/envelope biogenesis	106	5.23	107	5.28
N	Cell motility	9	0.44	6	0.29
O	Posttranslational modification, protein turnover, chaperones	53	2.61	59	2.91
P	Inorganic ion transport and metabolism	102	5.04	109	5.38
Q	Secondary metabolites biosynthesis, transport and catabolism	23	1.13	20	0.98
R	General function prediction only	213	10.52	215	10.62
S	Function unknown	167	8.25	172	8.50
T	Signal transduction mechanisms	76	3.75	83	4.10
U	Intracellular trafficking, secretion, and vesicular transport	26	1.28	23	1.13
V	Defence mechanisms	44	2.17	47	2.32
-	Not in COGs	217	10.72	208	10.28

Table 4-3. Functional genome annotations through COGs of the GBS strains GMFR293 and CMFR30

Strain	Total number of genes	Genes with assigned function	Function unknown	Not in COGs	Assigned to COGs
GMFR293	2023	1639 (81%)	167 (8.2%)	217 (10.7%)	1806 (89.3%)
CMFR30	2064	1684 (81.6%)	172 (8.3%)	208 (10%)	1856 (89.9%)

Almost one third of the genes in each of the two GBS genomes were predicted as hypothetical proteins when they were annotated by RAST. The statistical values of this annotation process are presented in table 4-4, including the description of the steps that the RAST server implemented to automatically produce the two classes of asserted gene functions: subsystem-based assertions are based on recognition of functional variants of subsystems (Collection of functional roles jointly involved in a biological process) while non-subsystem based assertions are filled in using more common approaches based on integration of evidence from a number of tools.

In the genome of GMFR293 400 CDSs were annotated as hypothetical proteins; 16 CDSs as surface-associated, 65 CDSs as membrane associated and 11 CDSs as cell wall associated by annotation with RAST,. In comparison, in the CMFR30 genome 390 CDSs were annotated as hypothetical proteins, 18 CDSs as cell surface-associated 69 CDSs as membrane associated and 15 CDSs as cell wall associated proteins. The latter were the first CDSs evaluated as potential CG for R3, Z1 and Z2 surface exposed protein.

Table 4-4. Statistics of the annotation process through RAST pipeline annotation.

Strain	Total number of CDSs	Coverage	Annotation	Number of Hypothetical Proteins	Number of non-Hypothetical proteins
GMFR293	2023	In subsystems 1102 (55%)	Hypothetical 46 (4.2%)	400 (19.8 %)	1623 (80.2%)
			Non hypothetical 1056 (95.8%)		
		Non-in subsystems 921 (45%)	Hypothetical 354 (38.43%)		
			Non hypothetical 567 (61.6%)		
CMFR30	2064	In subsystems 1139 (56%)	Hypothetical 48 (4.2%)	390 (18.9%)	1674 (81.1%)
			Non hypothetical 1091 (95.8%)		
		Non-in subsystems 925 (44.8%)	Hypothetical 342 (37%)		
			Non hypothetical 583 (63%)		

In addition, molecular weights (MW) were calculated for all CDSs that were present in both genomes. This was done in an attempt to identify the R3, Z1, and Z2 by their molecular weight, which had been estimated to around 140 kDa for R3, 250 kDa for Z1 and 135 kDa for Z2 in a previous study.

Most predicted proteins of the GMFR293 and GBS CMFR30 genomes were in the range of 4.1 ±1 to 172.3 kDa. From the 2,024 CDSs that constituted the complete GMFR293 genome and the 2,064 CDSs that constituted the genome CMFR30, 29 and 30 CDSs, respectively, had molecular weight of more than 100 kDa. However, there were no proteins with a molecular weight as high as that predicted for Z1 (250 kDa). Based on this result, the range of MW used as filter to target CDSs for CG was adjusted to higher

than 50 kDa. Based on this filter, 321 CDSs of GMFR293 and 242 CDSs of CMFR30 were selected as candidate genes for for R3, Z1 and Z2.

4.2.1. Genomic islands (GIs)

Strains GBS GMFR293 and CMFR30 possess several virulence factors, GIs, transposons and insertion sequence (IS) elements, distributed over their genomes. It is well known that genes contributing to pathogenesis frequently are located in such genomic Islands.

Strain GMFR293 contained six putative genomic islands (see figure 4-6) incorporating 91 predicted genes, many of which were mobile elements. The genomic islands were composed of 10 to 25 genes with molecular weights between 4.7 kDa and 93 kDa. 43 of these genes were predicted to encode hypothetical proteins and 25 were predicted to be transmembrane proteins. Nine of the 25 genes were hypothetical proteins and predicted to be transmembrane proteins. We also checked if the gene *sar5* encoding the R5 surface protein was part of a genomic island, but it was not present in any of the predicted genomic island in GMFR293.

Isolate CMFR30 contained seven putative genomic island (see figure 4-7) incorporating 79 predicted genes. The islands were composed of 6-22 genes with molecular weight in the range of 4.4 kDa to 128 kDa. 29 genes were predicted to encode hypothetical proteins and 19 were predicted to be transmembrane. 13 of the predicted CDSs were classified both as hypothetical and transmembrane proteins.

4.2.2. Known surface proteins in GBS GMFR293 and GBS CMFR30

Surface proteins in Gram-positive bacteria are frequently implicated in virulence. In GBS, numerous genes have been identified as genes encoding surface proteins. These proteins together with secreted products are identified as potential virulence factors.

After annotation of the GMFR293 and CMFR30 genomes, it was possible to identify previously sequenced GBS surface proteins. Some of the known surface proteins found in both GBS genomes were: C5a peptidase, cold shock protein CspA, surface protein Rib, sortase A (one in CMFR30 and three in GMFR293), fibronectin/fibrinogen-binding protein, hyaluronate lyase precursor, laminin-binding surface protein, group B streptococcal surface immunogenic protein and the CAMP factor.

4.2.3. Prediction of surface exposed proteins

The prediction of proteins carrying signature motifs to Gram positive surface proteins is important because the carriage of signal peptides is involved in the protein secretion and surface display in such bacteria. Therefore an attempt was done to predict potential subcellular locations of the proteins encoded by the GMFR293 and CMFR30 genomes. The aim was to identify CDSs encoding potential surface exposed and secreted proteins. The results for both GBS strains are presented in the table 4-5.

Table 4-5 Results of the prediction of transmembrane helix (TMH) and signature motifs in the GMFR293 and CMFR30 genomes.

Parameter Predicted	GMFR293		CMFR30		Meaning
	Total CDSs	CDSs >50kDa	Total CDSs	CDSs >50 kDa	
TMH	536 (26.5%)	113 (5.58%)	545 (26.40%)	70 (3.39%)	Characteristic of membrane proteins
Signal Peptides	114 (5.36%)	31 (1.53%)	114 (5.52%)	27 (1.30%)	Found in proteins that are secreted, retained or proteins that cross the membrane only once (single pass).
YSIRK Signal	7 (0.34%)	4 (0.19%)	7 (0.33%)	6 (0.29%)	Found in protein with potential to be secreted into the cell wall.
LPxTG Signal	58 (2.86%)	13 (0.64%)	64 (3.10%)	26 (1.26%)	
Lipoproteins	111 (5.48 %)	14 (0.69%)	108 (5.23%)	16 (0.77%)	Lipoproteins

CDs predicted to encode TM, and/or proteins carrying signals peptides were selected, and included in the final list of CG for R3, Z1 and Z2 proteins (Appendix D contains the full list of CDSs) as well as for the Z1 protein (Appendix E contains the full list of CDSs).

4.3. Comparison of the GMFR293 and CMFR30 genomes against reference GBS genomes

The pan-genome is the entire gene repertoire in a selection of a strain or a species, representing the sum of the above mentioned core genome and the dispensable genome. In previous studies it was found that strain GMFR293

expressed the proteins R3, Z1 and Z2 and contained the *sar5 gene*; strain CMFR30 expressed Z1 but was *sar5* negative by PCR analysis, and the GBS reference strains A909, NEM316, 2603 V/R and 515 were negative for the expression of all the three proteins and were *sar5* negative by PCR analysis. In this study we did a comparative analysis based on protein sequence homology. CDSs of the GMFR293 and CMFR30 genomes were compared against the five complete GBS genomes A909, NEM316, 2603 V/R, 515 and CMFR30 and/or GMFR293. The aim of the comparison was to identify candidate genes for the R3, Z1 and Z2 proteins by analysis of the occurrence pattern (absence/presence) and the grade of similarity between the genes.

In general, the number of genomes that are included in a comparison influences on the distribution of CDSs between the core and dispensable genome of each strain, and the number of genes which are unique to each genome. (See figure 4-7).

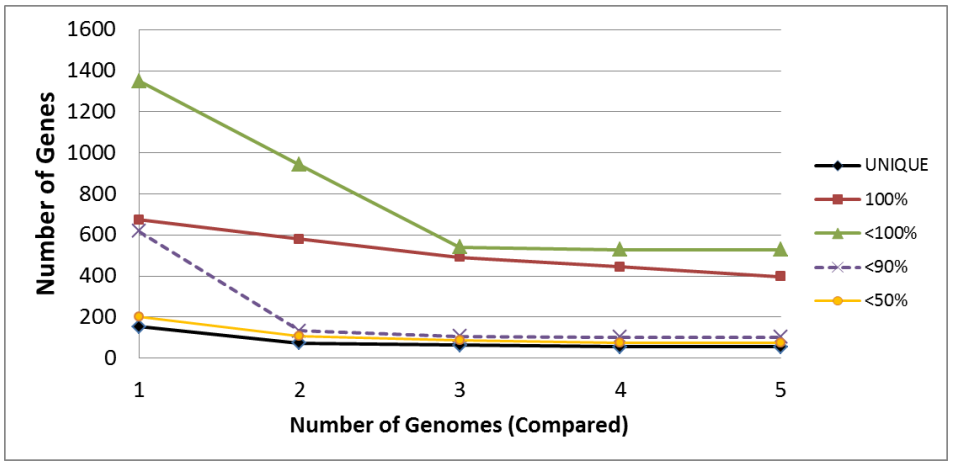


Figure 4-7 Comparative analysis of CDSs of the GMR293 genome with five GBS reference genomes. Colours indicate the number of genes that were present in all or just a subset of the genes, depending on how many genomes that were compared

CDSs highly conserved in the GBS genomes which were compared, CDSs shared between the genomes and strain specific genes in GBS GMFR293, and CMFR30 could be identified after the sequence-based comparison (see figure 4-8 for an illustration of the comparison of GMFR293).

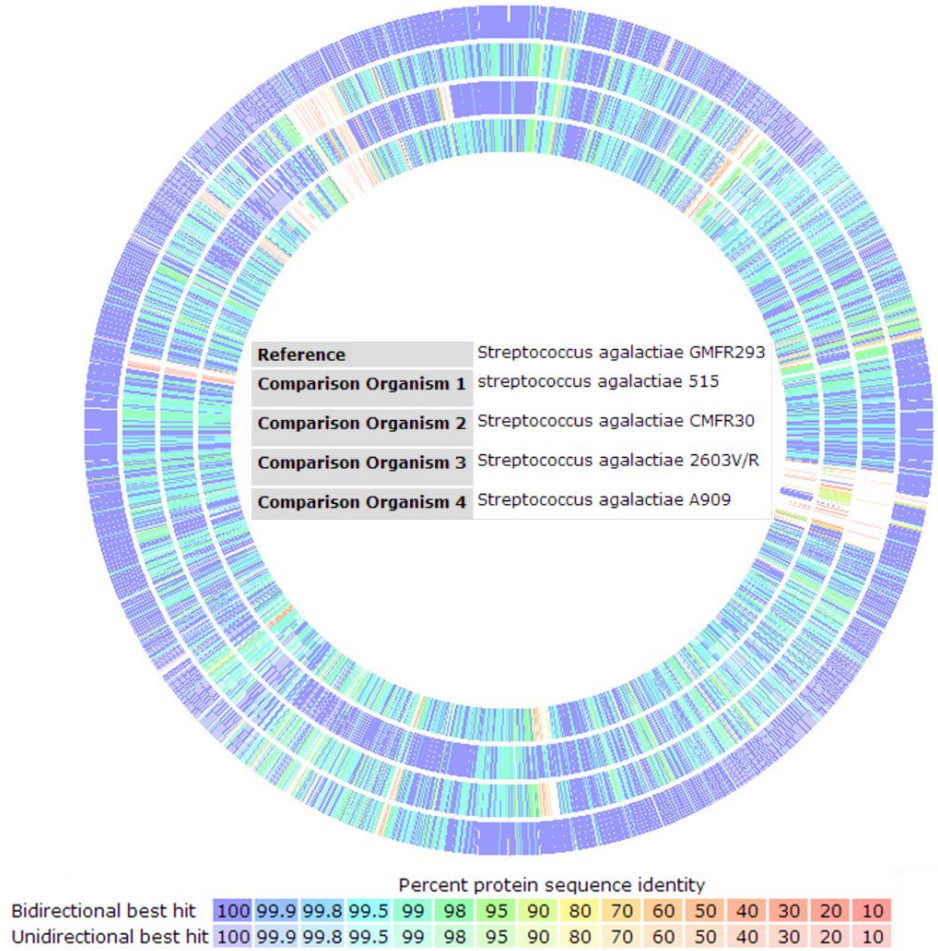


Figure 4-8 Circular map with color-coded table showing sequence identity of four reference GBS genomes compared to GMFR293, using the RAST sequence based comparison tool. The colours represent changes in conservation relative to the reference genome GMFR293. Colours going from blue representing highest protein sequence similarity to red representing the lowest. Each gene is marked as being unique, a unidirectional best hit or a bidirectional best hit in comparison to the reference genome. The order of the circles from the inner to the outer is as follow: A909, 2603V/R, CMFR30 and 515.

4.3.1. GBS GMFR293 genome comparison

Compared to the CMFR30, 515, A909, 2603 V/R and NEM316 GBS genomes, 14 genes were identified as strain specific genes for GMFR293 (see table 4-6), and the proteins encoded by these genes were estimated to be in the range of 4.25 kDa to 35.09 kDa. Most of the CDSs were annotated as hypothetical protein encoding genes, four were predicted to be part of the genomic island III, and none of the CDs in this group of strain specific genes were classified as transmembrane or carrier of signal peptides.

Table 4-6 GMFR293 strain specific genes.

CDS	Start	End	Annotation	MW (kDa)	Topology	Genomic Island (GI)
303	320618	320505	hypothetical protein	4.25	outside	
559	569880	570050	hypothetical protein	6.33	inside	
560	570227	570667	Phage protein	17.24	outside	
562	571695	572537	DNA replication protein	31.89	outside	III
563	572537	572683	hypothetical protein	5.75	outside	III
564	572673	572948	hypothetical protein	10.86	outside	III
581	580027	580935	Phage protein	35.09	outside	III
892	906624	906755	hypothetical protein	5.25	outside	
1189	1206597	1206220	hypothetical protein	14.82	outside	
1384	1411217	1411354	hypothetical protein	5.32	outside	
1671	1680553	1680675	hypothetical protein	4.75	outside	
1744	1761221	1761352	hypothetical protein	4.98	outside	
1874	1882814	1882647	hypothetical protein	6.59	outside	
1891	1894488	1894631	hypothetical protein	5.59	outside	

A total of 180 CDSs from GMFR293 had a similarity percentage less than 100% when compared with CDSs in the other five GBS genomes. There were 57 CDSs which were predicted as transmembrane and annotated as hypothetical proteins, 43 CDSs had a molecular weight higher than 50 kDa, and these were therefore selected as candidate CDSs for R3 and Z2, especially those predicted to be potential surface exposed from the previous analyses. The features presented by the members of this group are represented in the table 4-7.

Table 4-7 Features of the target CDs for R3 and Z2 CGs obtained through the GMFR293 CDs comparative analysis.

CDSs Features	No of CDSs	Signal peptide	Lipo-proteins	YSIRK signal	LPxTG signal
CDSs with molecular weight higher than 50 kDa	43	5	2	3	2
CDSs with molecular weight lower than 50 kDa	137	7	6	4	1
TOTAL of CDSs with similarity less than 100%	180	12	8	7	3

4.3.2. CMFR30 genome comparison

The comparative analysis of the CMFR30 CDSs against the GMFR293, 515, A909, 2603 V/R and NEM316 GBS genomes was done in order to identify candidate CGs for the Z1 protein, especially searching for CDSs more similar with CDSs in GMFR293, and absent or less similar with CDSs in the other genomes.

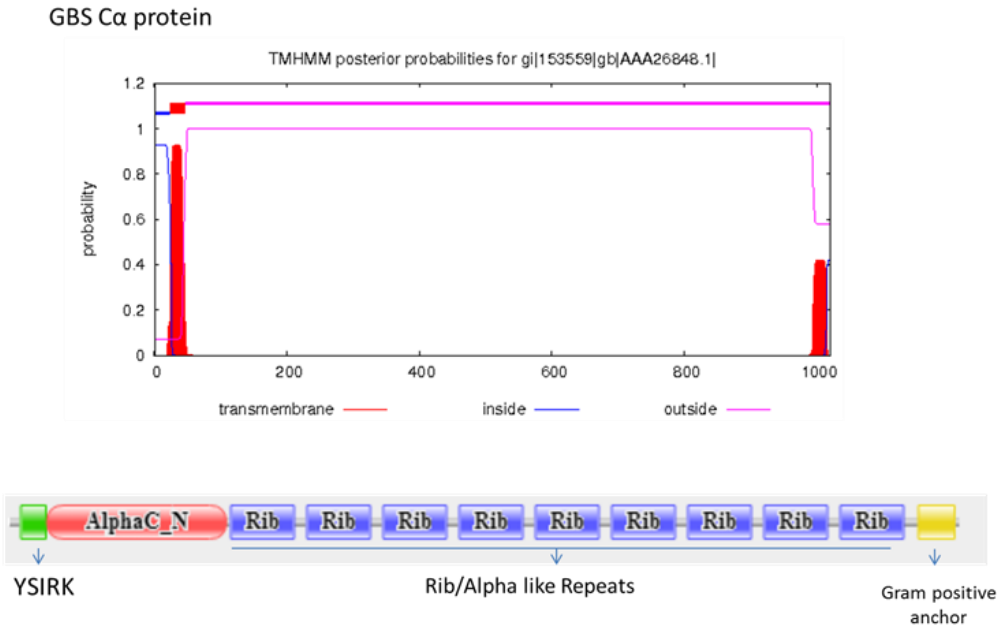
After the comparison with the other GBS genomes, 48 CDSs were identified as CMFR30 strain specific based on absence or similarity to other genomes

of less than 50%. The molecular weight of the encoded proteins in this group was in the range of 3.8 kDa to 63.2 kDa. 25 CDs were annotated as hypothetical proteins and six were predicted to belong to a genomic island (GI) (Appendix C contains the full list of the CMFR30 strain specific CDSs).

All the CDSs that were 100% identical between the CMFR30 and GMFR293 genomes and a similarity less than 100% with the reference genomes were selected for further analyses. Using that criterion, eight genes were identified with MWs in the range of 4.1 kDa to 80.01 kDa. However, only one of the CDSs had a molecular weight higher than 50 kDa.

4.4. R3, Z1 and Z2 candidate genes

The *in silico* approach allowed the identification of 32 CDSs in the CMFR30 genome with potential to be CGs for the Z1 protein. 26 of these were annotated to have a putative function and six as hypothetical proteins or proteins of unknown function. Many of them exhibited features similar to GBS surface proteins previously identified (see figures 4-9 and 4-10 for some examples). 14 CDSs were found sharing a similar organizational pattern: a N-terminal signal peptide and a C-terminal LPxTG motif. Five of them carried an YSIRK motif which is positioned within the signal peptide at the start of the transmembrane domain and six CDSs were predicted as carriers of the consensus sequence of lipoprotein precursors.



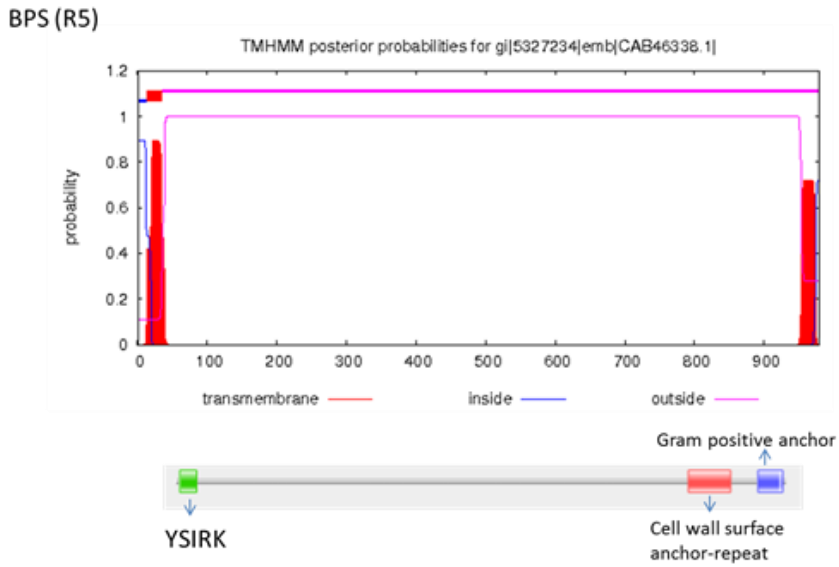
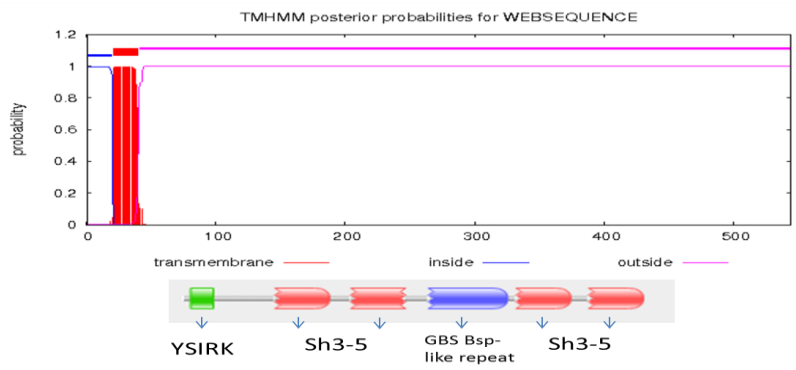
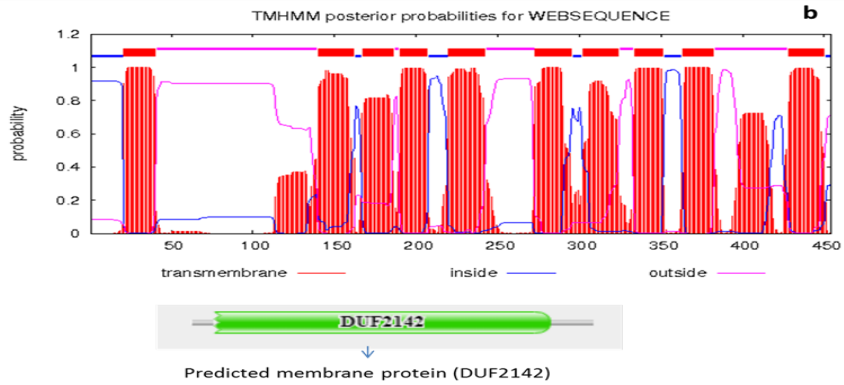


Figure 4-9 Graphs showing the prediction of transmembrane regions (TMHMM) and prediction of the domain architecture (Pfam) of the GBS proteins Cα (GenBank: M97256.1) and R5 (BPS) (GenBank: CAB46338.1).

CDS_1242_GMFR293



CDS_1305_GMFR293



CDS_159_GBS _CMFR30

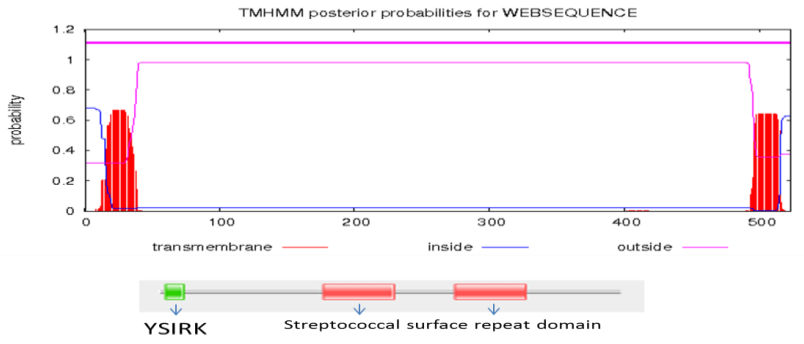


Figure 4-10 Graphs showing the prediction of TM regions (TMHMM) and the domain architecture (Pfam) from two CGs (CDS-1242 and CDS-1305) obtained from the GMFR293 sequence analysis and one of the selected CGs for the Z1 surface protein (CDS-159) obtained from the CMFR30 sequence analysis.

51 CDSs were identified with potential to be CGs for the R3 and Z1 and Z2 proteins in the GMFR293 genome. 36 of these were annotated to have a putative function and 15 as hypothetical proteins or proteins of unknown function. Similar to the CGs in the CMFR30 genome, many of them exhibited features similar to GBS surface proteins previously identified (see figures 4-9 and 4-10 for some examples). Four CDSs had a N-terminal signal peptide and a C-terminal LPxTG motif and one of them was predicted to carry additionally an YSIRK signal. Five CGs had an YSIRK signal motif and were predicted to carry a signal peptide. Fourteen CDSs were predicted as carriers of the consensus sequence of lipoprotein precursors. N-terminal signal peptides and a C-terminal LPxTGs are characteristic of cell wall associated proteins. Some proteins have in addition, an YSIRK signal positioned within the signal peptide at the start of the transmembrane domain. Lipoproteins are considered to be directly anchored to the cytoplasmic membrane.

4.5. Sar5 as candidate gene for the R3 surface display protein

Reference and prototype GBS strains that in a previous study expressed one or more of the surface exposed proteins in question (R3, Z1, and Z2) by immunofluorescence were tested by PCR for the presence of the *sar5* gene.

In *sar5* positive samples a PCR product of 417 bp, as expected for this gene was detected by gel electrophoresis (see figure 4-11). All the strains previously serotyped as R3 positive were positive for the R5 PCR, including

the GBS strain 2603V/R which had been reported previously as *sar5* negative[24] (See table 4-7).

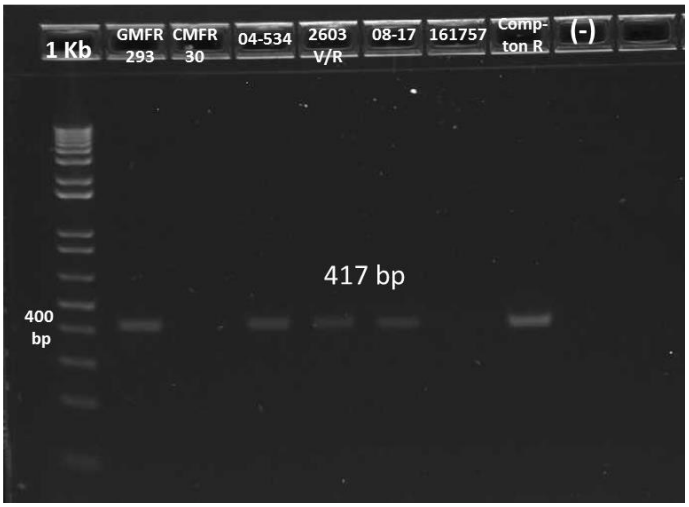


Figure 4-11 Electrophoresis gel obtained from *sar5* PCR. Ladder 1Kb.

Table 4-7. PCR results for the *sar5* gene.

GBS Strain	Serotype	*Sar*5 PCR
GMFR293	V/R3,Rib, Z1,Z2	+
CMFR30	Ib/Cα,Cβ,Z1	-
04-534	XI/ Cα, Cβ, R3, Z1, Z2	+
2603V/R	V/R4/ Rib	+
08-17	V/R3, Z1, Z2	+
161757	V/alp3	-
*Compton R (NCTC 9828/Prage 2560)	NT/R3, R4, R5	+

*Strain used as PCR positive control.

In addition, the *sar5* appeared as one of the CG for R3 (CDS-1223) in the *in silico* analysis of the CDSs from the GMFR293 genome. This CG was

annotated as a hypothetical protein of 105.62 kDa, predicted as TM, and carrier of LPxTG motif and of a 39-residue signal peptide.

In the genome comparison the *sar5* showed similarities less than 27% with CDSs in four of the five genomes used for the comparison. However, it was 100% similar with a protein from GBS 2603V/R. This is in agreement with the result of 100% similarity obtained between protein alignment of the CDS-1223 from GMFR293 with hypothetical proteins of 2603V/R, and 95% with the BPS protein (same R5) from GBS strain Compton R.

Since our results from strain 2603V/R were different regarding *sar5* from those published previously, we also retested this strain for R3 expression by immunofluorescence. However, the result obtained was unclear due to weak fluorescence signals that appeared in just few of the bacterial cells tested, while there was not fluorescence from the majority of cells. Thereafter, lack of R3 expression was further confirmed by Western blotting using polyclonal anti-R3 antibodies. Nevertheless, there is a possibility that strain 2603 V/R tested negative for R3 expression because of gene expression failure, which is known to occur in GBS[10]. To further clarify the relationship between R3 and R5 surface display proteins, we cloned the gene encoding the R5 surface protein (*sar5*) behind an inducible promoter on plasmids pET15. The resulting plasmid (pET15*sar5*) was introduced into *E. coli* BL21 cell and the strains containing the plasmid we streaked on agar plates containing the IPTG inducer. However, the *E. coli* BL21 cells transformants were negative when tested for R3 expression by immunofluorescence microscopy.

5 DISCUSION

The elucidation of the genes encoding the proteins R3, Z1 and Z2 is important since they could have a potential role in GBS serotyping and as vaccine candidates. In this study, we identified candidate genes (CGs) for the R3, Z1 and Z2 GBS surface exposed proteins, through the sequencing of the genome of the strains GMFR293 known to express R3, Z1 and Z2, and CMFR30 known to express the Z1 proteins, followed by the sequence analysis of their genomes by the use of *in silico* tools.

First, we used two different NGS platforms to obtain two complete GBS genome sequences. Strain GMFR293 was sequenced through 454 pyrosequencing, and strain CMFR30 was sequenced by Pacific Biosciences (PacBio) technology. To obtain a complete genome for strain GMFR293, the genome assembly was assisted by optical mapping and alignments to its closest reference genome. Together these approaches led to a complete draft genome, although with few regions that need experimental verification.

It is known that NGS technologies are developing very rapidly in terms of sequence output and cost reduction, which allows that draft genome sequences can be obtained easily and at low cost[109]. However, within these NGS technologies, each platform presents their respective advantages and disadvantages. The PacBio platform has been reported to have benefits like the highest N50, 99.99 % accuracy, and to produce fewest contigs. In addition, it has a relatively low cost per run, which may benefit studies that require only few samples to be sequenced. In terms of systemic error, PacBio has high error rates, but through the use of circular consensus reads, and

because errors are randomly distributed, the error rates are strongly reduced. In contrast, 454 have low error rates, but the errors are positionally increasing distally, with guanine-cytosine (GC) content, or with homopolymers [110; 111]. Our results showed that although 454 pyrosequencing might be a good choice to obtain a genome draft assembly, this technology led to many contigs, increasing both the likelihood of errors in the assembled genome and the effort needed to obtain a closed genome. In comparison, assembly of PacBio sequence reads led to a single contig, obtaining a complete CMFR30 genome. We conclude that less effort was needed to get a complete draft genome using PacBio compared with the 454 pyrosequencing method. However, both methods allowed obtaining the whole GBS genome sequences, and their availability allowed the identification of putative candidate genes coding for R3, Z1 and Z2 proteins.

An interesting question is to what degree analysis of draft genomes as compared to complete genomes, could result in introduction of errors from the sequencing and/or assembly process. Finished data of the highest quality is the most desirable state for a genome sequence, but draft quality sequence can provide a powerful resource for many genomic studies. In this study, we used the GMFR293 draft genome (closed complete genome but with few regions that need experimental verification), since we believe that they did not influence negatively on the identification of candidate genes, because even if there were errors in the sequencing and/or assembly process like lack of contig order, most genes were represented in the draft sequence. In addition, the GBS draft sequences also provided important information such as a comprehensive estimate of the number of genes in the GMFR293 and CMFR30 genomes and their classifications. The limitations for using a draft genome in genomic analysis has been observed more when using a draft sequence as a reference in comparative studies[112]. It could be one of the

reasons why the GBS genome comparison approach in this study was less successful for the identification of candidate genes than targeted search for specific characteristics of the proteins in question. However, the results from the genome comparisons increased our knowledge about special features of our sequences compared with other GBS genomes, and thereby this approach supported the identification of candidate genes.

Due to a few uncertain regions the GMFR293 draft genome needs to be experimentally verified for this genome can be considered a finished genome.

Second, this work describes the methodological model that we proposed for identification of CGs for the R3, Z1 and Z2 streptococcal surface-exposed proteins. The criteria used for selecting the CGs were based on previous knowledge about some characteristics of the R3, Z1 and Z2 proteins. The rationale for candidate gene selection was based on the following criteria: **(a)** CDSs encoding proteins with a MW higher than 50 kDa. This criterion was based in the assumption that R3, Z1 and Z2 are high molecular weight proteins, according to results obtained from Western blotting in previous studies. **(b)** CDSs with predicted functional annotations as membrane associated, surface associated or hypothetical proteins. This criterion was chosen from the knowledge that genes encoding the proteins of interest have not been previously identified. Therefore, these proteins can be encoded by CDSs without known function, or CDSs classified as membrane or surface associated proteins without any putative name or function. **(C)** CDSs encoding proteins predicted as potential surface located or secreted. This criterion was based on the knowledge of surface exposition of R3, Z1 and Z2 proteins from their detection by immunological test. This criterion allowed us to characterize each CDS in the GMFR293 and CMFR30 genomes,

according to its surface exposition potential. It also permitted us to identify those CDSs with common features to other GBS surface proteins previously identified, which were selected as potential CGs for the R3, Z1 and Z2 proteins. This methodology allowed us to select the group of potential candidate genes.

Computational approaches for sequence analysis obtained from the sequencing processes and their annotation were used to identify CGs. Automated annotation of the draft genomes of GMFR293 and CMFR30 provided valuable preliminary information about their genomes. However, annotations from both genomes should preferably be curated manually.

Approximately 19% of the complete genomes were assigned as hypothetical, uncharacterized or putative proteins. Some of the CDSs were specific for GMFR293 and CMFR30. It has been reported that hypothetical genes and genes with unknown function represent the vast majority of the dispensable GBS genome. Our findings are similar to results reported for other bacterial genomes including GBS, where around 20% of the predicted CDSs did not match any database entries, and an additional 15 to 20% CDSs were similar to genes with unknown function, many of them belonging to the dispensable genome [113; 114]. This shows that in spite of the increasing number of sequenced genomes, the assignment of function to a sequence remains in many cases a challenge, since this will require laboratory experiments which are complicated, time consuming and expensive. Several of the hypothetical proteins annotated in the GMFR293 and CMFR30 genomes belonged to strain specific gene clusters. They were identified through genome comparison used in this study to assist the selection of CGs.

Based on genome comparison analyses, we concluded that the genes encoding R3, Z1 and Z2 most probably did not belong to the group of “strain

unique genes" from the disposable genome. This conclusion was based on the observation that none of these genes fulfilled the criteria defined for CGs for the R3, Z1 and Z2 proteins. If these proteins were not encoded by genes of the disposable genome, we had to search among genes of the core genome. it has been reported that some proteins could be encoded in the genome without detectable expression. This fact raises the possibility that one or more of the GBS genomes used in the comparison (A909, 515, 2603 V/R and NEM316) could have the gene, in spite of being serotyped as negative for the expression of the R3, Z1 and Z2 proteins. In fact, previous research indicates that bacterial genes may not always be expressed, or be expressed in quantities insufficient for detection of the gene product [61]. For instance, GBS strains can possess an alpha-like-protein, even if the protein is not expressed on the bacterial surface [45]. This has been reported for R3, where strains previously tested and found to be negative for the expression of R3, later were found to produce antigen at low level[24]. However, the mechanism behind this has not been reported. This result suggested that genome comparison might not be the most suitable method to identify CGs for these three GBS proteins, and therefore needs to be complemented with other approaches.

In an attempt to reduce the number of candidate genes, molecular weight filtering criteria were applied. However, the molecular weight calculated for each of the GBS CDSs did not correspond with the expected MWs estimated from previous experiments with the R3, Z1 and Z2 proteins. None of the CDSs in GMFR293 and CMFR30 had a molecular weight as high as that expected for the Z1 protein. This finding could suggest that the molecular weight of at least the Z1 protein, previously estimated to be higher than 250 kDa had been overestimated. If so, a similar molecular weight overestimation could have been done for the R3 and Z2 proteins molecular

weights as well. The previous estimates of the molecular weights of the three proteins were obtained through Western blotting experiments which showed multiple bands [11]. It is, however, known that the analysis of a protein can be difficult if multiple bands appear on the blot. There are several possible reasons that could explain this kind of pattern. Multiple bands could be due to technical artefacts or could represent true variants of the protein of interest, like for instance repeats, which has been reported for other GBS surface proteins. Usually, higher molecular weight bands than the real molecular weight of the target protein may be seen when there is presence of unresolved multimers (protein complexes), or when the target protein is postranslationally modified (PTM).

A similar pattern was reported in the identification of the Srr-2 GBS surface protein[115], where a band was detected at >250 kDa, and the real molecular weight determined was 125 kDa. Several smaller molecular mass bands appeared also in the gel. The real Srr-2 protein molecular weight was resolved by adding urea (9M) to the buffer. This suggested that the protein could exist as a dimer in the absence of strongly denaturing conditions. Abnormal migration could also be attributed to the highly repetitive nature of the protein. Posttranslational modifications were thought to occur only rarely in bacteria. However, mass spectrometry (MS)-based proteomics has shown that prokaryotes are capable of modifying proteins with an extensive array of posttranslational modifications, and that these may have a profound influence on bacterial physiology and virulence[116], as shown for protein phosphorylation in *Streptococcus pneumoniae*[117]. Based on *in silico* analyses in this study, we conclude that the molecular weights for the R3, Z1 and Z2 proteins is uncertain and that further studies should be done in order to clarify this aspect. However, based on the characteristics previously reported

regarding molecular weight for these proteins together with the results obtained after characterization of Mws of CDSs of the two genomes, as well as the considerations mentioned above, we chose to define a filter for molecular weight for CGs for R3, Z1 and Z2 higher than 50 kDa.

The R3, Z1 and Z2 proteins were previously reported to be surface-exposed proteins[11]. Our use of *in silico* methods permitted the identification of CGs with potential of surface exposition. Proteins of Gram-positive bacteria destined for transport across the cytoplasmic membrane, frequently contain a hydrophobic N-terminal signal sequence. The approach described here provides an approximation to potential surface-exposed proteins in GMFR293 and CMFR30. Any of the identified CGs could be the genes encoding R3, Z1 and Z2. However, it is also important to be aware that not all membrane associated proteins have the structural elements used for identification by *in silico* analyses in this study. Some GBS surface proteins reported previously have been identified as having atypical structure characteristics, which basically means that exceptions could occur. For instance, proteins of Gram-positive bacteria destined for transport across the cytoplasmic membrane, frequently contain a hydrophobic N-terminal signal sequence[118]. Peculiarly, a number of secreted streptococcal proteins lack apparent secretion signal sequences[119]; thus the mechanism by which these proteins are transported to the bacterial cell surface is yet to be elucidated.

A further characterization of the R3, Z1 and Z2 proteins using for instance proteomics approaches could contribute to a more suitable prioritization to target a more limited group of CGs, or even to identify some CDSs which were discarded by the filters used in this study. In fact, similar studies combining proteomics and *in silico* prediction methods have been reported

for the identification of vaccine candidate genes for Group A Streptococcus (GAS)[120], *Streptococcus pyogenes*[121] and other bacterial genera[122].

In addition, based on CG prioritization by *in silico* analysis of the CDSs in the GMFR293 genome, we found that the gene encoding to R5 appears to be a strong CG for R3 (CDS-1223). From the beginning of the project, the possibility that R3 could be identical to R5 was one of the formulated hypotheses. During this work several analyses were done to attempt to clarify the R3-R5 relationship. The negative results by immunofluorescence testing for R3 expression from strain 2603V/R by Western blotting, and from *E. coli* BL21 cells transformants could indicate that they are different proteins. However, as discussed above, low level expression of the protein below the detection limit could be an explanation for the negatives results. In addition, there may be several other ways to explain the negative result for R3 expression, by immunofluorescence testing of *E. coli* BL21 cells transformants. First, a not recombinant protein may have been produced by the transformed *E. coli* LB 21 cells, or that *E. coli* LB21 cells had been transformed, but they did not express the R3 protein. Second, the recombinant protein may have been located intracellularly. Finally, the protein may not have been secreted. Unfortunately, time restrictions did not permit further testing of these possibilities Future studies along these lines are therefore recommended.

6 CONCLUSIONS

The work described in this thesis is the first attempt to identify the genes encoding the R3, Z1 and Z2 surface exposed protein of GBS. It was done through the sequencing of two GBS genomes and the analysis of their genome sequences. From the results obtained in this thesis, we conclude the following aspects:

The GBS strains GMFR293 and CMFR30 were sequenced and assembled into complete genomes, annotated and characterized. The availability of these strains allowed making the analysis needed for selection of the candidate genes for R3, Z1 and Z2 proteins. However, to obtain finished genomes, uncertain regions of the draft genomes need to be verified experimentally.

Genome comparison analysis were not a suitable approach to select candidates genes coding for R3, Z1 and Z2 surface exposed proteins of GBS. However, the comparison of GMFR293 and CMFR30 against other GBS genomes allowed identification of the strain-specific genes from both Zimbabwean strains.

The genome analysis using *in silico* tools was a rapid and inexpensive approach to target CGs for the R3, Z1 and Z2 GBS surface proteins. Additionally, a relevant conclusion from this work is the demonstration that a comprehensive characterization *in silico* of surface-exposed proteins can lead to candidate gene discovery of surface exposed proteins.

Finally, 51 CGs were chosen as CGs for R3, Z1 and Z2 in the GMFR293, and 32 CDSs were chosen as CGs for Z1 in GMFR30 genome. Among them, there were CDSs annotated as hypothetical protein with putative function, and some with predicted function. The results presented in this study represents an interesting first stage in the way for discovering the genes encoding the R3, Z1 and Z2 GBS surface exposed proteins. However, CGs identified by *in silico* analyses need to be tested further via experimental analyses for validation of the results. Further outcomes may be obtained if more information about the proteins becomes available in future.

The relationship between the R5 and R3 GBS proteins could not be clarified in this study, in spite of the experiments done. However, the gene coding for the R5 protein appears as one of the potential CGs for the R3 and Z2 surface exposed proteins by *in silico* analysis. Unfortunately, time restrictions did not permit further testing.

7 REFERENCES

1. Dramsi, S., Caliot, E., Bonne, I., Guadagnini, S., Prevost, M.C., Kojadinovic, M., Lalioui, L., Poyart, C., and Trieu-Cuot, P. (2006). Assembly and role of pili in group B streptococci. Molecular microbiology 60, 1401-1413.
2. Bolduc, G.R., and Madoff, L.C. (2007). The group B streptococcal alpha C protein binds alpha1beta1-integrin through a novel KTD motif that promotes internalization of GBS within human epithelial cells. Microbiology (Reading, England) 153, 4039-4049.
3. Ramaswamy, S.V., Ferrieri, P., Flores, A.E., and Paoletti, L.C. (2006). Molecular characterization of nontypeable group B streptococcus. Journal of clinical microbiology 44, 2398-2403.
4. Lindahl, G., Stalhammar-Carlemalm, M., and Areschoug, T. (2005). Surface proteins of Streptococcus agalactiae and related proteins in other bacterial pathogens. Clinical microbiology reviews 18, 102-127.
5. Mavenyengwa, R.T., Maeland, J.A., and Moyo, S.R. (2009). Putative novel surface-exposed Streptococcus agalactiae protein frequently expressed by the group B streptococcus from Zimbabwe. Clinical and vaccine immunology : CVI 16, 1302-1308.
6. Erdogan, S., Fagan, P.K., Talay, S.R., Rohde, M., Ferrieri, P., Flores, A.E., Guzman, C.A., Walker, M.J., and Chhatwal, G.S. (2002). Molecular analysis of group B protective surface protein, a new cell surface protective antigen of group B streptococci. Infection and immunity 70, 803-811.
7. Kvam, A.I., Bevanger, L., and Maeland, J.A. (1999). Properties and distribution of the putative R3 protein of Streptococcus agalactiae. APMIS : acta pathologica, microbiologica, et immunologica Scandinavica 107, 869-874.
8. Ferrieri, P., Baker, C.J., Hillier, S.L., and Flores, A.E. (2004). Diversity of surface protein expression in group B streptococcal colonizing & invasive isolates. Indian J Med Res 119 Suppl, 191-196.
9. Brzychczy-Wloch, M., Gosiewski, T., Bodaszewska-Lubas, M., Adamski, P., and Heczko, P.B. (2012). Molecular characterization of capsular polysaccharides and surface protein genes in relation to genetic similarity of group B streptococci isolated from Polish pregnant women. Epidemiol Infect 140, 329-336.
10. Mavenyengwa, R.T., Maeland, J.A., and Moyo, S.R. (2008). Distinctive features of surface-anchored proteins of Streptococcus agalactiae strains

from Zimbabwe revealed by PCR and dot blotting. Clinical and vaccine immunology : CVI 15, 1420-1424.

11. Maeland, J.A., Radtke, A., Lyng, R.V., and Mavenyengwa, R.T. (2013). Novel Aspects of the Z and R3 Antigens of Streptococcus agalactiae Revealed by Immunological Testing. Clinical and vaccine immunology : CVI.

12. Glazer, A.N., and Nikaido, H. (2007). Microbial biotechnology: fundamentals of applied microbiology.(Cambridge University Press).

13. Patterson, M.J. (1996). Streptococcus. In Medical Microbiology, S. Baron, ed. (Galveston TX, The University of Texas Medical Branch at Galveston.

14. Krzysciak, W., Pluskwa, K.K., Jurczak, A., and Koscielniak, D. (2013). The pathogenicity of the Streptococcus genus. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology 32, 1361-1376.

15. Hardie, J.M., and Whiley, R.A. (1997). Classification and overview of the genera Streptococcus and Enterococcus. Society for Applied Bacteriology symposium series 26, 1S-11S.

16. Garrity, G.M., and Bergey, D.H. (2001). Bergey's manual of systematic bacteriology.(New York: Springer).

17. Gibson, R.L., Nizet, V., and Rubens, C.E. (1999). Group B streptococcal beta-hemolysin promotes injury of lung microvascular endothelial cells. Pediatric research 45, 626-634.

18. Le Doare, K., and Heath, P.T. (2013). An overview of global GBS epidemiology. Vaccine 31 Suppl 4, D7-12.

19. Rato, M.G., Bexiga, R., Florindo, C., Cavaco, L.M., Vilela, C.L., and Santos-Sanches, I. (2013). Antimicrobial resistance and molecular epidemiology of streptococci from bovine mastitis. Veterinary microbiology 161, 286-294.

20. Vandamme, P., Devriese, L.A., Pot, B., Kersters, K., and Melin, P. (1997). Streptococcus difficile is a nonhemolytic group B, type Ib Streptococcus. International journal of systematic bacteriology 47, 81-85.

21. Younan, M., and Bornstein, S. (2007). Lancefield group B and C streptococci in East African camels (Camelus dromedarius). The Veterinary record 160, 330-335.

22. Mahmmod, Y.S., Toft, N., Katholm, J., Gronbaek, C., and Klaas, I.C. (2013). Estimation of test characteristics of real-time PCR and bacterial culture for diagnosis of subclinical intramammary infections with Streptococcus agalactiae in Danish dairy cattle in 2012 using latent class analysis. Preventive veterinary medicine 109, 264-270.

23. Amal, M.N., Zamri-Saad, M., Siti-Zahrah, A., Zulkafli, A.R., and Nur-Nazifah, M. (2013). Molecular characterization of Streptococcus agalactiae strains isolated from fishes in Malaysia. Journal of applied microbiology 115, 20-29.

24. Mavenyengwa, R.T., Maeland, J.A., and Moyo, S.R. (2010). Serotype markers in a Streptococcus agalactiae strain collection from Zimbabwe. Indian journal of medical microbiology 28, 313-319.
25. Johri, A.K., Paoletti, L.C., Glaser, P., Dua, M., Sharma, P.K., Grandi, G., and Rappuoli, R. (2006). Group B Streptococcus: global incidence and vaccine development. Nature reviews Microbiology 4, 932-942.
26. Yook, J.H., Kim, M.Y., Kim, E.J., Yang, J.H., Ryu, H.M., Oh, K.Y., Shin, J.H., Foxman, B., and Ki, M. (2013). Risk factors associated with group B streptococcus resistant to clindamycin and erythromycin in pregnant korean women. Infection & chemotherapy 45, 299-307.
27. Verani, J.R., McGee, L., and Schrag, S.J. (2010). Prevention of perinatal group B streptococcal disease--revised guidelines from CDC, 2010. MMWR Recommendations and reports : Morbidity and mortality weekly report Recommendations and reports / Centers for Disease Control 59, 1-36.
28. Boyer, K., and Gotoff, S. (1985). Strategies for chemoprophylaxis of GBS early-onset infections. Antibiotics and chemotherapy 35, 267.
29. Edmond, K.M., Kortsalioudaki, C., Scott, S., Schrag, S.J., Zaidi, A.K., Cousens, S., and Heath, P.T. (2012). Group B streptococcal disease in infants aged younger than 3 months: systematic review and meta-analysis. Lancet 379, 547-556.
30. McQuaid, F., Jones, C., Stevens, Z., Plumb, J., Hughes, R., Bedford, H., Heath, P.T., and Snape, M.D. (2013). Attitudes towards vaccination against group B streptococcus in pregnancy. Archives of disease in childhood.
31. Schrag, S., Gorwitz, R., Fultz-Butts, K., and Schuchat, A. (2002). Prevention of perinatal group B streptococcal disease. MMWR Recommendations and reports : Morbidity and mortality weekly report Recommendations and reports / Centers for Disease Control 51, 1-22.
32. Lawrence C Paoletti, L.C.M., Carol J Baker. Vaccines for the prevention of group B streptococcal disease. In. (
33. Deutscher, M., Lewis, M., Zell, E.R., Taylor, T.H., Jr., Van Beneden, C., and Schrag, S. (2011). Incidence and severity of invasive Streptococcus pneumoniae, group A Streptococcus, and group B Streptococcus infections among pregnant and postpartum women. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 53, 114-123.
34. Valls-Pascual, E., Alegre-Sancho, J.J., Ivorra-Cortes, J., Roman-Ivorra, J.A., Fernandez-Llanio-Comella, N., Chalmeta-Verdejo, I., Munoz-Gil, S., and Senabre-Gallego, J.M. (2008). [Joint Infections Due to Streptococcus agalactiae in Non Immunocompromised Adults: Presentation of Two Cases]. Reumatologia clinica 4, 155-158.
35. Sunkara, B., Bheemreddy, S., Lorber, B., Lephart, P.R., Hayakawa, K., Sobel, J.D., Kaye, K.S., and Marchaim, D. (2012). Group B Streptococcus infections in non-pregnant adults: the role of immunosuppression.

International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases 16, e182-186.

36. Sitkiewicz, I., and Hryniewicz, W. (2010). Pyogenic streptococci--danger of re-emerging pathogens. Polish journal of microbiology / Polskie Towarzystwo Mikrobiologow = The Polish Society of Microbiologists 59, 219-226.

37. Rajagopal, L. (2009). Understanding the regulation of Group B Streptococcal virulence factors. Future microbiology 4, 201-221.

38. Caliot, E., Dramsi, S., Chapot-Chartier, M.P., Courtin, P., Kulakauskas, S., Pechoux, C., Trieu-Cuot, P., and Mistou, M.Y. (2012). Role of the Group B antigen of Streptococcus agalactiae: a peptidoglycan-anchored polysaccharide involved in cell wall biogenesis. PLoS pathogens 8, e1002756.

39. Radtke, A. (2012). Molecular Methods for Typing of Streptococcus agalactiae with Special Emphasis on the Development and Validation of a Multi-Locus Variable Number of Tandem Repeats Assay (MLVA).

40. Maisey, H.C., Doran, K.S., and Nizet, V. (2008). Recent advances in understanding the molecular basis of group B Streptococcus virulence. Expert reviews in molecular medicine 10, e27.

41. Sagar, A., Klemm, C., Hartjes, L., Mauerer, S., van Zandbergen, G., and Spellerberg, B. (2013). The beta-hemolysin and intracellular survival of Streptococcus agalactiae in human macrophages. PloS one 8, e60160.

42. Otaguiri, E.S., Morguette, A.E., Tavares, E.R., dos Santos, P.M., Morey, A.T., Cardoso, J.D., Perugini, M.R., Yamauchi, L.M., and Yamada-Ogatta, S.F. (2013). Commensal Streptococcus agalactiae isolated from patients seen at University Hospital of Londrina, Parana, Brazil: capsular types, genotyping, antimicrobial susceptibility and virulence determinants. BMC microbiology 13, 297.

43. Hensler, M.E., Quach, D., Hsieh, C.J., Doran, K.S., and Nizet, V. (2008). CAMP factor is not essential for systemic virulence of Group B Streptococcus. Microbial pathogenesis 44, 84-88.

44. Nuccitelli, A., Rinaudo, C.D., Brogioni, B., Cozzi, R., Ferrer-Navarro, M., Yero, D., Telford, J.L., Grandi, G., Daura, X., Zacharias, M., et al. (2013). Understanding the molecular determinants driving the immunological specificity of the protective pilus 2a backbone protein of group B streptococcus. PLoS computational biology 9, e1003115.

45. Creti, R., Fabretti, F., Orefici, G., and von Hunolstein, C. (2004). Multiplex PCR assay for direct identification of group B streptococcal alpha-protein-like protein genes. Journal of clinical microbiology 42, 1326-1329.

46. Protein rib: a novel group B streptococcal cell surface protein that confers protective immunity and is expressed by most strains causing invasive infections.

47. Brodeur, B.R., Boyer, M., Charlebois, I., Hamel, J., Couture, F., Rioux, C.R., and Martin, D. (2000). Identification of group B streptococcal Sip protein, which elicits cross-protective immunity. Infection and immunity 68, 5610-5618.

48. Lalioui, L., Pellegrini, E., Dramsi, S., Baptista, M., Bourgeois, N., Doucet-Populaire, F., Rusniok, C., Zouine, M., Glaser, P., Kunst, F., et al. (2005). The SrtA Sortase of Streptococcus agalactiae is required for cell wall anchoring of proteins containing the LPXTG motif, for adhesion to epithelial cells, and for colonization of the mouse intestine. Infection and immunity 73, 3342-3350.

49. Papasergi, S., Lanza Cariccio, V., Pietrocola, G., Domina, M., D'Aliberti, D., Trunfio, M.G., Signorino, G., Peppoloni, S., Biondo, C., Mancuso, G., et al. (2013). Immunogenic properties of Streptococcus agalactiae FbsA fragments. PloS one 8, e75266.

50. van Sorge, N.M., Quach, D., Gurney, M.A., Sullam, P.M., Nizet, V., and Doran, K.S. (2009). The group B streptococcal serine-rich repeat 1 glycoprotein mediates penetration of the blood-brain barrier. Journal of Infectious Diseases 199, 1479-1487.

51. Seifert, K.N., Adderson, E.E., Whiting, A.A., Bohnsack, J.F., Crowley, P.J., and Brady, L.J. (2006). A unique serine-rich repeat protein (Srr-2) and novel surface antigen (ε) associated with a virulent lineage of serotype III Streptococcus agalactiae. Microbiology (Reading, England) 152, 1029-1040.

52. Bryan, J.D., and Shelver, D.W. (2009). Streptococcus agalactiae CspA is a serine protease that inactivates chemokines. Journal of bacteriology 191, 1847-1854.

53. Gase, K., Ozegowski, J., and Malke, H. (1998). The Streptococcus agalactiae hylB gene encoding hyaluronate lyase: completion of the sequence and expression analysis. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression 1398, 86-98.

54. Persson, E., Berg, S., Bevanger, L., Bergh, K., Valso-Lyng, R., and Trollfors, B. (2008). Characterisation of invasive group B streptococci based on investigation of surface proteins and genes encoding surface proteins. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases 14, 66-73.

55. Jerlstrom, P.G., Chhatwal, G.S., and Timmis, K.N. (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. Molecular microbiology 5, 843-849.

56. Michel, J.L., Madoff, L.C., Olson, K., Kling, D.E., Kasper, D.L., and Ausubel, F.M. (1992). Large, identical, tandem repeating units in the C protein alpha antigen gene, bca, of group B streptococci. Proceedings of the

National Academy of Sciences of the United States of America 89, 10060-10064.
57. Lachenauer, C.S., Creti, R., Michel, J.L., and Madoff, L.C. (2000). Mosaicism in the alpha-like protein genes of group B streptococci. Proceedings of the National Academy of Sciences of the United States of America 97, 9630-9635.
58. Gherardi, G., Imperi, M., Baldassarri, L., Pataracchia, M., Alfarone, G., Recchia, S., Orefici, G., Dicuonzo, G., and Creti, R. (2007). Molecular epidemiology and distribution of serotypes, surface proteins, and antibiotic resistance among group B streptococci in Italy. Journal of clinical microbiology 45, 2909-2916.
59. Moyo, S.R., Maeland, J.A., and Lyng, R.V. (2001). The putative R1 protein of Streptococcus agalactiae as serotype marker and target of protective antibodies. APMIS : acta pathologica, microbiologica, et immunologica Scandinavica 109, 842-848.
60. Bevanger, L., and Maeland, J.A. (1977). Type classification of group B streptococci by the fluorescent antibody test. Acta pathologica et microbiologica Scandinavica Section B, Microbiology 85B, 357-362.
61. Radtke, A., Kong, F., Bergh, K., Lyng, R.V., Ko, D., and Gilbert, G.L. (2009). Identification of surface proteins of group B streptococci: serotyping versus genotyping. Journal of microbiological methods 78, 363-365.
62. Wilkinson, H.W. (1972). Comparison of streptococcal R antigens. Applied microbiology 24, 669-670.
63. Dobrindt, U., and Hacker, J. (2001). Whole genome plasticity in pathogenic bacteria. Current opinion in microbiology 4, 550-557.
64. Klein, J., Munch, R., Biegler, I., Haddad, I., Retter, I., and Jahn, D. (2009). Strepto-DB, a database for comparative genomics of group A (GAS) and B (GBS) streptococci, implemented with the novel database platform 'Open Genome Resource' (OGeR). Nucleic acids research 37, D494-498.
65. Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences of the United States of America 102, 13950-13955.
66. Glaser, P., Rusniok, C., Buchrieser, C., Chevalier, F., Frangeul, L., Msadek, T., Zouine, M., Couve, E., Lalioui, L., Poyart, C., et al. (2002). Genome sequence of Streptococcus agalactiae, a pathogen causing invasive neonatal disease. Molecular microbiology 45, 1499-1513.
67. Richards, V.P., Lang, P., Bitar, P.D., Lefebure, T., Schukken, Y.H., Zadoks, R.N., and Stanhope, M.J. (2011). Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted Streptococcus agalactiae. Infection, genetics and evolution : journal of molecular

epidemiology and evolutionary genetics in infectious diseases 11, 1263-1275.

68. Jain, R., Rivera, M.C., Moore, J.E., and Lake, J.A. (2002). Horizontal gene transfer in microbial genome evolution. Theoretical population biology 61, 489-495.

69. Kunin, V., and Ouzounis, C.A. (2003). The balance of driving forces during genome evolution in prokaryotes. Genome research 13, 1589-1594.

70. Broker, G., and Spellerberg, B. (2004). Surface proteins of Streptococcus agalactiae and horizontal gene transfer. International journal of medical microbiology : IJMM 294, 169-175.

71. Brochet, M., Rusniok, C., Couve, E., Dramsi, S., Poyart, C., Trieu-Cuot, P., Kunst, F., and Glaser, P. (2008). Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of Streptococcus agalactiae. Proceedings of the National Academy of Sciences of the United States of America 105, 15961-15966.

72. Kvam, A.I., Mavenyengwa, R.T., Radtke, A., and Maeland, J.A. (2011). Streptococcus agalactiae alpha-like protein 1 possesses both cross-reacting and Alp1-specific epitopes. Clinical and vaccine immunology : CVI 18, 1365-1370.

73. Janulczyk, R., Masignani, V., Maione, D., Tettelin, H., Grandi, G., and Telford, J.L. (2010). Simple sequence repeats and genome plasticity in Streptococcus agalactiae. Journal of bacteriology 192, 3990-4000.

74. Facklam, R. (2002). What happened to the streptococci: overview of taxonomic and nomenclature changes. Clinical microbiology reviews 15, 613-630.

75. Radtke, A., Lindstedt, B.A., Afset, J.E., and Bergh, K. (2010). Rapid multiple-locus variant-repeat assay (MLVA) for genotyping of Streptococcus agalactiae. Journal of clinical microbiology 48, 2502-2508.

76. Jones, N., Bohnsack, J.F., Takahashi, S., Oliver, K.A., Chan, M.S., Kunst, F., Glaser, P., Rusniok, C., Crook, D.W., Harding, R.M., et al. (2003). Multilocus sequence typing system for group B streptococcus. Journal of clinical microbiology 41, 2530-2536.

77. Sun, Y., Kong, F., Zhao, Z., and Gilbert, G.L. (2005). Comparison of a 3-set genotyping system with multilocus sequence typing for Streptococcus agalactiae (Group B Streptococcus). Journal of clinical microbiology 43, 4704-4707.

78. Edwards, D.J., and Holt, K.E. (2013). Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microbial informatics and experimentation 3, 2.

79. Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nature biotechnology 26, 1135-1145.

80. Metzker, M.L. (2010). Sequencing technologies - the next generation. Nature reviews Genetics 11, 31-46.

81. Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annual review of genomics and human genetics 9, 387-402.
82. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science (New York, NY) 269, 496-512.
83. Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., et al. (1995). The minimal gene complement of Mycoplasma genitalium. Science (New York, NY) 270, 397-403.
84. Fraser, C.M., and Fleischmann, R.D. (1997). Strategies for whole microbial genome sequencing and analysis. Electrophoresis 18, 1207-1216.
85. Tettelin, H., and Feldblyum, T. (2009). Bacterial genome sequencing. Methods in molecular biology (Clifton, NJ) 551, 231-247.
86. Stothard, P., and Wishart, D.S. (2006). Automated bacterial genome analysis and annotation. Current opinion in microbiology 9, 505-510.
87. Zhang, T., Luo, Y., Chen, Y., Li, X., and Yu, J. (2012). BIGrat: a repeat resolver for pyrosequencing-based re-sequencing with Newbler. BMC research notes 5, 567.
88. Aston, C., Mishra, B., and Schwartz, D.C. (1999). Optical mapping and its potential for large-scale sequencing projects. Trends in biotechnology 17, 297-302.
89. Fraser, C.M., Eisen, J.A., Nelson, K.E., Paulsen, I.T., and Salzberg, S.L. (2002). The value of complete microbial genome sequencing (you get what you pay for). Journal of bacteriology 184, 6403-6405.
90. Chain, P., Grafham, D., Fulton, R., Fitzgerald, M., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D., and Buhay, C. (2009). Genome project standards in a new era of sequencing. Science (New York, NY) 326.
91. Samad, A., Huff, E.F., Cai, W., and Schwartz, D.C. (1995). Optical mapping: a novel, single-molecule approach to genomic analysis. Genome research 5, 1-4.
92. Riley, M.C., Lee, J.E., Lesho, E., and Kirkup, B.C., Jr. (2011). Optically mapping multiple bacterial genomes simultaneously in a single run. PloS one 6, e27085.
93. Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Parrello, B., Shukla, M., et al. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic acids research 42, D206-214.
94. Onmus-Leone, F., Hang, J., Clifford, R.J., Yang, Y., Riley, M.C., Kuschner, R.A., Waterman, P.E., and Lesho, E.P. (2013). Enhanced de novo assembly of high throughput pyrosequencing data using whole genome mapping. PloS one 8, e61762.

95. Latreille, P., Norton, S., Goldman, B.S., Henkhaus, J., Miller, N., Barbazuk, B., Bode, H.B., Darby, C., Du, Z., Forst, S., et al. (2007). Optical mapping as a routine tool for bacterial genome sequence finishing. BMC genomics 8, 321.

96. Zhou, S., Deng, W., Anantharaman, T.S., Lim, A., Dimalanta, E.T., Wang, J., Wu, T., Chunhong, T., Creighton, R., Kile, A., et al. (2002). A whole-genome shotgun optical map of Yersinia pestis strain KIM. Applied and environmental microbiology 68, 6321-6331.

97. Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., et al. (2008). The RAST Server: rapid annotations using subsystems technology. BMC genomics 9, 75.

98. Prentice, M.B. (2004). Bacterial comparative genomics. Genome biology 5, 338.

99. Tettelin, H., Masignani, V., Cieslewicz, M.J., Eisen, J.A., Peterson, S., Wessels, M.R., Paulsen, I.T., Nelson, K.E., Margarit, I., Read, T.D., et al. (2002). Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V Streptococcus agalactiae. Proceedings of the National Academy of Sciences of the United States of America 99, 12391-12396.

100. Patnala, R., Clements, J., and Batra, J. (2013). Candidate gene association studies: a comprehensive guide to useful in silico tools. BMC genetics 14, 39.

101. Lin, F.P., Coiera, E., Lan, R., and Sintchenko, V. (2009). In silico prioritisation of candidate genes for prokaryotic gene function discovery: an application of phylogenetic profiles. BMC bioinformatics 10, 86.

102. Wu, S., Zhu, Z., Fu, L., Niu, B., and Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC genomics 12, 444.

103. Wu, S., and Zhu, Y. (2012). ProPAS: standalone software to analyze protein properties. Bioinformation 8, 167-169.

104. Juhas, M., Der Meer, V., Roelof, J., Gaillard, M., Harding, R.M., Hood, D.W., and Crook, D.W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS microbiology reviews 33, 376-393.

105. Langille, M.G., and Brinkman, F.S. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics 25, 664-665.

106. Barinov, A., Loux, V., Hammani, A., Nicolas, P., Langella, P., Ehrlich, D., Maguin, E., and van de Guchte, M. (2009). Prediction of surface exposed proteins in Streptococcus pyogenes, with a potential application to other Gram-positive bacteria. Proteomics 9, 61-73.

107. Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature methods 8, 785-786.

108. Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., and Markowitz, S. (2011). Geneious, version 5.4. Geneious, Auckland, New Zealand.

109. Kisand, V., and Lettieri, T. (2013). Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. BMC genomics 14, 211.

110. Fichot, E.B., and Norman, R.S. (2013). Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. Microbiome 1, 10.

111. Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics 13, 341.

112. Mardis, E., McPherson, J., Martienssen, R., Wilson, R.K., and McCombie, W.R. (2002). What is finished, and why does it matter. Genome research 12, 669-671.

113. Grandi, G. (2001). Antibacterial vaccine design using genomics and proteomics. Trends in biotechnology 19, 181-188.

114. Fraser, C.M. (2000). Microbial genome sequencing: new insights into physiology and evolution. Novartis Foundation symposium 229, 54-58; discussion 58-62.

115. Seifert, K.N., Adderson, E.E., Whiting, A.A., Bohnsack, J.F., Crowley, P.J., and Brady, L.J. (2006). A unique serine-rich repeat protein (Srr-2) and novel surface antigen (epsilon) associated with a virulent lineage of serotype III Streptococcus agalactiae. Microbiology (Reading, England) 152, 1029-1040.

116. Cain, J.A., Solis, N., and Cordwell, S.J. (2014). Beyond gene expression: the impact of protein post-translational modifications in bacteria. Journal of proteomics 97, 265-286.

117. Sun, X., Ge, F., Xiao, C.L., Yin, X.F., Ge, R., Zhang, L.H., and He, Q.Y. (2010). Phosphoproteomic analysis reveals the multiple roles of phosphorylation in pathogenic bacterium Streptococcus pneumoniae. Journal of proteome research 9, 275-282.

118. Chhatwal, G.S. (2002). Anchorless adhesins and invasins of Gram-positive bacteria: a new class of virulence factors. Trends in microbiology 10, 205-208.

119. Lei, B., Mackie, S., Lukomski, S., and Musser, J.M. (2000). Identification and Immunogenicity of Group AStreptococcus Culture Supernatant Proteins. Infection and immunity 68, 6807-6818.

120. Rodríguez-Ortega, M.J., Norais, N., Bensi, G., Liberatori, S., Capo, S., Mora, M., Scarselli, M., Doro, F., Ferrari, G., and Garaguso, I. (2006).

Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome. Nature biotechnology 24, 191-197.

121. Severin, A., Nickbarg, E., Wooters, J., Quazi, S.A., Matsuka, Y.V., Murphy, E., Moutsatsos, I.K., Zagursky, R.J., and Olmsted, S.B. (2007). Proteomic analysis and identification of Streptococcus pyogenes surface-associated proteins. Journal of bacteriology 189, 1514-1522.

122. Ariel, N., Zvi, A., Makarova, K., Chitlaru, T., Elhanany, E., Velan, B., Cohen, S., Friedlander, A., and Shafferman, A. (2003). Genome-based bioinformatic selection of chromosomal Bacillus anthracis putative vaccine candidates coupled with proteomic identification of surface-associated antigens. Infection and immunity 71, 4563-4579.

8 *Appendix A*. Uncertain regions detected due to differences between the optical restriction map and the *in silico* restriction map patterns.

No	GMFR293 *in silico* fragment number	GMFR293 *in silico* fragment size	GMFR293 Optical Map matching fragment	GMFR293 Optical Map fragment size
1	5	12.948	4	12.856
2	6	5.873	5	6.641
3	7	1.296	no match	0
4	8	0.987	no match	0
5	9	1.694	no match	0
6	10	1.337	no match	0
7	11	6.326	8	7.008
8	12	0.471	no match	0
9	13	1.492	no match	0
10	23	3.356	18	3.574
11	24	12.163	19	15.024
13	28	26.121	22	25.702
15	32	0.207	no match	0
16	33	0.941	no match	0
17	34	8.38	no match	0
18	35	16.994	no match	0
19	38	12.584	29	13.369
20	39	0.717	no match	0
21	40	0.416	no match	0
22	42	19.395	31	19.049
23	50	7.387	38	7.434
24	56	4.824	43	5.104
25	57	1.433	no match	0
26	58	1.588	no match	0

27	59	2.923	45	3.189
28	61	4.385	47	4.494
29	62	1.934	no match	0
30	63	0.001G	no match	0
31	64	0.002	no match	0
32	65	0.001G	no match	0
33	66	0.397	no match	0
35	73	4.886	54	5.101
36	74	4.871	55	5.532
37	79	5.516	no match	0
40	86	13.728	64	14.085
41	91	4.631	68	5.437
42	111	12.973	87	12.28
43	126	13.127	101	13.797
44	133	6.892	107	6.979
45	134	1.583	no match	0
47	136	3.593	109	3.719
48	137	0.123	no match	0
49	148	19.42	120	18.805
50	163	6.049	134	6.603
59	198	11.949	164	12.042
61	202	0.403	no match	0
63	209	0.122	no match	0
65	228	1.679	no match	0
66	229	396	no match	0
67	230	8972	no match	0

9 *Appendix B.* Uncertain regions identified from the GBS GMFR293 genome assembly.

No	GMFR293 Frag. No	Size (Kb)	Start	End	Size (bp)	No match	DRP 1	Size Error %	DRP 2	Diff. size Ref. gen
1	63	0.001	392420	392421	1	x			x	
2	65	0.001	392423	392424	1	x			x	
3	64	0.002	392421	392423	2	x			x	
4	137	0.123	1113140	1113263	123	x				X
5	66	0.397	392424	392821	397	x			x	
6	33	0.941	224317	225258	941	x			x	
7	83	1.379	551674	553053	1379		x	x		
8	58	1.588	373160	374748	1588	x			x	
9	228	1.679	1978622	1980301	1679	x			x	
10	62	1.934	390486	392420	1934	x			x	
11	183	9.027	1520311	1522769	2458		x	x		
12	167	2.579	1368071	1370650	2579		x	x		
13	59	2.923	374748	377671	2923		x		x	
14	135	3.188	1106359	1109547	3188		x	x		
15	23	3.356	153085	156441	3356		x			X
16	61	4.385	386101	390486	4385		x		x	
17	91	4.631	725427	730058	4631		x	x	x	
18	74	4.871	506746	511617	4871		x	x		
19	6	5.873	59432	65305	5873		x	x		
20	11	6.326	70619	76945	6326		x	x		
21	189	6.404	1563347	1569751	6404		x	x		
22	34	8.38	225258	233638	8380	x			x	

23	198	11.949	1652179	1664128	11949		x		x	
24	24	12.163	156441	168604	12163		x	x		x
25	5	12.948	46484	59432	12948		x			x
26	86	13.728	593344	607072	13728		x		x	
27	35	16.994	233638	250632	16994	x			x	
28	227	17.95	1960672	1978622	17950		x			x
29	68	28.186	394379	422565	28186		x			x

*(DRP 1: different restriction pattern compared to that of the optical map, DRP 2: different restriction pattern compared to that of the optical map of reference genomes).

10 *Appendix C.* GBS CMFR30 strain specific genes.

CDs	Annotation	MW (kDa)	Topology	GEIs
83	FIG01114815: hypothetical protein	26	Inside	
295	Hypothetical protein	4.1	outside	
421	Hypothetical protein	3.8	TM	I
602	Hypothetical protein	6.2	outside	
606	Conserved hypothetical protein - phage associated	7.6	TM	
608	Hypothetical protein	8.1	outside	
613	Hypothetical protein	9.1	Inside	III
616	hypothetical protein	5.8	TM	III
620	Hypothetical protein	4.2	TM	
716	Hypothetical protein	4.9	outside	
734	Hypothetical protein	4	TM	IV
778	Hypothetical protein	4.3	TM	
782	Hypothetical protein	4	TM	
812	Hypothetical protein	4.2	outside	
939	Conserved hypothetical protein	4.6	outside	
941	Hypothetical protein	11.7	outside	
1146	Hypothetical protein	8.2	outside	
1455	Hypothetical protein	4.7	TM	
1605	Hypothetical protein	8.8	outside	
1665	Hypothetical protein	7.3	TM	
1666	Hypothetical protein	7.2	TM	
1668	Hypothetical protein	13.7	Inside	
1669	Hypothetical protein	11.6	outside	
1670	Hypothetical protein	4.2	outside	
1796	Hypothetical protein	4.2	TM	

11 *Appendix D.* Candidate genes for the R3, Z1 and Z2 surface exposed proteins in the GBS GMFR293 genome.

No	CDS	MW (Kda)	Topology	Signal peptide	YSIRK	LPxTG	Lipo-protein	annotation
1	137	72.41	TM	x		x		Hypothetical protein
2	189	61.05	TM	x		x	x	Hypothetical protein
3	298	54.41	TM					Hypothetical protein SPy1643
4	384	56.34	outside	x	x			Cell wall surface anchor family protein
5	415	118.16	TM					Cell surface protein
6	431	91.30	TM					Surface protein Rib
7	532	53.13	outside					Hypothetical protein
8	601	71.83	outside					Membrane proteins related to metalloendopeptidases
9	651	78.40	outside					Hypothetical protein
10	747	54.38	TM	x				Cell wall surface anchor family protein
11	758	85.54	TM				x	Late competence protein ComEC, DNA transport
12	965	54.49	TM				x	Carbon starvation protein A
13	991	52.03	TM	x				Putative secretion accessory protein EsaA/YueB
14	1039	74.35	TM			x		Kup system potassium uptake protein
15	1154	94.16	TM	x				Conserved domain protein
16	1163	139.74	TM	x	x			Pullulanase
17	1185	119.70	outside	x				C5a peptidase
18	1186	53.14	outside					Hypothetical protein
19	1223	105.62	TM	x	x	x		Hypothetical protein
20	1229	66.24	TM				x	Lipid A export ATP-binding/permease protein MsbA
21	1230	64.89	TM			x		Lipid A export ATP-binding/permease protein MsbA

22	1242	60.12	TM	x	x			Surface antigen-related protein
23	1282	60.32	TM					Membrane protein involved in the export of O-antigen, teichoic acid lipoteichoic acids
24	1299	73.49	TM	x				Cell wall surface anchor family protein
25	1300	101.01	TM	x				Cell wall surface anchor family protein, FPXTG motif
26	1304	52.84	TM			x		Membrane protein involved in the export of O-antigen, teichoic acid lipoteichoic acids
27	1305	51.73	TM			x	x	Membrane protein, putative
28	1311	66.65	TM	x			x	Lipoprotein involved in the synthesis of group B streptococcal carboyhdrate antigen
29	1357	79.05	TM	x		x		Glutamine ABC transporter, glutamine-binding protein/permease protein
30	1365	74.58	TM	x			x	Amidase family protein
31	1376	64.35	TM					Membrane protein, putative
32	1434	83.13	TM					Hypothetical protein
33	1435	50.28	outside					Hypothetical protein
34	1469	52.61	TM				x	Potassium uptake protein TrkH
35	1501	57.51	TM					Transmembrane histidine kinase CsrS
36	1734	52.71	outside					Hypothetical protein
37	1761	52.15	TM				x	PTS system, galactose-specific IIC component
38	1766	53.32	TM	x				Streptococcal histidine triad protein
39	1779	80.77	TM	x				Membrane protein, putative
40	1780	93.31	outside					Hypothetical protein
41	1786	53.37	TM					Hypothetical protein
42	1808	77.75	TM				x	PTS system, maltose

								and glucose-specific IIC component
43	1837	74.74	TM				x	Membrane protein, putative
44	1854	94.20	TM	x				Hypothetical protein
45	1855	81.69	TM	x			x	Conserved domain protein
46	1899	172.29	TM	x	x			Serine endopeptidase ScpC
47	1914	61.03	TM					Hypothetical protein
48	1934	59.33	TM				x	Competence-induced protein Ccs4
49	1960	54.76	outside					Hypothetical protein
50	1979	96.67	TM			x		Membrane protein, putative
51	1989	74.56	TM				x	Phosphoesterase, DHH family protein

12 *Appendix E.* Candidate genes for Z1surface exposed protein obtained from GBS CMFR30 genome.

CG No	CDS	MW (Kda)	Topology	Signal Peptide	YSIRK	LPxTG	Lipo-protein	Annotation
1	112	102.1	outside		x	x		Surface protein Rib
2	128	105.6	TM			x		Cell surface protein
3	133	80.1	TM			x		C5a peptidase
4	159	52.4	outside	x	x	x	x	Cell wall surface anchor family protein
5	305	54.5	Inside					Hypothetical protein
6	399	55.2	TM	x			x	Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA
7	410	62.8	TM	x				Hypothetical protein
8	594	66.1	TM				x	Phosphoesterase, DHH family protein
9	628	84.7	outside			x		Membrane protein, putative
10	671	51.4	TM				x	Competence-induced protein Ccs4
11	690	54.7	TM					Hypothetical protein
12	694	73.2	TM			x		Putative peptidoglycan linked protein (LPXTG motif)
13	702	157.2	TM	x	x	x		Serine endopeptidase ScpC
14	801	80.1	TM	x		x		Cyclic-nucleotide-phosphodiesterase
15	944	63	TM					Hypothetical protein
16	1025	88.2	TM				x	Pyruvate,phosphate dikinase
17	1071	50.2	TM					Transmembrane histidine kinase CsrS
18	1140	72	TM					Hypothetical protein
19	1191	53.1	TM	x				SLH, S-layer homology domain W
20	1196	54.9	TM					Membrane protein, putative
21	1207	68	TM	x		x	x	Amidase family protein

22	1271	89.7	TM	x		x		Cell wall surface anchor family protein, FPXTG motif
23	1272	68.3	outside	x		x		Cell wall surface anchor family protein
24	1289	54.5	outside					Membrane protein involved in the export of O-antigen, teichoic acid lipoteichoic acids
25	1328	54.5	outside	x	x	x		Surface antigen-related protein
26	1424	85.5	TM	x				Conserved domain protein
27	1593	100.6	TM	x				Putative secretion accessory protein EsaA/YueB
28	1770	81.6	TM					Hypothetical protein
29	1771	72.6	outside	x				Membrane protein, putative
30	1957	88.6	outside					Lactocepin (Cell wall-associated serine proteinase)
31	1981	90.3	outside			x		Cell wall surface anchor family protein, FPXTG motif
32	1985	55.5	outside	x		x		Cell wall surface anchor family protein