The role of the mobilome in the evolution and host-adaptation of *Streptococcus agalactiae*

UNIVERSITY
of
GLASGOW

Chiara Crestani, DVM

Institute of Biodiversity, Animal Health & Comparative Medicine

College of Medical, Veterinary & Life Sciences

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

July 2021

Abstract

Streptococcus agalactiae, also known as group B *Streptococcus* (GBS), is a complex multi-host opportunistic bacterial pathogen. In human medicine it is recognised as a leading invasive neonatal pathogen, an emerging pathogen of non-pregnant adults and a newly-emerged foodborne pathogen. In veterinary medicine, GBS is a well-known mastitis-causing agent in dairy cattle, an important invasive pathogen of warm-water fish species in aquaculture, and an emerging pathogen of dromedary camels. Adaptation to new hosts and ecological niches of several bacterial pathogens has been linked with the acquisition of various types of mobile genetic elements (MGE), dynamic molecular parasites that can be transferred between bacterial cells, which together from the 'mobilome'. In GBS, certain MGE have been associated with host-adaptation, with high pathogenicity (e.g. bacteriophages and insertion sequences) and with remodelling of population structure due to positive selection (e.g. integrative conjugative elements, ICE, for tetracycline resistance); however, most studies to date primarily focused on human GBS.

The overarching aim of this work was to assess the role of the mobilome in host-adaptation and evolution of GBS with extensive comparative genomic analyses across host groups. This was carried out through specific objectives: i) Fill knowledge gaps with regards to presence and distribution of various classes and types of MGE among GBS lineages and host groups. An implementation and evaluation of existing methods for the detection of MGE in GBS was carried out to facilitate subsequent analyses, and a new typing and detection method for GBS prophages and phage-inducible chromosomal islands (PICI) was developed. Findings show a high diversity of prophage types and of their relative insertion sites, as well as of ICE. One PICI type appears to be ubiquitous in GBS, but PICI as a class show low diversity in GBS compared to other bacterial species, except for GBS from camels. Overall, few known plas-

mids were detected among GBS isolates of human origin, but thanks to long-read sequencing, novel plasmids with homologs in other streptococci were identified, one of which was highly prevalent among bovine GBS. ii) Improving our understanding of the GBS population structure both at the national (bovine GBS in Sweden, camel GBS in Kenya) and global levels, and the genetic background associated with host-specialist and host-generalist lineages, through analysis of core and accessory genome content. Results support the possibility of reverse zoonotic transmission, with introduction of new lineages of human origin in dairy cattle and subsequent adaptation to the bovine niche thanks to the acquisition of relevant MGE. A high genome plasticity of host-generalists was detected, suggesting these lineages might have a superior ability to uptake and retain foreign DNA compared to host-specialists, from which they differ considerably in terms of recombinogenic potential. Host-specialists and generalists seem to largely evolve independently of each other. iii) Investigate the association of accessory genes and MGE, which could be exerting an impact in host-adaptation, with the major GBS host groups (humans, bovines, fishes and camels) through large-scale genome-wide association studies (GWAS). Findings indicate that a limited number of genomic islands (GEI), some of which are recognised MGE, are associated with each host group: *scpB-lmb* transposon in humans, Lac.2 in cattle, locus 3 in fishes and two major GEI in camels. The distribution of the former three elements among host-specialist and host-generalist lineages within GBS, and among other streptococci that affect the same hosts, suggests they are potentially major drivers of host-adaptation in GBS and in streptococci more widely. The presence vs absence of these host-associated genetic markers demarcates separate 'ecotypes', i.e. groups of bacterial species and strains that are well adapted to a certain ecological niche.

Overall, this work shows that the pangenome cannot be understood without a focus on all affected host species, that the GBS mobilome comprises many types of MGE, and that a select group of MGE may drive host-adaptation both within and beyond GBS.

Contents

List of Tables

List of Figures

Acknowledgements

Scientific Supervision

First and foremost, to my PhD project supervisors. To Ruth Zadoks, an incredibly bright mind and an impossible combination of humanity and intelligence, full of passion for the complex world surrounding us. Thank you for accepting me as your PhD student when you didn't know anything about me, for guiding me through this challenging journey, for teaching me how to be a better researcher and how to ask the right questions. Thank you for believing in me when I did not, for the sensitivity with which you managed to understand my feelings when I was not listening to myself, and for always caring for the person behind the student. You are full of rare qualities, and I couldn't have asked more from a supervisor.

To Taya Forde, who has always been able to 'infect' me with her never-ending positive attitude. Thank you for the incalculable support you have given me in the past two years: I couldn't have done it without you. Thank you for being a role model and an inspiration for the young Italian veterinarian who arrived in the lab without knowing anything about bioinformatics. Thank you for your endless availability, for always providing detailed feedback, and for supporting me in navigating all the aspects of academic life, from science to administration. You are a great inspiration for young female scientists and I am sure you have a bright future ahead.

Scientific Collaborations

I would like to heartily thank the many collaborators that I have had the privilege to work with in the past four years. In particular Dr Samantha Lycett, who kindly taught me BEAST (or at least a small but very important part of it) and who was always available when I asked for help. To Dr Nicola Lynskey, who I wish I met four years ago, for running preliminary lab

experiments and for her contagious passion for science and streptococci. To the people who provided isolates and/or generated whole genome sequence data first described in this thesis: Prof Karin Persson-Waller, Dr Katrina Bosward, Dr Derek Brown, Prof Andrew Smith, Dr Jørgen Katholm, Dr Ulrike Lyhs, Dr Nguyen Ngoc Phuoc, Dr Wanna Siramanapong, Prof Mark Holmes and Prof Swaine Chen. I also want to thank the Wellcome Sanger Institute, specifically Prof Stephen Bentley and his group, in particular Dr Dorota Jamrozy, for bioinformatic training and for giving me access to computational resources that were essential to the completion of this work. To Dr Nuria Quiles-Puchalt, who taught me about the fantastic world of prophages and PICI: you are one of the best people in science I met during my PhD.

A special mention goes to Dinah Seligsohn: our scientific collaboration turned into a trusted friendship. You are a bright veterinarian and a wonderful person, and I hope our professional and personal relationship will last for years to come.

Financial Support

I would like to acknowledge the University of Glasgow and in particular the College of Medical, Veterinary and Life Sciences for funding my PhD project with the Doctoral Training Programme studentship, and for awarding me travel and training grants throughout my PhD journey.

Family

Ai miei genitori. Grazie per essere sempre presenti, disponibili e per tutto l'incoraggiamento che mi avete sempre dato. Grazie per avermi insegnato che l'impegno e il duro lavoro, al pari di camminare a lungo e faticosamente su per un pendio erto, portano più lontano e più in alto del talento.

A mia sorella Irene, che mia ha insegnato che il cielo è dappertutto. Grazie di essere stata la mia maestra, la mia compagna di giochi e di avventure, ma soprattutto grazie di essere la mia migliore amica.

Friends

Un grazie di cuore a Davide Pagnossin, per essere stato negli ultimi tre anni l'amico a cui potevo sempre rivolgermi in momenti difficili e in quelli gioiosi. A Michele De Noia: grazie per le risate, per le passeggiate, per le imprecate, per i traslochi fatti insieme, per le piante e il gin and tonic, per la leggerezza e il gusto della vita. To Eleni Christoforu, thank you because I couldn't have hoped to find such compatibility with a person so far from home: you have been the best of friends. Also thanks to Christodoulos Karakannas, a wonderful person whom I got to know better in the past year: you two make people around you feel great, and I am very happy you are together. To my loves Emily Hoyt and Paula Struthoff: thank you for being an incredible support in dark times, and thank you for existing as these two genuine wonderful and fierce human beings. I cherish our friendship very deeply. To Kate Ings, thank you for being the best climbing buddy, for the van talks, the anti-patriarchy talks, and for being just this cool strong independent woman. I will miss you. Grazie a tutte le ragazze dello Zoo, Angelica, Vale e Betty, colleghe diventate amiche: la nostra chimica è stata una sorpresa meravigliosa. Grazie a Elena Costa, un'altra donna in gambissima che stimo molto e a cui voglio veramente bene.

Ultima non per importanza, alla mia cara amica Nadia Degregorio, ti faccio i ringraziamenti che erano un po' mancati anni fa. Grazie di essere una persona speciale e di essere mia amica da tanto tempo. Grazie di avermi accompagnata negli anni dell'università, di tutte le risate e le avventure insieme. Sei una roccia dal cuore dolce e morbido.

General Assistance

Heartily thanks to all the fantastic OHRBID lab members in particular Katarina, Joel, Assel, Chris and Ryan, but also to the lovely Marie and Marg. To the wonder ladies of the writing club, Gemma, Stef, Diana, Ellen, it was a pleasure to write my thesis in your company.

Everyone else

To Jamie the cactus, to pomofocus.io, the god of the twenty-five minutes, to FRIENDS, Jane the Virgin, to the MinION (both the machine and the cartoons) and to the EURO 2020 friends. To Paesano pizza. To myself.

Author's declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. It is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University and no part has already been, or is concurrently being, submitted for any degree, diploma, or other qualification. It does not exceed 80,000 words, excluding references, bibliography, table of contents and appendices.

Chapter 2: Development and application of methods for the identification of mobile genetic elements in group B *Streptococcus* genomes from multiple host species. Initial concept developed by C. Crestani and R. N. Zadoks. Data collection of available sequenced data was carried out by C. Crestani and by V. P. Richards (Clemson University); the latter was also responsible for generation of new sequenced data. Analysis was conducted by C. Crestani with initial advice from N. Quiles-Puchalt and J. R. Penadés (Imperial College London). The chapter was written by C. Crestani, and revised by R. N. Zadoks and T. L. Forde.

Chapter 3: Genomic explanations for the temporal shift in bovine group B *Streptococcus* subpopulations in Sweden. Initial conceptualisation by R. N. Zadoks and K. Persson-Waller (SVA -National Veterinary Institute Sweden), with development by C. Crestani. Bacterial isolates were provided by K. Persson-Waller and C. Fasth (SVA). Sequenced data was generated by staff at the Moredun Research Institute and M. A. Holmes (University of Cambridge) for Illumina and by C. Crestani for Oxford Nanopore. Analysis was conducted by C. Crestani, with advice from R. N. Zadoks. The chapter was written by C. Crestani, and revised by R. N. Zadoks and T. L. Forde.

Chapter 4: Host-specialist and generalist lineages of group B *Streptococcus* are associated with distinct patterns of core and accessory genome content and variable levels of genome plasticity. Initial concept developed by C. Crestani, R. N. Zadoks and T. L. Forde. Data collection of available sequenced data was carried out by C. Crestani. R. N. Zadoks and M. A. Holmes provided new sequenced data. Analysis was conducted by C. Crestani, with advice from R. N. Zadoks and T. L. Forde. The chapter was written by C. Crestani, and revised by R. N. Zadoks and T. L. Forde.

Chapter 5: A limited number of mobile genetic elements are associated with host adaptation in group B *Streptococcus*, GWAS reveals. The initial concept of the study was developed by C. Crestani, and R. N. Zadoks and T. L. Forde provided technical advice on improving the study. For data collection and generation, refer to chapter 4. Analysis was conducted by C. Crestani, with advice from R. N. Zadoks and T. L. Forde. The chapter was written by C. Crestani, and revised by R. N. Zadoks and T. L. Forde.

Chapter 6: Characterisation and identification of niche-associated genes of group B *Streptococcus* from camels. The initial concept of the study was developed by C. Crestani, and R. N. Zadoks and T. L. Forde provided technical advice on improving the study. Data collection and sequencing was carried out by D. Seligsohn (SVA). Analysis was conducted by C. Crestani. The chapter was written by C. Crestani, revised by R. N. Zadoks, T. L. Forde and approved by D. Seligsohn.

Abbreviations

AA	Amino Acid
AMR	Antimicrobial Resistance
att	attachment site
BAPS	Bayesian Analysis of Population Structure
BEAST	Bayesian Evolutionary Analysis Sampling Trees
β-h/c	β-haemolysin/cytolysin
BLAST	Basic Local Alignment Search Tool
BURST	Based Upon Related Sequence Types
CC	Clonal Complex(es)
CDC	Centers for Disease Control and Prevention
CIME	*cis*-Mobilisable Element(s)
CMT	California Mastitis Test
CPS	Capsular Polysaccharide
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
EOD	Early Onset Disease
ErmR	Erythromycin Resistance
GEI	Genomic Islands
GAS	Group A *Streptococcus*
GBS	Group B *Streptococcus*
GTR	General Time-Reversible model
GWAS	Genome-Wide Association Study
HGT	Horizontal Gene Transfer
IAP	Intrapartum Antibiotic Prophylaxis
ICE	Integrative and Conjugative Element(s)
ID	Identity

IEP	Intron-Encoded Protein
IME	Integrative and Mobilisable Element(s)
IS	Insertion Sequence(s)
LOD	Late Onset Disease
MAF	Minor Allele Frequency
MGE	Mobile Genetic Element(s)
MI	Mobile Integron(s)
ML	Maximum Likelihood
MLST	Multilocus Sequence Typing
MME	Minimal Mobile Element(s)
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NN	Nucleotide
NPV	Negative Predictive Value
NT	Non-Typeable
ORF	Open Reading Frame
PAI	Pathogenicity Island(s)
PICI	Phage-Inducible Chromosomal Island(s)
PFGE	Pulsed-Field Gel Electrophoresis
PPV	Positive Predictive Value
RAPD	Random Amplified Polymorphic DNA
RE	Restriction Enzymes
RMS	Restriction Modification Systems
QC	Query Coverage
SaPI	*Staphylococcus aureus* Pathogenicity Island(s)
SCC	Somatic Cell Count
SD	Standard Deviation
SEB	Staphylococcal Enterotoxin B
SLV	Single Locus Variant(s)
SNP	Single Nucleotide Polymorphism(s)
SSTI	Skin and Soft Tissue Infections
ST	Sequence Type(s)
TA	Toxin/Antitoxin

TcR	Tetracycline Resistance
TE	Transposable Elements
TSST1	Toxic Shock Syndrome Toxin 1
UTI	Urinary Tract Infections
WGS	Whole Genome Sequence
WSI	Wellcome Sanger Institute

Chapter 1

Introduction

1.1 Group B *Streptococcus* (GBS): an important multi-host pathogen

Group B *Streptococcus* (GBS), better known in the field of veterinary medicine as *S. agalactiae*, is an opportunistic bacterial pathogen with a wide spectrum of host species, ranging from humans, to cattle and fishes, which represent the three major host groups (Richards et al., 2011; Garcia et al., 2008; Wilkinson et al., 1973). A fourth host group, which has been gaining more attention in recent years, is that of camels (Seligsohn et al., 2020; Fischer et al., 2013; Younan & Bornstein, 2007). GBS is also occasionally isolated from cats and dogs (Yildirim et al., 2002b), sea mammals such as seals and dolphins (Delannoy et al., 2016; Evans et al., 2006), horses (Yildirim et al., 2002a), monkeys (Lämmler et al., 1998), reptiles such as crocodiles (Bishop et al., 2007), amphibians, notably frogs, (Elliott et al., 1990) and rodents (Hetzel et al., 2003; Elliott et al., 1990). Despite its wide range of host species, GBS has never been described in birds.

1.1.1 GBS in humans

In humans, GBS is a common commensal bacterium of the gastrointestinal and genitourinary tract. However, it is also the leading global cause of early and late onset neonatal invasive diseases (EOD: 0-6, LOD: 7-89 days after birth, respectively) (Seale et al., 2017; Centers for Disease Control and Prevention, 2005). These two clinical syndromes can result in neonatal

death or long-term disability and impairment of the newborn. In most EOD cases, the appearance of clinical signs, which are primarily ascribable to pneumonia and sepsis, usually occurs within the first 24h after birth (Melin, 2011). Compared to EOD cases, infants with LOD more commonly develop meningitis due to bacteremia. Within the gynaecologic and obstetric domains, GBS can also be responsible for maternal disease (estimated incidence of 0.38 cases per 1,000 pregnancies) (Hall et al., 2017) and stillbirths (57,000 cases/year worldwide) (Seale et al., 2017).

Despite the adoption in many countries of either risk-based or microbiological screening programs, the latter including administration of intrapartum antibiotic chemoprophylaxis (IAP) (Le Doare et al., 2017; Ohlsson & Shah, 2014), GBS is still the global leading cause of neonatal invasive disease since its emergence in humans in the 1960s (Fig. 1.1). Seale et al., 2017, estimated a global annual burden of 319,000 cases/year of neonatal invasive GBS disease, of which 90,000 resulted in death and 10,000 in neurodevelopmental impairment. The majority of cases were EOD (64%), the highest proportion of which were reported in Asia (30%), in particular India, and in Africa (26%), while developed countries accounted for only 3% of the overall cases (Seale et al., 2017). Nonetheless, GBS remains the primary cause of neonatal invasive infections even in high-income countries, right before *Escherichia coli* (Shane et al., 2017).

An important risk factor for EOD is maternal carriage of GBS in the genitourinary tract (estimated global prevalence among pregnant women of 17.9%) (Russell et al., 2017), as EOD is acquired vertically; transmission can occur both *in utero* or during birth through inhalation of contaminated maternal secretions (Melin, 2011). For LOD, the mode of transmission is still poorly understood (Mukhopadhyay & Puopolo, 2019), but it is attributed to horizontal transmission during the perinatal period; this can be from the mother, from the hospital, from other community sources, or, less commonly, from breastfeeding (Collin et al., 2019; Melin, 2011).

Even though a lot of attention within the GBS scientific community is focused on systemic infections in newborn babies, more than 50% of GBS deaths in the United States are reported among adults (High et al., 2005), which emphasises the importance of GBS in-

fections beyond the paediatric field. In addition to neonatal and maternal disease, GBS in adults can be responsible for skin and soft tissue infections (SSTI), urinary tract infections (UTI), bacteremia, osteomyelitis and, more rarely, meningitis, endocarditis and necrotising fasciitis (Lyhs et al., 2016; Le Doare & Heath, 2013; High et al., 2005). Such cases of GBS-associated disease have been most commonly observed among the elderly and adults with underlying medical conditions (Chaiwarith et al., 2011; Skoff et al., 2009; Farley et al., 1993). However, more recently there has also been a rise in cases of disease caused by GBS among immunocompetent non-pregnant adults (Lambertsen et al., 2010; High et al., 2005). Finally, GBS has also been implicated as a cause of invasive foodborne disease, with the first such outbreak reported in Singapore in 2015; this was attributed to consumption of contaminated raw fish (Kalimuddin et al., 2017; Tan et al., 2016; Rajendram et al., 2016). Foodborne GBS has since been shown to be caused by a hypervirulent clone (see subsection 1.2) that appears to be widespread in the South-East Asian region (Barkham et al., 2019), and which was also recently reported in Brazil (Leal et al., 2019).

1.1.2 GBS in dairy cattle

GBS does not cause invasive infections in bovines. It is a common cause of mastitis[1] in dairy cattle (Nocard & Mollereau, 1887), but it does not give rise to systemic disease, in contrast to other mastitis-causing pathogens - typically gram-negatives such as *E. coli* and *Klebsiella pneumoniae*. GBS usually causes subclinical infections, which can be hard to detect as they do not lead to changes in the appearance of the mammary gland or milk[2]. However, subclinical mastitis should not be underestimated, as it is the most frequent cause of mastitis (Forsbäck et al., 2009) and it can lead to significant economic losses due to reduction of milk quality and quantity (Gonçalves et al., 2018; Huijps et al., 2008). Only occasionally, GBS causes oedema of the mammary gland, pain and the appearance of fibrin clots in milk.

[1] Inflammation of the mammary gland.

[2] Subclinical mastitis can only be diagnosed if a somatic cell count (SCC) is performed, either in the laboratory or if an in-parlour automated sensor is available on farm. This test gives an estimate of the number of inflammatory cells and epithelial cells, which tend to increase during inflammatory processes, per ml of milk. Subclinical mastitis is characterised by >200,000 cells/ml, in the absence of clinical signs. Bacteriological culture tests may be either positive or negative.

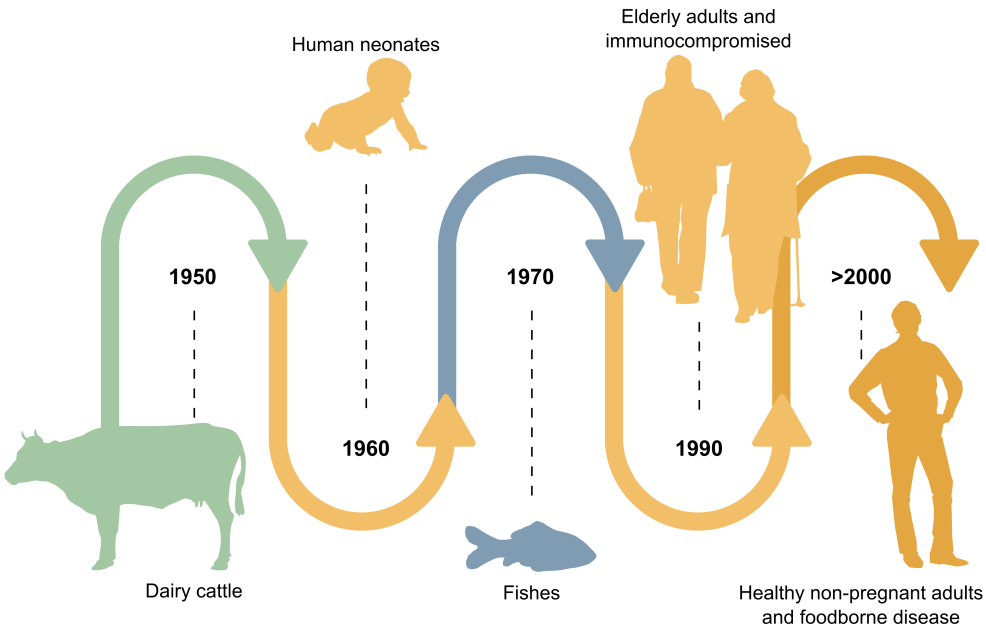


Figure 1.1: Timeline of recognition of group B *Streptococcus* (GBS) as a pathogen in the three major host groups. The first description of GBS as a mastitis-causing agent in dairy cows dates back to 1887 (Nocard & Mollereau, 1887), which precedes the Lancefield classification (Lancefield, 1933). GBS was well-known in veterinary medicine by the 1950s, and it became known among medical doctors during the 1960s and 1970s as a neonatal pathogen. Around the same period, the first reports of GBS diseases in fishes were published (Robinson & Meyer, 1966). During the 1990s and 2000s, the impact of GBS invasive infections was also recognised in elderly people and in non-pregnant adults, first among those with underlying medical conditions, and then also in otherwise-healthy individuals. In 2015, GBS was described for the first time as a food-borne pathogen, following an outbreak of invasive disease in Singapore linked to the consumption of raw fish.

Streptococcal species, including GBS, were the most common mastitis-causing agents together with staphylococci during the 1950s and 1960s. After the 1960s, their prevalence was drastically reduced in several high-income countries thanks to the introduction of mastitis control programs, which focused particularly on increasing hygiene of the milking process and removal of infected animals (through either treatment or culling) (Nielsen & Emanuelson, 2013; Neave et al., 1969). Different implementation of these measures at the local and national level led to variable changes in the spread of GBS worldwide: very low prevalences were reached in northern European countries such as Belgium (Piepers et al., 2007), Denmark (Andersen et al., 2003), the Netherlands (Sampimon et al., 2009), Norway (Østerås et

al., 2006), Sweden (Persson et al., 2011), Finland (Pitkälä et al., 2004) and the UK (Bradley et al., 2007), but also in some areas of North America, particularly in Canada (Riekerink et al., 2010). However, in recent years, re-emergence of GBS in dairy herds has been documented in Europe's Nordic countries, including Denmark, Norway, Finland and Sweden (Lyhs et al., 2016; Jørgensen et al., 2016; Katholm et al., 2012). In other countries, GBS did not reach such low prevalences. Examples in Europe are Germany (Tenhagen et al., 2006), Spain (Las Heras et al., 1999) and Italy (Zecconi & Zanirato, 2013); in the Americas, Brazil (Duarte et al., 2004), Colombia (Keefe et al., 2010), Uruguay (Gianneechini et al., 2002) and New York State (USA) (Wilson et al., 1997).

Two modes of transmission exist for GBS, the major one being contagious transmission. This occurs when a pathogen is spread from an infected cow to a healthy one during milking via contamination of the milking systems, or of milkers' hands and wiping towels (Zadoks et al., 2011; Keefe, 1997; Neave et al., 1969). This process by definition results in the spread of the same clone to several animals on the farm. The implementation of control measures such as the ones cited above can greatly reduce the prevalence of GBS on farm. The existence of a second transmission cycle, the environmental one, with gastrointestinal carriage and faecal shedding of GBS, has been recently demonstrated (Cobo-Ángel et al., 2018; Jørgensen et al., 2016; Manning et al., 2010). This mode of transmission, which is less known and potentially underestimated, could have an impact on elimination efforts, although the major problem with GBS control in most countries is lack of compliance with measures to prevent contagious transmission.

There are certain factors relative to the host which increase the risk of development of GBS mastitis. Older age and a higher number of pregnancies are among the factors linked with a higher frequency of mastitis from contagious pathogens, including GBS (Tenhagen et al., 2006), which probably just reflects a longer time at risk. In addition, the selection of certain breeds, such as the Holstein, over more traditional and rustic breeds, has led to a reduction of the genetic pool, which can exert a negative effect on the immune response (Zadoks & Fitzpatrick, 2009). Finally, factors relative to the environment (type of bedding and cleanliness) and general management (biosecurity measures, implementation of mastitis control programs, types of milking machines such as traditional systems or automated

milking systems) also play an important role in the risk of development of GBS infections.

1.1.3 GBS in fishes

GBS infections have been described in many fish[3] species, from farmed fishes, most notably tilapia (*Oreochromis* spp.), and wild fishes (Bowater et al., 2012; Jafar et al., 2008; Glibert et al., 2002), to ornamental (Delannoy et al., 2013; Ferguson et al., 1994) and pedicure fish species (Verner-Jeffreys et al., 2012). It can affect both bony (e.g. tilapia) and cartilaginous fishes (e.g. rays) (Bowater et al., 2012). Typically GBS causes invasive disease in these hosts, with the most common clinical signs being erratic swimming, loss of appetite, haemorrhages (gill, eye, opercula), corneal opacity, bi-lateral exophthalmia and abdominal swelling (Zamri-Saad et al., 2010; X. Y. Zhang et al., 2008)[4]. In aquaculture, one of the fastest growing animal-based food producing sectors (FAOSTAT, 2018), streptococcal infections represent a major economic threat (Amal & Zamri-Saad, 2011).

In farmed fish settings, outbreaks of GBS can have rates of morbidity and mortality up to 50% in severe acute infections and 70% accumulated mortality in chronic infections over several weeks (Yanong & Francis-Floyd, 2010). Fish streptococcosis in aquaculture has been reported in Central and South America (Barony et al., 2017; Asencios et al., 2016; Hernández et al., 2009), South and Southeast Asia (Rahman et al., 2021; Phuoc et al., 2021; Jantrakajorn et al., 2014), China (L. Liu et al., 2014; D. Zhang et al., 2013; Geng et al., 2012; Guo et al., 2012; Lu, 2010), Kuwait Bay (Jafar et al., 2008; Glibert et al., 2002) and West Africa (Verner-Jeffreys et al., 2018).

Newly-introduced carrier individuals represent the main risk factor for the dissemination of GBS, but vertical transmission has also been described (Zamri-Saad, 2018; Pradeep et al., 2016; Pereira et al., 2010). Summer months pose an increased risk for the development of GBS invasive disease[5], as the water temperature, fish density and stress are at their highest, and as the water quality drops substantially (Zamri-Saad et al., 2010).

[3] In the interest of simplification, the term 'fish' as a species is sometimes used throughout this thesis to refer to poikilotherm species, including multiple fish and frog species.

[4] Author's initials appear throughout this thesis when multiple authors with the same surname are included.

[5] Streptococcosis in fish is also referred to as 'summer streptococcosis'.

1.1.4 GBS in camels

The major clinical syndrome caused by GBS in camels (*Camelus dromedarius*) is mastitis, which, as in dairy cattle, is typically subclinical (Seligsohn et al., 2020). Considering the growing importance of camel milk production in some countries (see below), particularly in pastoralist settings (Elhadi et al., 2015), it is clear how GBS mastitis represents a threat to the livelihood of these communities. GBS has also been isolated from camels with chronic cough, wound infections, abscesses/peri-arthricular abscesses, gingivitis and vaginal discharge (Fischer et al., 2013; Zubair et al., 2013). A high prevalence of nasal carriage among healthy camels has also been described (Seligsohn et al., 2021b; Younan & Bornstein, 2007).

Reported prevalences of GBS mastitis vary substantially based on country and farming conditions (pastoralist, semi-pastoralist, ranch). GBS has been described in camels in East Africa, notably Kenya (Younan & Bornstein, 2007), Somalia (Fischer et al., 2013) and Ethiopia (Husein et al., 2013; Bekele & Molla, 2001), in the Middle East in the United Arab Emirates (El Tigani-Asil et al., 2020), and in Asia, in particular in India (Sena et al., 2001). In a pastoralist setting in Kenya, Seligsohn et al., 2020 reported a prevalence at the individual level of 32% GBS subclinical mastitis, while in a nearby region with a predominant ranching system, GBS mastitis was detected in only 11% of the sampled individuals (Seligsohn et al., 2021b).

The epidemiology and routes of transmission of GBS in dairy camels are still being elucidated. Similar to dairy cattle, low hygiene standards, in particular the absence of hand washing practices and poor water quality, which can be observed in low-resource settings (Seligsohn et al., 2020), could promote the horizontal spread of GBS during milking. In addition to the contagious spread of mammary-adapted clones (Seligsohn et al., 2021a), GBS strains typically isolated from the nose could represent a risk for the development of mastitis (Seligsohn et al., 2021b); however, environmental transmission is unlikely, considering that camels are mostly kept in an arid environment and that their faeces are dry, while transmission from skin and mucosa, which has been described for *Staphylococcus aureus* (Zadoks et al., 2002), seems more plausible. Other important risk factors for GBS mastitis in camels are the introduction of infected animals from other herds, as well as the intermixing of in-

dividuals belonging to different herds/owners (which is a common practice in pastoralist communities), and an older age of the animal (Seligsohn et al., 2020).

1.2 Microbiological and molecular characterisation of GBS

1.2.1 General microbiological characteristics of GBS

GBS is a gram-positive bacterium that belongs to the phylum of Firmicutes, Streptococcaceae family (De Vos et al., 2009; Glaser et al., 2002; Lancefield, 1933). Its name derives from the presence of the group-specific B antigen in the capsule, identified by Dr Rebecca Lancefield in the 1930s (De Vos et al., 2009; Glaser et al., 2002; Lancefield, 1933). In contrast to the sialylated capsular polysaccharide (CPS) (see subsection 1.3), the B antigen is common to all strains of the species as well as unique to the species. GBS cells are spherical or ovoid, 0.6-1.2 μm in diameter, often arranged in pairs (diplococci) or chains, the latter form appearing more evident when grown in broth culture (De Vos et al., 2009). It is virtually the only streptococcal species to include strains that appear pigmented when cultured on solid media (De Vos et al., 2009), thanks to a yellow-orange to red pigment recently named *granadene* (Six et al., 2015; Rosa-Fraile et al., 2014; Whidbey et al., 2013), from the Granada medium where this shade is observed. Originally thought to be a distinct molecule (M. Liu et al., 2018), this pigment corresponds to the β-haemolysin/cytolysin (β-h/c) (Six et al., 2015; Rosa-Fraile et al., 2014; Whidbey et al., 2013), a potent exotoxin that causes complete lysis of red blood cells. The β-h/c creates a clear zone of total haemolysis (β-haemolysis) around GBS colonies when cultured on blood agar (see subsection 1.3). Partially-haemolytic (α-haemolysis) and non-haemolytic (γ-haemolysis) strains are quite common in some groups of GBS, e.g. those derived from fish and belonging to the fish specific clonal complex CC552 (Delannoy et al., 2013).

Several methods for GBS typing, which have evolved in parallel with the development of new technologies, have been adopted throughout the years. Bacterial typing systems for GBS can be broadly divided into two main categories: serotyping and genotyping.

1.2.2 GBS serotyping

Serotyping techniques are traditionally based on antibody-antigen reactions with surface proteins of the CPS (e.g. latex agglutination or reactions with monospecific rabbit antisera) (Slotved et al., 2003; D. R. Johnson & Ferrieri, 1984), but serotypes can also be determined through PCR and real-time PCR (Breeding et al., 2016; Yao et al., 2013), or extracted from next generation sequencing (NGS) data (Metcalf et al., 2017; A. E. Sheppard et al., 2016). The CPS varies among GBS strains and allows distinction among ten different serotypes (Ia, Ib, II-IX) (Breeding et al., 2016; Le Doare & Heath, 2013; Glaser et al., 2002); non-typeable (NT) isolates are also reported, mostly in dairy cattle (Dogan et al., 2005). Serotype prevalence is different based on geographical origin, host species and host-specific characteristics such as age (Seale et al., 2017; Lyhs et al., 2016). In humans, serotypes Ia, Ib, II, III and V are responsible for most cases of GBS disease (Le Doare & Heath, 2013), with serotype III being the primary cause of neonatal invasive disease, accounting for 48% of EOD and 74% LOD (Seale et al., 2017). In recent years, serotype IV has been described as an emerging serotype in both neonates and adults in Europe (Lyhs et al., 2016; Florindo et al., 2014) and North America (Teatero, Athey, et al., 2015; Diedrick et al., 2010). In non-pregnant adults, serotypes Ia, IV and V are the most common (Edwards et al., 2016; Dogan et al., 2005; High et al., 2005). In dairy cattle, serotypes Ia, II and III occur with higher frequencies (Hernandez et al., 2021; Lyhs et al., 2016; Duarte et al., 2005); among these it is worth highlighting how, in the global GBS population, serotype II belongs primarily to a bovine-specific lineage. In fishes, three serotypes have been described: Ia, which is found in a primarily fish-associated lineage, Ib, which is associated to a fish-specific lineage, and III, which is found in a lineage that also affects humans (see below) (Ong et al., 2018; Chideroli et al., 2017; Delannoy et al., 2013).

A vaccine against GBS in humans is still not available, but CPS conjugate vaccines that target different serotypes are currently being developed by various pharmaceutical companies and academic groups (Kobayashi et al., 2019; Heath, 2016). Vaccine development currently focuses on five main serotypes (Ia, Ib, II, III, V) (Kobayashi et al., 2019), which are in accordance with the most common types reported in neonates and pregnant women (Seale et al., 2017). As in humans, there are no commercial vaccines to prevent GBS infec-

tions available in animals at the moment. Despite the need for immunisation, particularly in farmed fish (Munang'andu et al., 2016), most studies on vaccine development for the veterinary field are being performed by academic groups, rather than industry, working separately on different formulations. For example, in fish, injection of inactivated (Pretto-Giordano et al., 2010), live attenuated (L. P. Li et al., 2015) and protein vaccines (e.g. GapA protein) (Z. Zhang et al., 2017) has been carried out with variable levels of efficacy. However, injection of vaccines in aquaculture settings is not practical, and feed-based vaccination has also been attempted with good results (Zamri-Saad, 2018). In dairy cattle, studies have been carried out with first-generation vaccines, e.g. inactivated (Magaš et al., 2013), as well as with more modern types of vaccines, e.g. microspheres-based vaccine with encapsulated CAMP factor (G. Liu et al., 2017).

1.2.3 GBS genotyping

Bacterial genotyping methods include those based on DNA band patterns (comparative typing methods) and on DNA sequencing (library typing methods) (Zadoks & Schukken, 2006). The majority of band pattern methods are based on DNA digestion with restriction enzymes and migration of the fragments on a gel that is subjected to an electrical current, whereas some are based on PCR-amplification (Ochoa-Díaz et al., 2018); they offer a lower resolution compared to sequencing, and often suffer from poor reproducibility, thereby limiting options for isolate comparisons to be made between laboratories (i.e. there is no definitive classification that can be easily applied to different contexts). Some examples of comparative typing methods are: Pulsed-Field Gel Electrophoresis (PFGE) and Random Amplified Polymorphic DNA (RAPD) (the latter is based on PCR amplification, rather than restriction). Both RAPD (Zhao et al., 2006; Sukhnanand et al., 2005; Martinez et al., 2000) and PFGE (Duarte et al., 2005) have been used in the past to compare human and bovine GBS populations, reporting a clear distinction between the two, which was later disproved.

DNA sequence typing methods include lower-resolution methods, which are based on the determination of the nucleotide sequence of a selected number of highly conserved genes, and methods with a higher discriminatory power, which involve the sequencing of large

sections of the genome[6]. Among the former group, the most widely used is MultiLocus Sequence Typing (MLST) (Maiden, 2006), which was originally based on the amplification and sequencing of a few housekeeping genes (usually seven), to determine allelic profiles. Sequence types (ST) are now routinely extracted from NGS data. Some ST are uniquely detected in one host species, such as ST260/261 and ST552 in fishes, ST61/67 in cattle and ST17 in humans (a serotype III strain that is responsible for most neonatal infections). Other ST are shared between hosts, an example in humans and fish being ST283, a serotype III hypervirulent clone responsible for foodborne infections linked to the consumption of raw fish (Kalimuddin et al., 2017; Ong et al., 2018). Shared ST between humans and dairy cattle have also been observed in studies that have sampled animals together with their herds-persons (Cobo-Ángel et al., 2019; Sørensen et al., 2019) or that included sympatric and contemporaneous isolates (Fig 1.2) (Lyhs et al., 2016), which supports the possibility of inter-species transmission. ST can be classified in groups of related isolates that likely derive from a common ancestor, the so-called clonal complexes (CC) (Pavón & Maiden, 2009). This central allelic profile, or ancestral genotype, is usually assigned with the BURST (Based Upon Related Sequence Types) algorithm; the eBURST program (Feil et al., 2004) is able to assign ST into groups according to user-defined criteria of different alleles in common to at least one other member of the group. An example in GBS is CC552 from fish, which comprises ST260/261 and ST552, along with their single locus variants (SLV) (Delannoy et al., 2013). A major CC is that of CC1, a very diverse lineage which comprises several ST. Although ST and CC represent universal classification methods that can be applied in different contexts and still be comparable, they do not offer high discriminatory power. For example, phylogenetic analyses carried out from MLST data in GBS led to the false conclusion that the neonatal-associated hypervirulent clone ST17 derived from a bovine-specific lineage (CC61/67) (Héry-Arnaud et al., 2007; Bisharat et al., 2004), which was later disproved (Sørensen et al., 2010). With the advancements of NGS, and the possibility to reconstruct phylogenetic trees from the whole repertoire of core genes to identify clusters of related isolates, MLST became secondary, although this nomenclature is often included in WGS

[6]Modern DNA sequencing obtained with Next Generation Sequencing (NGS) technologies is often improperly referred to as Whole Genome Sequencing (WGS), although in many cases sequencing of an isolate results only in a partial genome sequence divided into large fragments, known as contigs.

phylogenies to help scientists identify lineages they are already familiar with. In addition to core genome phylogenies, clustering algorithms such as Bayesian Analysis of Population Structure (BAPS) (Corander et al., 2003; Corander & Marttinen, 2006; Corander et al., 2008) and hierBAPS (Cheng et al., 2013) provide alternative methods of genotyping for the identification of groups of similar sequences.

Another PCR-based genotyping method that was shown to be strongly correlated with both serotype and ST in *Salmonella* spp. (Fabre et al., 2012; F. Liu et al., 2011) is CRISPR typing (clustered regularly interspaced short palindromic repeats) (Barrangou & Dudley, 2016). CRISPR are discussed in more detail in section 1.4.

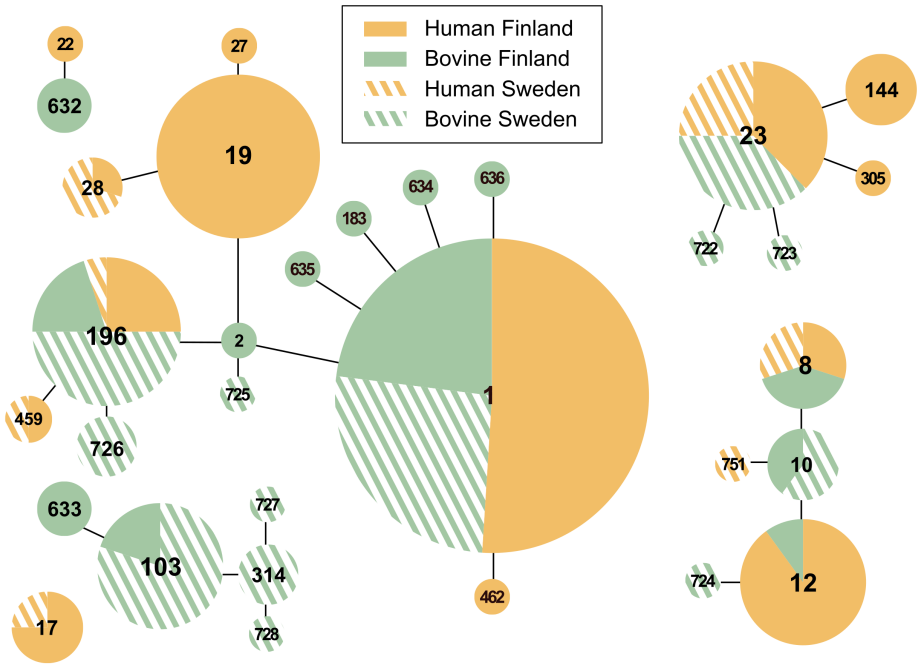


Figure 1.2: Distribution of group B *Streptococcus* (GBS) isolates from people and cattle in two countries (Finland and Sweden) across clusters of different sequence types (ST). Clusters include single- and double-locus variants (connected by black lines). Each circle represents an ST, with size of the circle and its coloured segments proportional to the number and origin of isolates, respectively. Figure has been adapted from Lyhs et al., 2016, with permission from the journal.

1.3 Virulence factors

Several virulence factors have been identified in GBS over the years (Lin et al., 2018; Maisey et al., 2008; Herbert et al., 2004a; Tettelin et al., 2002). Particularly, they have been extensively studied and well-characterised in isolates of human origin, even though their relative contribution to disease development is still not fully understood. Based on their function, virulence genes can be classified into genes that have a role in:

1. Regulation

   GBS encode a number of regulatory genes that are known to contribute to virulence (Herbert et al., 2004a). One of the most well-characterised, the two component system CovS/CovR, has been shown to play a role in regulating haemolytic activity (S. M. Jiang et al., 2005; Lamy et al., 2004) and adherence to epithelial cells (Patras et al., 2013; Lembo et al., 2010).

2. Adherence

   Adhesion to the epithelial surface is necessary for colonisation of the epithelial cells of the human vagina. Among the wide variety of adhesins encoded by GBS are: fibronectin-binding proteins (*pavA* and *scpB*, also known as C5a peptidase), fibrinogen binding proteins (*fbsA, fbsB*), laminin-binding protein (*lmb*) and lipotheicoic acid (Maisey et al., 2008; Herbert et al., 2004a; Tettelin et al., 2002). In addition to these, an ST17-specific surface-anchored protein, the hypervirulent GBS adhesin (*hvgA*), is a critical virulence gene in neonatal infections, particularly for meningeal tropism of these strains (Tazi et al., 2010).

   ScpB and *lmb* are variably present in GBS among different host species: they are present in a large proportion of human strains (Morach et al., 2018), whereas they are lacking from most bovine (Morach et al., 2018; Rato et al., 2013) and fish isolates (Morach et al., 2018; Delannoy et al., 2016; Kayansamruaj et al., 2014). Additionally, the expression of *scpB* is solely induced by human serum, but not by bovine serum (Gleich-Theurer et al., 2009).

   Also playing a role in attachment to host cells are *pili*, encoded by the pilin gene

clusters (PI-1, PI-2a, PI-2b) (Rosini et al., 2006; Lauer et al., 2005); their prevalence is variable among human, bovine and fish genomes (Morach et al., 2018; Delannoy et al., 2016; Kayansamruaj et al., 2014).

3. Invasion

   Invasion of the host is promoted by toxins and cell-surface proteins (Maisey et al., 2008; Herbert et al., 2004a; Tettelin et al., 2002). Among toxins, the β-h/c pigment (Rodriguez-Granger et al., 2015), encoded by the *cyl* locus, is regarded as one of the main factors that lead to the development of human invasive disease, although its production is not absolutely necessary to establish systemic infections (Gendrin et al., 2017; Six et al., 2015). The prevalence of non-haemolytic non-pigmented strains among the human population is approximately 5-8% (Nickmans et al., 2012). However, it is hard to have an accurate estimate: since human non-haemolytic strains have long been considered avirulent, most routine diagnostic procedures exclude them. In contrast to GBS from humans, a high proportion of fish and cattle pathogenic strains are non-haemolytic (Delannoy et al., 2016; Lusiastuti et al., 2013; Ebrahimi et al., 2013).

   Also important among toxins are the exfoliative toxin A and the CAMP factor (*cfb*). The latter forms pores in the host cells and it is present in almost all pathogenic isolates across different host species (Brochet et al., 2006; Herbert et al., 2004a), except for fish isolates from lineage CC552 (Bowater et al., 2012; Evans et al., 2006). In rare cases, human bovine GBS isolates tested negative to the CAMP phenotypic test and to PCR of *cfb* (Kong et al., 2002; Hassan et al., 2000; Podbielski et al., 1994). The CAMP factor test has long been used for GBS species confirmation, together with other biochemical tests (e.g. bile esculin negativity) (Darling, 1975; Munch-Petersen et al., 1945).

   Several cell-surface secreted proteins such as peptidases, proteases, collagenases, nucleases and amidases have been described as virulence factors in human GBS isolates (Herbert et al., 2004a; Tettelin et al., 2002). Among these, the hyaluronate lyase (*hylB*) is able to degrade certain components of the extracellular matrix and is believed to

contribute significantly to invasion (Maisey et al., 2008; Herbert et al., 2004a; Tettelin et al., 2002; Glaser et al., 2002; Rolland et al., 1999), although it is not necessary (Domelier et al., 2006). *HylB* also occurs with high frequency among bovine (Sukhnanand et al., 2005) and piscine isolates (Delannoy et al., 2016; Kayansamruaj et al., 2014; Godoy et al., 2013).

4. Immune response evasion

   In humans, the CPS is of great importance for limiting complement deposition and phagocytosis and it is also able to mask surface proteins and therefore avoid stimulation of the host immune response (Maisey et al., 2008; Herbert et al., 2004a). In a bovine-specific lineage (CC61/67), pseudogenisation of genes in the *cps* locus has been described (Almeida et al., 2016); therefore the CPS is thought to be less important in the establishment of infections in dairy cattle.

   Immunoprotective surface proteins also play a role in immune response evasion (Herbert et al., 2004a; Tettelin et al., 2002). One example is the Alpha-like protein Rib (*rib*), (Brochet et al., 2006; Lachenauer et al., 2000), which is absent from fish strains (Morach et al., 2018; Delannoy et al., 2016; Rosinski-Chupin et al., 2013) and present in approximately 50% of human and 25% of cattle isolates (Morach et al., 2018). The Rib protein is among the candidate protein-based vaccine targets, together with other surface proteins (Kobayashi et al., 2019).

1.4 The mobilome

1.4.1 The importance of mobile genetic elements in GBS

In GBS, a number of virulence genes, especially those involved in pathogen-host interaction, are associated with confirmed or putative mobile genetic elements (MGE) (Brochet et al., 2006; Herbert et al., 2004a). Examples of this are the *scpB* and *lmb* genes, which are usually co-located on a transposon (Kayansamruaj et al., 2014; Franken et al., 2004, 2001), the pilin gene clusters (Rosini et al., 2006) and the Alpha-like protein Rib (*rib*) (Brochet et al., 2006; Lachenauer et al., 2000). The CPS is also located on a putative MGE, with capsular

switching events being well-documented in GBS (Neemuchwala et al., 2016; Bellais et al., 2012; Martins et al., 2010).

The repertoire of all MGE encoded in a genome is known as the 'mobilome' (Rouli et al., 2015; Siefert, 2009; Frost et al., 2005), which is part of the accessory or dispensable genome. The accessory genome is composed of the genes present in some but not all the members of a bacterial species, as opposed to the core genome, which is the genetic fraction shared by all the sequenced genomes of a given species (Soucy et al., 2015; Tettelin et al., 2008, 2005; Medini et al., 2005). All the known accessory and core genes of a bacterial species together form the pangenome, from the Greek 'pan' ('$\pi\alpha\nu$') which means 'whole'. The more strains that are sequenced and added to the analysis, the wider the accessory genome and the smaller the core genome will be, until a predicted plateau is reached (Tettelin et al., 2005; Medini et al., 2005). In bacterial species with an open pangenome, like GBS (Tettelin et al., 2008; Medini et al., 2005), the pangenome size increases indefinitely when a new sequence is added to the database. By contrast, closed pangenomes are quickly saturated with the entire spectrum of genes present in the species (Rouli et al., 2015; Tettelin et al., 2008; Medini et al., 2005).

With the remarkable amount of genomic data available nowadays thanks to NGS technologies, it has become more evident how horizontal gene transfer (HGT) plays a pivotal role in the evolution and niche[7] adaptation of bacteria. HGT is defined as the lateral transmission of genetic material between both closely and distantly-related bacteria (P. H. Oliveira et al., 2017; C. M. Johnson & Grossman, 2015; Siefert, 2009; Wozniak et al., 2009; Thomas & Nielsen, 2005; Boucher et al., 2003; Hilario & Gogarten, 1993). It is strongly believed that HGT, through the mobilome, plays a significant role in GBS species evolution and diversity, contributing to host and niche adaptation, strain virulence and antibiotic resistance (Richards et al., 2019; Tettelin et al., 2008; Davies et al., 2005). Strains that share the same niche or host environment with other bacterial species, either commensal or pathogenic, can

[7] In ecology, the term 'niche' is used to indicate the role of an organism in an ecosystem and in particular an environment with which the organism is associated. 'Niche' is a broad term that can indicate both a host species (host-adaptation) or a particular organ/tissue within that host (tissue tropism) to which a bacterial species is adapted.

acquire fragments of DNA that contain useful genes for survival and adaptation, such as new metabolic pathways and transporters for various substrates (S. K. Sheppard et al., 2018; Davies et al., 2005). As an example, a remarkable difference between GBS from dairy cattle and from humans is the presence of the Lac.2 operon in most bovine isolates and its absence from most human isolates (Fig. 1.3) (Lyhs et al., 2016; Richards et al., 2013; Ferretti et al., 2001). Human isolates often encode only the Lac.1 operon, whereas bovine isolates have both Lac.1 and Lac.2 operons (Ferretti et al., 2001). The latter is believed to be the major factor responsible for host and niche adaptation of bovine strains to the mammary gland (Lyhs et al., 2016; Richards et al., 2013). Because Lac.2 promotes the metabolism of lactose (Lac.2+ isolates show phenotypic lactose fermentation (Lyhs et al., 2016)), isolates carrying Lac.2 will have a fitness advantage when in a lactose-rich environment (e.g. the bovine udder). Lac.2 is carried by a putative integrative conjugative element (ICE), showing signatures of mobility (e.g. integrase gene) (Lyhs et al., 2016; Richards et al., 2011), and it

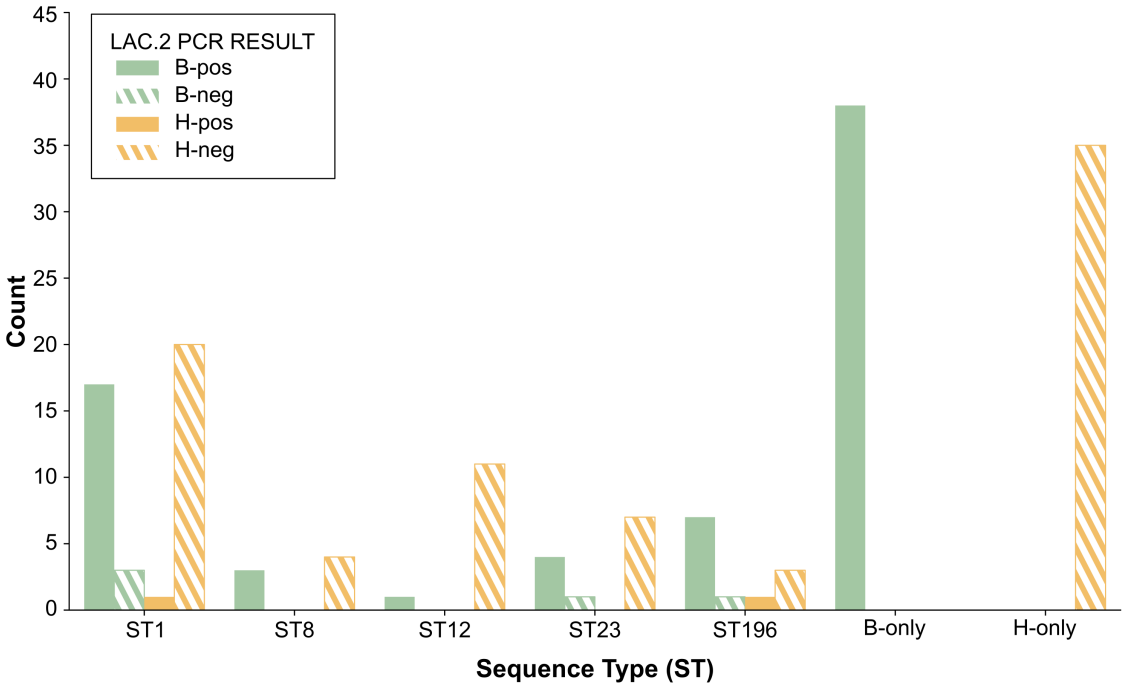


Figure 1.3: Distribution of Lac.2 (*lacEFG* genes) PCR–positive (pos) and negative (neg) human (H) and bovine (B) group B *Streptococcus* (GBS) isolates across sequence types (ST). ST found in both host species (host-generalist lineages) are shown individually, whereas ST that are found in a single species (host specialist lineages) are grouped by species. Figure has been adapted from Lyhs et al., 2016, with permission from the journal.

is found in other streptococcal species that are important mastitis-causing agents: *Streptococcus dysgalactiae* subsp. *dysgalactiae* and *Streptococcus uberis*. Cross-bacterial species genetic transfer mediated by MGE is very common between streptococci that share the same niche (e.g. between GBS and *Streptococcus pyogenes* in the human oropharynx) and it is thought to play an important role in their evolution and pathogenicity (Davies et al., 2005). Recombination of horizontally-acquired DNA and MGE is an important driving evolutionary force in the bacterial universe in general (Siefert, 2009; Guttman & Dykhuizen, 1994) and among streptococcal species in particular (Lefébure & Stanhope, 2007; Brochet et al., 2006). There are two types of DNA recombination mechanisms: homologous, which is the substitution of DNA segments that share high sequence similarity, and non-homologous, which occurs between DNA sequences that do not share sequence similarity (e.g. integration in the chromosome of MGE that encode an integrase gene) (Frost et al., 2005). Homologous recombination in the core genes poses a challenge to phylogenetic reconstruction because it can affect large sections of the chromosome and introduce several new genes or single nucleotide polymorphisms (SNP) in a single event. A 'net-like model' of evolution, as opposed to the classical 'tree-like' Darwinian model, can therefore be a more accurate way of displaying relationships among bacteria (Olendzenski & Gogarten, 2009; Kunin et al., 2005; Bapteste et al., 2004; Hilario & Gogarten, 1993; Spencer, 1864); this is true especially for those species that frequently undergo recombination, like streptococci, including GBS (Sørensen et al., 2010; Lefébure & Stanhope, 2007; Brochet et al., 2006).

Although the flow of MGE between bacterial cells can promote the acquisition of useful genes for adaptation to specific conditions (S. K. Sheppard et al., 2018), this can sometimes be detrimental (e.g. virulent bacteriophages that lead to lysis of the cell). Therefore, bacteria have evolved different mechanisms to protect themselves from invading DNA and MGE. These include restriction modification systems (RMS), which are rudimentary immune systems that cleave unmethylated alien DNA (Rodic et al., 2017; Ershova et al., 2015), and CRISPR, which are more sophisticated systems of adaptive immunity that modulate integrated MGE by cleaving their double stranded DNA (Barrangou & Dudley, 2016). Both these systems are able to shape bacterial chromosomes thanks to the regulation of their mobilomes. In particular, CRISPR loci are composed of three main elements: direct repeated se-

quences, spacer sequences (which are responsible for recognising the protospacer, the DNA target region that is cleaved by CRISPR), and an AT-rich leader sequence (which contains the promoter for the CRISPR locus) (Karimi et al., 2018; Lopez-Sanchez et al., 2012). Bacteria can acquire new spacer sequences over time, and different isolates will have different CRISPR profiles (recently used for genotyping) based on the DNA invaders they have come in contact with (acquired immunity). In GBS, analysis of the ubiquitous CRISPR1 locus showed a remarkable diversity in the repertoire of MGE within the GBS population (Lopez-Sanchez et al., 2012), confirming the importance of MGE in this bacterium.

Genes can be laterally exchanged between bacterial cells via three mechanisms: transduction, conjugation and transformation (Tab. 1.1) (P. H. Oliveira et al., 2017; Soucy et al., 2015; C. M. Johnson & Grossman, 2015; Olendzenski & Gogarten, 2009; Frost et al., 2005). The latter, which entails the uptake of exogenous 'naked' DNA from the environment, was the first one to be discovered (Frost et al., 2005), but it is not as frequent as the first two, which involve the activity of several MGE (C. M. Johnson & Grossman, 2015). MGE are described as any type of DNA that can move within a genome (intracellular mobility) or between genomes (intercellular mobility) (Bellanger et al., 2014; Siefert, 2009; Frost et al., 2005; Toussaint & Merlin, 2002); different MGE propagate between cells via different mechanisms, based on their distinct nature (Tab. 1.1) (Siefert, 2009; Frost et al., 2005).

1.4.2 MGE transferred via transduction

Bacteriophages, which are viruses that infect bacterial cells, package their DNA in viral particles that transfer their genetic material to a new host cell through transduction. Phages range in size from 5 to 500 kbp (Siefert, 2009; Frost et al., 2005) and are highly bacterial species-specific, even though phages infecting different bacterial species can show a certain degree of similarity. They have a well conserved genomic organisation, with a modular structure (Frost et al., 2005; Iandolo et al., 2002). Phage predation can affect bacterial evolution and fitness in different ways: when a temperate bacteriophage, which is a virus that does not immediately lyse its host after infection and production of viral particles, injects its DNA into a bacterial cell, a state of site-specific integration is established, called 'lysogeny', in which the phage co-exists with the host. 'Lysogenic conversion' is achieved when an

Table 1.1: Description of the different mechanisms of horizontal gene transfer (HGT) and the mobile genetic elements (MGE) that propagate via each of them. A specific example in group B *Streptococcus* (GBS) is given.

Mechanism	Description	MGE	Example in GBS
Transformation	Uptake of free DNA from the environment	MME	locus 1 and 8 (Delannoy et al., 2016)
Transduction	Injection of DNA mediated by a virus	Bacteriophages	Prophages A-F (van der Mee-Marquet et al., 2018)
		PICI	PICI1 and PICI2 (Crestani et al., 2020)
Conjugation	Transfer of DNA through direct contact between cells	Plasmids	pCCH208 *erm*(T) (Compain et al., 2014)
		ICE	ICE Tn*916* *tet*(M) and ICE Tn*5801* *tet*(M) (Da Cunha et al., 2014)
Mobilisation	Elements not capable of intercellular self-transfer	IS	IS*1458* (Domelier et al., 2006)
		Transposons	Many misclassified (see ICE)
	Insertion and carriage by other elements	Introns	GBSi1 (Domelier et al., 2006)

integrated dormant phage confers a new phenotype through the introduction of new fitness factors and/or the disruption of host genes (Siefert, 2009; Domelier et al., 2006; Frost et al., 2005; Bossi et al., 2003; Hendrix et al., 1999). When an integrated prophage excises, random pieces of host DNA can be incorporated into phage particles during cell lysis (generalised transduction). Another possibility is that host DNA regions flanking the prophage can be packaged (specialised transduction) (P. H. Oliveira et al., 2017; Soucy et al., 2015; Frost et al., 2005). Recently, another mechanism has been discovered and named lateral transduction: prophages have been shown to package host DNA to at least +300 kbp from their integration sites, with massive impact on bacterial recombination events (Chen et al., 2018). All of these mechanisms are key to HGT, as they promote the exchange of genes between diverse strains. In human GBS, a higher prevalence of prophages has been associated

with greater virulence, particularly in the ability to cause invasive infections (Salloum et al., 2011, 2010; Domelier et al., 2009; van der Mee-Marquet et al., 2006); a number of these phages carry genes associated with virulence, toxicity and host adaptation, suggesting that lysogeny might play an important role in the biological success of the bacterial host (van der Mee-Marquet et al., 2018).

Transduction is also used by phage-inducible chromosomal islands (PICI), small MGE (12-16 kbp) that hijack phages' packaging systems in order to be transferred to a new cell (Martínez-Rubio et al., 2017; Penadés & Christie, 2015; Novick et al., 2010). PICI are extremely successful, highly specialised molecular parasites with a very well conserved structure that can carry important virulence genes and toxins (Martínez-Rubio et al., 2017; Penadés & Christie, 2015; Novick et al., 2010). As an example, in *S. aureus*, pathogenicity islands (SaPI) can encode the toxic shock syndrome toxin 1 (TSST1) or the staphylococcal enterotoxin B (SEB). PICI are widespread in the bacterial world, as they have been described in both gram-positive and gram-negative bacteria (Fillol-Salom et al., 2018; Martínez-Rubio et al., 2017; Novick et al., 2010). However, their presence and importance in GBS is unknown, as no previous studies have focused on their detection in this species.

1.4.3 MGE transferred via conjugation

Unlike transformation and transduction, conjugation takes place when two cells make contact and the genetic material travels from the donor to the recipient cell through a specialised transfer pore (Soucy et al., 2015; Frost et al., 2005). This transfer mechanism can be mediated by conjugative MGE such as conjugative plasmids or ICE (see subsection 1.4.5).

Plasmids are extra-chromosomal DNA molecules, ranging in size from 1-400 kbp. Their genes are organised into a stable structure capable of replicating independently from the host chromosome (replicon) (Siefert, 2009; Frost et al., 2005). Plasmids can be differentiated into conjugative and non-conjugative. In the former, a gene cluster known as transfer locus (*tra*) promotes the contact between donor and recipient cell by means of the *sex pilus*, through which genetic material is exchanged (Siefert, 2009; Carattoli, 2009; Frost et al., 2005); non-conjugative plasmids lack these systems, and they need the help of conjugative plasmids to

transfer to a new host cell (John et al., 1981). Plasmids can be classified using 'functional groups' into fertility plasmids, resistance plasmids, degradative plasmids (which metabolise unusual compounds) and Col-plasmids (which can kill other bacteria). Additionally, they can be classified based on 'incompatibility groups' (Inc plasmids); incompatible plasmids are usually related to each other and unable to co-exist within the same host cell, as they share the same replication mechanism (Siefert, 2009; Carattoli, 2009; Frost et al., 2005; DeNap & Hergenrother, 2005; Datta & Hedges, 1971). Plasmids can be present in circular free-form within the cytoplasm or they can integrate into the bacterial chromosome. In the latter case, they promote recombination events and genetic transfer of parts of the host chromosome, as well as of other MGE like insertion sequences (IS) and transposons (see subsection 1.4.4). Because of their mobile nature, plasmids do not usually encode any genes essential for cell survival (Siefert, 2009; Carattoli, 2009; Frost et al., 2005); hence, they can be easily lost unless they provide a genetic advantage over competing strains, like carriage of antimicrobial resistance (AMR) genes (Carattoli, 2013, 2009). In GBS, plasmids are rarely reported, and their prevalence among human isolates is low (Compain et al., 2014). Prior to this PhD project, no studies had focused on the detection and on the prevalence of plasmids in animal GBS isolates. Most GBS plasmids reported in the literature, which are all of human origin, carry AMR genes for erythromycin, chloramphenicol and gentamicin (Sendi et al., 2016; Compain et al., 2014; DiPersio et al., 2011; Horodniceanu et al., 1976).

1.4.4 MGE mobilised by other MGE

Some types of MGE can move within a genome but are not capable of self-transfer to a new host cell; thus, they require other MGE to be mobilised. This is the case for transposable elements (TE) (C. M. Johnson & Grossman, 2015), which include transposons, IS, mobile integrons (MI) and introns (C. M. Johnson & Grossman, 2015; Frost et al., 2005). Both transposons and IS are able to move or copy themselves ('cut-and-paste' and 'copy-and-paste' mechanisms, respectively) into a new site of the genome, with no DNA homology required and variable site-specificity (C. M. Johnson & Grossman, 2015; Curcio & Derbyshire, 2003). Different types of transposons can be identified based on how they move (Siefert, 2009; Curcio & Derbyshire, 2003):

1. Retrotransposons, which invade the genome in RNA form that is transcribed back to DNA by reverse transcription;
2. DNA transposons (or type II transposons or DDE-transposons), which encode a DDE-transposase;
3. DNA serine-transposons;
4. DNA tyrosine-transposons.

The latter two encode for site-specific recombination and are often referred to as 'conjugative transposons', but have recently been reclassified as ICE (C. M. Johnson & Grossman, 2015; Siguier et al., 2014; Wozniak et al., 2009; Frost et al., 2005; Curcio & Derbyshire, 2003; Burrus et al., 2002) (see subsection 1.4.5).

IS are the simplest, smallest and most abundant type of autonomous TE. They can vary in terms of genetic organisation and target sequences of integration (different degree of site-specificity) (Siguier et al., 2014). Transposons and IS can have a remarkable impact on their host genome, thanks to a number of different mechanisms: they can act as vectors of virulence genes, they can reorganise host genes when transposing flanking DNA sequences, and they can silence or activate genes, based on whether the integration site is located within or upstream the gene, respectively (Siguier et al., 2014; Curcio & Derbyshire, 2003). This latter characteristic, which causes gene disruption, can lead to an increased pathogenicity of the strains, as has been described in various bacterial species, including GBS (Domelier et al., 2006; Héry-Arnaud et al., 2005; Granlund et al., 2001; Spellerberg et al., 2000; Rolland et al., 1999). As an example, the presence of a copy of the IS*1458* within the *hylB* (hyaluronate lyase) gene is one of the markers associated with high risk of neonatal meningitis (Domelier et al., 2006).

TE also include mobile integrons (MI) and introns. MI, first discovered on conjugative plasmids, are made up of a tyrosine site-specific recombinase and one or consecutive gene cassette arrays; these, in gram-negative bacteria, often carry AMR genes (Cambray et al., 2010; Siefert, 2009; Domelier et al., 2006). Introns can be divided in group I and II introns. Mobile group II introns are transposable retroelements which consist of a highly organised

catalytic RNA and a multifunctional intron-encoded protein (IEP) (Siefert, 2009; Lambowitz & Zimmerly, 2004), which also acts as a retrotranscriptase. HGT of group II introns is cross-specific and conjugative (Siefert, 2009; Lambowitz & Zimmerly, 2004; Belhocine et al., 2004). As for IS, the site of integration of introns plays a role in gene expression and virulence; as an example, the integration of GBSi1 group II intron downstream the *scpB* gene in GBS is a marker of pathogenicity (Domelier et al., 2006). Both group I and II introns can encode homing endonucleases, nicking enzymes that recognise long (12-40 bp) cutting sequences and that can contribute in shaping the host genome (Siefert, 2009).

1.4.5 Non-conventional MGE

In addition to traditional classes of MGE, such as plasmids and prophages, other types of MGE, often hard to classify, have been described over the years. These include PICI, as well as genomic islands (GEI), comprising both full-length GEI (>10 kbp) and genomic islets (<10 kbp). GEI are a diverse group of elements consisting of chromosomal segments acquired by HGT; a high number of these segments are probably derived from consecutive HGT events with contiguous integration of different elements, which shape GEI's modular structure (Bellanger et al., 2014; Guglielmini et al., 2011). Among GEI, a group of self-conjugative, self-integrative elements, the ICE, has been defined (Burrus et al., 2002).

ICE are a diversified class of MGE with a size range of ∼20-500 kbp (C. M. Johnson & Grossman, 2015). They mostly reside within the host chromosome, occasionally excising for conjugation and transfer to a new cell (Wozniak et al., 2009; Frost et al., 2005). Although ICE can differ significantly among themselves, they all share the same modular structure composed of integration, excision, conjugation and regulation genes (Ambroset et al., 2016; Frost et al., 2005). Moreover, they all encode a type IV secretion system, used for conjugation (C. M. Johnson & Grossman, 2015; Wozniak et al., 2009). ICE have a wide spectrum of hosts and are unable to replicate autonomously (Wozniak et al., 2009). ICE can mobilise contiguous ICE, transposons, plasmids, segments of chromosomal DNA and other non-conjugative elements; among the latter we can find integrative and mobilisable elements (IME), which retain a functional integrase but no conjugation system, and *cis*-mobilisable elements (CIME), which lack both but include *attL* and *attR* sequences (left and right attach-

ment sites, respectively; see chapter 2) (C. M. Johnson & Grossman, 2015; Bellanger et al., 2014; Wozniak et al., 2009; Burrus et al., 2002).

Many ICE also carry genes for AMR (C. M. Johnson & Grossman, 2015; Klima et al., 2013; Mata et al., 2011; Michael et al., 2011; Wozniak & Waldor, 2010). These have mostly been reported among human isolates and carry genes for erythromycin resistance (ErmR), such as Tn*3872* and ICE*sp2905* (Oppegaard et al., 2020), or multidrug resistance, such as ICE*Sa2603* (ErmR, some tetracycline resistance (TcR) and streptothricin) (Oppegaard et al., 2020), which has also been reported in bovine GBS (Huang et al., 2016), ICE*Sag37* (K. Zhou et al., 2017) and ICESag(RR1) (ErmR, TcR and aminoglycosides) (Campisi et al., 2016). Notably, ICE Tn*916* is responsible for TcR in human (Da Cunha et al., 2014), cattle (Crestani et al., 2021), fish (Barkham et al., 2019) and camel GBS (Fischer et al., 2013). ICE Tn*916* and Tn*5801*, both carrying the tetracycline resistance gene *tet*(M), are examples in GBS of the crucial role AMR ICE can play in the evolution of bacterial populations. A replacement of the GBS population infecting humans is thought to have occurred during the 20th century and to have been caused by the introduction and extensive use of tetracycline in the clinical practice from 1948 onward; this likely led to the selection[8] of a few highly human-adapted clones that had previously acquired either of these two *tet*(M)-carrying ICE (estimated to have occurred around 1917-1935 for ST17) (Da Cunha et al., 2014). However, although the global spread of tetracycline resistant clones in the human population occurred in parallel with the first reports of neonatal invasive disease in the 1960s (Fig. 1.1), it does not explain the emergence of EOD and LOD (Da Cunha et al., 2014).

In the past, a number of MGE have been misclassified, especially as transposons, but are now considered part of the ICE group. Among these are Tn*916* and other Tn*916*-like elements, which are present in many different bacteria (Bellanger et al., 2014; A. P. Roberts & Mullany, 2009; Clewell et al., 1995). Other examples in GBS are Tn*GBS1* and Tn*GBS2* (Bellanger et al., 2014; Guérillot et al., 2013; Brochet et al., 2009), which however are still often referred to as 'conjugative transposons'.

[8]Selection, in biology, the preferential survival and reproduction or preferential elimination of individuals with certain genotypes, by means of natural or artificial controlling factors (Encyclopaedia Britannica, 2021).

Another type of non-conventional MGE is the minimal mobile element (MME) (Snyder et al., 2007; Saunders & Snyder, 2002). MME are thought to spread through natural transformation and integrate by homologous recombination, thus they do not require integrases, recombinases or conjugation machineries. MME probably originated as gene cassettes that, in rare events with little or no recombination, got inserted between well-conserved protein-encoding genes. This is the primary feature of MME, which are very efficiently mobilised thanks to homologous flanking regions that are shared among all the strains of one or more bacterial species (Snyder et al., 2007; Saunders & Snyder, 2002). MME have been described in *Neisseria* spp. and, more recently, in GBS fish strains (Delannoy et al., 2016).

1.5 Aim and objectives

The aim of this PhD project was to investigate and assess the role of the mobilome in GBS evolution and host-adaptation. As described above, MGE have been shown to impact on the ecological success, pathogenicity and host-switching events of multiple gram-positive and gram-negative pathogens. The mobilome is thought to have played a major role in GBS evolution (Lopez-Sanchez et al., 2012) and therefore knowing which types of MGE are found in GBS and the extent of their distribution across lineages is key to better understanding GBS as a pathogen.

The following specific objectives were covered:

1. Implementation and evaluation of existing methods for the detection of MGE in GBS to define preferred methods; development of a new typing and detection method for GBS prophages; development of an inventory of MGE in GBS isolated from across host species and lineages to facilitate subsequent analyses.

   This objective was addressed in chapter 2, in which I analysed two datasets of GBS genomes from multiple host species and countries for the presence of MGE (prophages, PICI, plasmids and ICE). Findings from this work were published in two papers: i) Richards et al., 2019, a large genomic study of GBS for which I contributed with the analysis of the presence of MGE (prophages, ICE, plasmids); ii) Crestani et al., 2020, a first-author paper in which I carried out an in-depth analysis of GBS prophages and in which I propose a new typing scheme and detection method for GBS prophages.

2. Analysis of the GBS population structure at a national level, and evaluation of the presence of known host-associated MGE to explain temporal lineage shifts.
   This objective was addressed in chapter 3, in which I analysed a dataset comprising both historical and contemporary GBS genomes (1953-1978) isolated from Swedish dairy cattle for population structure, and for the carriage of MGE and known molecular markers of host-adaptation. Findings from this work were published as a first-author paper in Microbial Genomics: Crestani et al., 2021.

3. Analysis of the global GBS population structure in terms of core and accessory genome content.
   In chapter 4, I assembled a large collection of GBS genomes from several countries and host species, and I developed a method for dataset curation and quality control to select a representative subset of the global GBS diversity. I applied methods for the analysis of population structure based on core and accessory genome, evaluation of recombination and distribution of RMS. The results were analysed in the context of ST/CC and previous knowledge on host-specialist[9] and generalist lineages.

4. Detection of host-associated accessory genome content and MGE with large-scale genome-wide association studies (GWAS).
   This objective was addressed in chapter 5, in which I carried out two GWAS on a vast, high-quality collection of genomes representative of the global GBS population (the same dataset used in chapter 4), in order to identify a comprehensive set of host-associated genes and MGE.

5. Analysis of the structure of the GBS population from camels, identification of camel-associated accessory genes with GWAS and detection of MGE.
   In chapter 6, I report the results from bioinformatic analyses of a GBS genomic collection from camels that was newly generated in collaboration with Dr Dinah Seligsohn and Dr Erika Chenais (National Veterinary Institute Sweden - SVA). This collaboration

[9]In this thesis, the terminology 'host-specialist' is used to indicate both host-associated lineages that are restricted to one host (i.e. uniquely found in one host group) and host-associated lineages that show predilection for one host (i.e. predominantly isolated from one host group, but that can occasionally occur in other hosts). Host-generalist lineages comprise strains that commonly affect multiple host groups.

resulted in the publication of two papers (Seligsohn et al., 2021a, 2021b), of which I am second author, and for which I contributed to genomic analyses including: MLST, serotyping, analysis of population structure, presence of molecular markers of host-adaptation and AMR genes.

Integrating approaches from all preceding chapters, camel-associated accessory genes were identified with GWAS within the context of the wider GBS population (genomes in dataset from chapter 4 were included in the comparison), and MGE (prophages, PICI and ICE) were detected with gold standard methods developed in previous chapters. A manuscript including the findings of this second set of analyses is currently in preparation for publication.

6. In chapter 7, the results of genomic analyses from previous chapters were integrated and discussed, also in the wider context of the genus *Streptococcus* (e.g. genetic species vs ecological species). Implications of bioinformatic findings, particularly on host-associated MGE and their distribution among host-specialist and generalist lineages, for human and animal health (including zoonotic and reverse zoonotic transmission) and future work are discussed.

Chapter 2

Development and application of methods for the identification of mobile genetic elements in group B *Streptococcus* genomes from multiple host species

2.1 Introduction

Group B *Streptococcus* (GBS) is a multi-host pathogen primarily affecting humans, dairy cattle and fishes. The GBS population comprises host-specialist and host-generalist lineages (see chapter 4), which are variably detected in these three major host groups. The existence of host-specific and host-generalist lineages has also been observed in other multi-host pathogens such as *Staphylococcus aureus* (Richardson et al., 2018), for which it has been hypothesised that successful inter-species transmission could be explained by the acquisition of MGE from an accessory gene pool that is present in the recipient host species, and/or by the loss of MGE of the source host species. As an example, the acquisition of novel MGE (bacteriophages, PICI and plasmids) from an avian-specific accessory gene pool, together with the loss of function of human-disease-specific genes, is thought to have caused

a recent human-to-poultry host jump in *S. aureus* (Lowder et al., 2009). In GBS, horizontal gene transfer (HGT) events could also be responsible for host-switching in the case of host-generalist lineages, for example for shared human-cattle sequence types (ST) (e.g. clonal complex CC1) (Lyhs et al., 2016; Manning et al., 2010; I. C. M. Oliveira et al., 2006), or contribute to the long-term host adaptation of host-specialist lineages (e.g. CC61/67 in dairy cattle).

Moreover, several types of MGE have been shown to impact on different features of GBS isolates, particularly on their pathogenicity. In GBS, human isolates carrying a high number of prophages showed greater virulence, and have been linked to a higher invasiveness compared to the ones with fewer prophages (Salloum et al., 2011, 2010; Domelier et al., 2009; van der Mee-Marquet et al., 2006). Integrative and conjugative elements (ICE) and plasmids have been reported to carry and spread antimicrobial resistance genes among GBS strains, although plasmid prevalence in GBS is usually low (Sendi et al., 2016; C. M. Johnson & Grossman, 2015; Compain et al., 2014; Klima et al., 2013; DiPersio et al., 2011; Mata et al., 2011; Michael et al., 2011; Wozniak & Waldor, 2010). Other elements, such as PICI, a family of small MGE that exploit bacteriophages for their own transmission (Martínez-Rubio et al., 2017; Penadés & Christie, 2015; Novick et al., 2010), have never been studied before in GBS. PICI play an important role in the epidemiology of *S. aureus*, the bacterium in which they were first discovered, carrying the toxic shock syndrome toxin-1 (TSST-1) (Lindsay et al., 1998), which is responsible for the clinical manifestations of the toxic shock syndrome (Todd et al., 1978).

In GBS, an extensive investigation of different types of MGE on a large genomic dataset comprising both human and animal sequences has never been carried out before. To my knowledge, no previous studies focused on the detection of prophages and plasmids in GBS of animal origin, or on the detection of PICI in isolates of any origin. Considering the impact of MGE in host-adaptation and pathogenicity of other bacterial species, such as *S. aureus*, it is important to determine their presence and distribution in GBS from different lineages and hosts.

Several bioinformatics programs are available for the detection of prophages (Amgarten

et al., 2018; A. L. de Sousa et al., 2018; Arndt et al., 2016; Cresawn et al., 2011; Lima-Mendez et al., 2008; Bose & Barber, 2006; Fouts, 2006), ICE (M. Liu et al., 2018; Langille & Brinkman, 2009) and plasmids (Galata et al., 2018; Carattoli et al., 2014a, 2014b), although plasmid searches are usually limited to databases of known plasmids which derive from studies on human isolates. The main limitations of bioinformatic programs for the detection of prophages are:

1. Low sensitivity due to a lack of GBS-specific prophages in the search databases. These databases are often built on a set of prophage sequences from the most commonly-studied bacteria, such as *Escherichia coli* (Javan et al., 2019). As prophages are considered highly species-specific, searching for prophage sequences from different bacterial species can increase the false negative rate;

2. Low sensitivity due to full-prophage sequence search on draft genome assemblies. Short-read sequencing technologies such as Illumina (Bennett, 2004) are currently the most cost-effective and widely employed for next generation sequencing. However, these generate fragmented genome assemblies and, in case of prophage assembly over multiple contigs, the search result can lead to false negatives (Jamrozy et al., 2017) (this also applies to most ICE-detection programs, such as ICEFinder (M. Liu et al., 2018));

3. Impracticality when dealing with large genomic datasets. Most of these programs follow either an on-line server queue or are local but semi-automated and require several additional steps from the user in order to inspect and download prophage sequences;

4. Lack of a reproducible and standardised classification method. This represents a limitation especially when trying to compare results from different studies, similarly to what happens for the comparison of bacterial genomes with comparative typing methods vs library typing methods (e.g. MLST) (see chapter 1 and 4).

A possible way to overcome these issues would be the adoption of a classification scheme based on bacterial species-specific prophage integrase types. This type of approach is already in place for some bacterial species, e.g. *S. aureus* (Goerke et al., 2009). These typing schemes are based on the concept that prophage integrases are site-specific (i.e. one type of

integrase is usually found at only one chromosomal insertion site) through the recognition of chromosomal attachment sites (*att*B) (Campbell, 1992), short nucleotide segments that are identical to phage attachment sites (*att*P). The *att*B corresponds to the insertion site where the phage recombines, becoming an integrated lysogenic prophage. Once the prophage is integrated, the *att* site is usually found at both ends of the prophage. Integrase-based typing schemes also exist for PICI (Fillol-Salom et al., 2018; Penadés & Christie, 2015). However, to date, no bioinformatic programs specifically designed for the detection of these MGE are available. Hence, manual inspection of whole genome sequence data is the only strategy for *in silico* identification of PICI.

The aim of this study was to investigate the presence of various types of MGE, including prophages, PICI, ICE and plasmids, in a large genomic dataset comprising both human and animal sequences. As MGE have been shown to impact on ecological success, pathogenicity and host-jumps of multiple gram-positive and gram-negative pathogens, knowing which types of MGE exist in GBS and their distribution across GBS strains is important. To this end, a GBS-specific prophage and PICI detection and typing method based on the integrase gene was developed, similar to what exists for *S. aureus* (Goerke et al., 2009). A database of known GBS plasmids was compiled and used to screen isolates, while existing online tools were applied to identify ICE in sequence data. New tools for prophage and PICI recognition (integrase typing schemes) were developed on a collection of closed genomes[1] available in the public domain (dataset 1) and further enhanced using a global genomic dataset providing coverage across a wider and more representative range of countries and host species (dataset 2); existing tools for the detection of ICE and IME, as well as blast for plasmid searches, were applied to dataset 2.

2.2 Materials and methods

All supplementary material for this chapter, including tables and figures, can be found in Appendix A (these are indicated with the letter A in front of the sequential number).

[1]Bacterial genome assemblies can either be draft genomes, nucleotide sequences fragmented in multiple sections known as contigs, or closed genomes, complete circularised sequences of the isolates' DNA.

2.2.1 Datasets included in this study

MGE analyses were initially carried out on a publicly available dataset, consisting of closed GBS genome sequences obtained from NCBI (dataset 1, Tab. A.1), and then on a custom dataset (dataset 2, Tab. A.2). Dataset 1 was used for methods development and implementation, and in particular to test the available pieces of software for MGE identification and to establish a classification scheme for prophages and PICI in GBS. These methods were then applied for the analysis of MGE in dataset 2. During the analysis of dataset 2, a further refinement of the prophage and PICI integrase typing method developed using dataset 1 was also carried out.

Dataset 1 comprised 69 closed genome sequences representing major and minor host species (humans, *n*=15; fishes, *n*=49; cattle, *n*=2; camel, *n*=1; frog, *n*=1; unknown, *n*=1) and five continents (Africa, *n*=1; South America, *n*=40; North America, *n*=4; Asia, *n*=22; Europe, *n*=1; unknown, *n*=1). Genomes originating from fish represented the majority of the dataset, and forty-one of them were published in one study from Brazil (Barony et al., 2017), hence genomic diversity was limited for this group. In fact, the main ST in this dataset were ST552 (*n*=26) and ST260 (*n*=9), which are part of CC552, a CC which is highly adapted to the aquatic environment and poikilothermic animals (Delannoy et al., 2016; Rosinski-Chupin et al., 2013). Despite this bias towards GBS fish genomes from Brazil, it was decided to include all of these sequences in dataset 1. This is because the evaluation of the prevalence of MGE based on host species or country was not the aim of this work, rather genomes were being screened for MGE inventory and typing scheme development.

Dataset 2 is a subset of 901 publicly available sequences included in the study by Richards et al., 2019, and includes all sequences with a maximum of 50 contigs (*n*=503). This filter was applied to the original dataset so that only genomes that were of high quality, albeit not necessarily closed, were included, as high genome fragmentation can lead to sub-optimal performance of certain bioinformatics programs (e.g. programs that search for long gene clusters, such as PhageMiner or ICEFinder, as detailed below). Additionally, these genomes provided more in-depth coverage of major and minor GBS host species, geographic diversity, and GBS clades. As for dataset 1, isolates in dataset 2 originated from major host species

(humans, *n*=486; fishes, *n*=8; and cattle *n*=5) and minor host species (camel, dog, dolphin and seal, *n*=1 per species). Geographical origins were diverse (Africa, *n*=1; the Americas, *n*=353; Asia, *n*=10; Europe, *n*=117; Oceania, *n*=18; unknown, *n*=4). Fourteen CC and fifty-three ST were represented in dataset 2 (Tab. A.3), with the most well-represented being common GBS clades from humans (CC1, *n*=260; CC17, *n*=90; CC23, *n*=56; CC12, *n*=38; CC19, *n*=29) or fishes (CC7, *n*=6). These data altogether show a bias of dataset 2 towards isolates deriving from humans, originating from North America and belonging to CC1.

2.2.2 Implementation of existing methods for the identification of prophages and development of GBS-specific prophage and PICI typing schemes based on the integrase gene

Detection of prophages, PICI and their integrase genes in dataset 1

Several bioinformatic tools for the identification of prophages in bacterial genomes are available on-line (Amgarten et al., 2018; A. L. de Sousa et al., 2018; Arndt et al., 2016). To obtain the most complete database possible, and to assess agreement between methods, closed genomes from dataset 1 were analysed with three methods, i.e manual screening of GenBank files, PHASTER, and PhageMiner. GenBank files were used for manual screening for potential phage sequences starting from genes annotated as "site-specific integrase", "integrase" or "recombinase". Manual inspections were also used to identify putative PICI, as there are no specific bioinformatic programs available for the detection of these MGE. A detailed description on how manual inspection of genomes for the presence of PICI was carried out can be found in section A.2, Appendix A. PHASTER (PHAge Search Tool Enhanced Release) (Arndt et al., 2016; Y. Zhou et al., 2011) is a widely used web-based integrated search and annotation tool for phage display. PhageMiner (Javan et al., 2019) is a user-supervised semi-automated computational tool that enables the identification of prophage sequences within complete or draft bacterial genomes. It allows for rapid identification, user inspection and curation of phage sequences from large numbers of genomes and has been validated on streptococci (Javan et al., 2019). For this study, PhageMiner was run locally on GenBank files annotated with one of the recommended annotation tools for this program, RAST v2.0 (Aziz

et al., 2008) or Prokka v1.11 (Seemann, 2014)[2]. A database of phage integrase types was built from the identified complete prophages. Incomplete prophage sequences, whether due to genome fragmentation or lack of essential genes such as the integrase, were not included in the analysis. Integrases were classified based on insertion site and percentage identity (% ID), using translated amino acid (AA) sequences, and numbered in order of detection. If a blastp (Camacho et al., 2009) comparison resulted in >90% ID (Fig. 2.1, Fig. A.1, Tab. A.4) and >95% query cover (QC), integrases were considered to belong to the same type. When an integrase did not meet these thresholds but occupied the same integration site as an integrase that had already been classified, a subtype number was added[3] (e.g. GBS*Int*2.1 and GBS*Int*2.2 represent integrases that both occupied integration site GBS2 but with <90% sequence similarity). The same set of prophages was identified with all three detection methods.

Detection of prophages and integrase genes in dataset 2

Because all methods identified the same prophages in dataset 1, only PhageMiner was used for dataset 2. This program can be run locally, eliminating waiting time for server queues that may affect analysis speed for server-based programs like PHASTER, which is particularly relevant for large batches of genomes such as dataset 2. In addition, PhageMiner can generate annotated maps of putative prophage sequences, allowing for almost instantaneous inspection, and it can automatically store the extracted prophage sequences. PhageMiner requires a threshold of fewer than fifty contigs for an optimal performance, as results obtained from fragmented (>50 contigs) and highly fragmented (>100 contigs) genomes could lead to false negative results in terms of MGE detection, as explained in section 2.1; this is why sequences that had over fifty contigs were filtered out during the curation process of dataset 2. For complete prophages identified in dataset 2, the integrase amino acid sequence was compared against the phage integrase database derived from dataset 1 using blastp to classify the phage integrase type, as detailed for dataset 1. PhageMiner searches often recognised phages

[2]The two annotation tools used did not lead to different results in terms of prophage detection.

[3]The existence of multiple integrase subtypes in some insertion sites is a sign that these are likely hotspots for bacteriophage chromosomal integration; this suggests a possible biological advantage for integration in these sites compared to other sites (e.g. highly conserved genes are recognised as insertion site rather than accessory/dispensable genes which are not present in all GBS isolates).

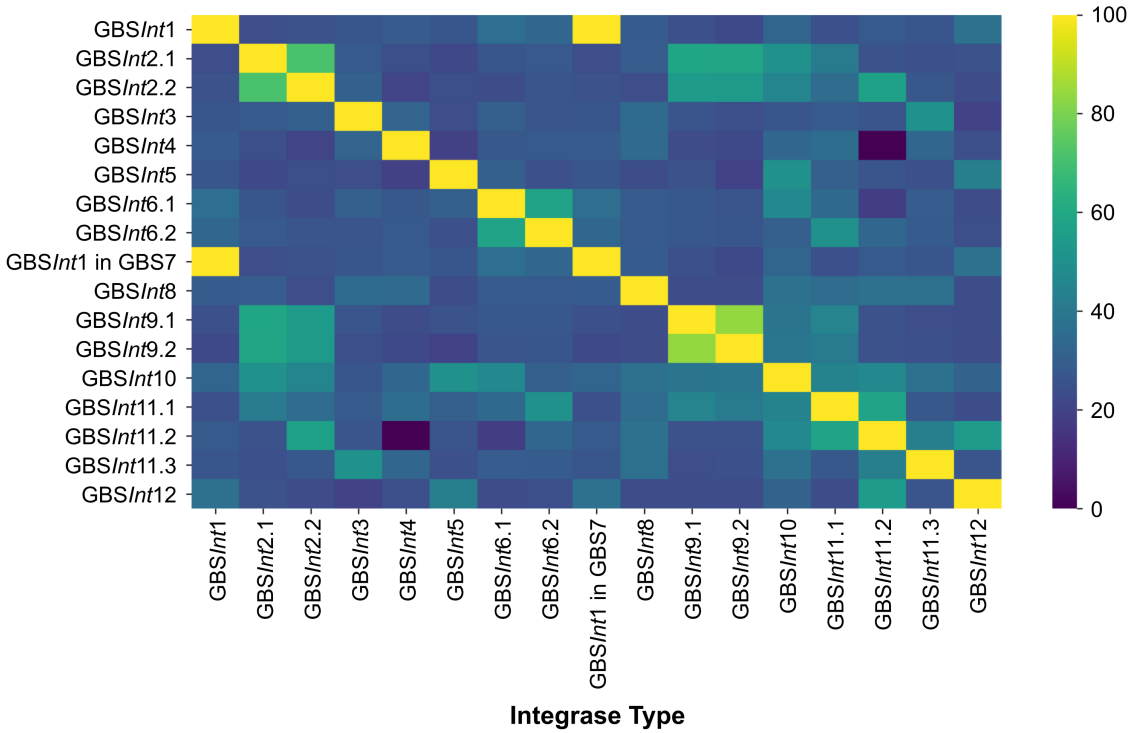


Figure 2.1: Heat-map showing the pairwise percentage of identities (% ID) at the amino acid sequence level between sixteen prophage integrase types identified in this work in group B *Streptococcus* (GBS). GBS*Int*1 was found in two different insertion sites: GBS1 and GBS7.

as partial rather than complete, even for full prophages, e.g. due to annotation of integrase genes and other prophage-related genes as hypothetical proteins. To allow for visual differentiation between partial and full prophages, the inspection window was widened, generally by 15 genes on either side of the sequence detected by PhageMiner.

Putative attachment sites were identified bioinformatically using blastn, through comparison of the site of integration in an empty genome (i.e. not harbouring the prophage, chosen among closed genomes of ideally the same ST and host species) and the regions at both ends of the integrated prophage in a genome harbouring the prophage. Closed genomes in dataset 2 (n=25) were scanned manually for PICI identification, whereas draft genomes were screened for PICI presence with tblastn, searching for the integrase amino acid sequences of already-identified PICI.

Whole-prophage and integrase gene phylogenies

Two hundred eighty-two complete lysogenic prophages were detected in dataset 2. Using PhageMiner, all complete prophage sequences were extracted from the genome assemblies (one complete prophage from 38% of genomes and two complete prophages from 9% of genomes) in dataset 2 and stored as GenBank files. Prophages that straddled two contigs were excluded from the phylogeny (n=16). Extracted prophage sequences (n=266) were manually inspected and curated with Geneious v2020.1.2 (Biomatters Ltd, `https://www.geneious.com`), as were n=22 prophage sequences identified in the study by van der Mee-Marquet et al., 2018, that were added to my phylogenetic analyses for comparison. Sequences were reverse-complemented as needed to start with the integrase gene, and all integrase protein sequences were also stored separately. Multiple sequence alignments were performed for whole-prophage sequences and for integrase genes, respectively, using ClustalW v2.1 (Thompson et al., 1994) with default settings (Gap opening penalty = 10, Gap extension penalty = 0.20). Approximate maximum-likelihood phylogenetic trees were constructed from the sequence alignments of the nucleotide sequences for prophages and of amino acid sequences for the integrase genes using FastTree v2.1.11 (Price et al., 2010) with the Jukes–Cantor model (default parameters).

Figures were edited using Inkscape (`www.inkscape.org`).

2.2.3 Application of existing methods for the detection of ICE and plasmids in GBS genomes

The software of choice for the identification of ICE in dataset 2 was ICEberg 2.0 (M. Liu et al., 2018), and in particular its tool ICEfinder, a program that predicts ICE or IME in bacterial genome sequences. Briefly, ICEfinder first detects recombination and conjugation gene cassettes using Hidden Markov Model profiles. The origin of transfer site (*ori*T) is detected using a homology search against a database of 1,074 *ori*T sequences. Elements carrying an integrase gene, a relaxase gene, and a type IV secretion system (T4SS) are considered T4SS-type ICE. Elements with an integrase and relaxase gene, but lacking a T4SS, are classified as IME. The program is available online and as a standalone version: the online tool allows

users to submit as a query either a GenBank file containing a nucleotide sequence and its annotation or a raw nucleotide sequence in fasta format, which is first annotated using the server annotation tool (CDSeasy) (J. Li et al., 2017), and is then used as the input for ICE detection. The local version, which works only on GenBank files, is available for Linux processors from the program developers. The local version was selected over the online one for various reasons: the possibility to run large batches of sequences with one command instead of manually uploading each file separately, no waiting time due to the server queue, and the automatic storage of output files in the processor. The standalone version was installed on a Linux Ubuntu v18.04.2 virtual machine.

For plasmid identification, a database of known GBS plasmids (n=5) was built. Plasmids that were investigated are: pCCH208 (GenBank accession n. KJ778678) (Compain et al., 2014), pGB2001 (accession n. JF308630) and pGB2002 (accession n. JF308629) (DiPersio et al., 2011), all harbouring *erm*(T) resistance, pPI502 (accession n. KP698941) which carries a gentamicin resistance gene (Sendi et al., 2016), and finally the pNEM316-1, which is a conjugative plasmid (Herbert et al., 2005).

A summary of the methods used to detect MGE in dataset 2 is presented in Tab. 2.1.

Table 2.1: Summary of identification methods used for for mobile genetic elements (MGE) detection in dataset 2. After development of a database of PICI integrases (amino acid sequences), tblastn was used to search for these elements in genome assemblies' nucleotide sequences and to classify their type. PhageMiner was used to identify prophages and blastp searches were used to compare the phage integrase type; if the integrase (amino acid sequence) shared >90% identity (ID) and >95% query coverage (QC) with one in the database, it was classified as the same type. ICE/IME were detected with ICEfinder local and results were expressed as total count. Blastn searches against the genomes were performed to identify the different plasmid types.

MGE type	Identification method	Result
PICI	tblastn integrase (ID >90%, QC >95%)	count/type
Prophage	PhageMiner/blastp integrase (ID >90%, QC >95%)	count/type
ICE/IME	ICEfinder local	count
Plasmid	blastn of known database (ID >95%, QC >80%)	count/type

2.3 Results

2.3.1 Prophages and PICI

Detection of insertion sites and integrases across the prophage phylogeny

Twelve prophage insertion sites were identified and progressively numbered as GBS1 to GBS12. The 12 insertion sites were occupied by 16 integrase types, implying that there were subtypes for some integration sites (Fig. 2.2, Tab. A.5). Ten integrase types were identified in dataset 1, with two additional types and four subtypes identified in dataset 2. The complete database of integrase types can be found in Appendix A, section A.3, and at the GitHub online repository: chcrestani/GBS_prophage_integrase_typing. Putative attachment sites (Tab. A.5) were identified bioinformatically for twelve integrase (sub)types, but the search was inconclusive for five (sub)types, as blast searches were unable to detect short sequences with similarity between the two sides of the prophages (*attL* and *attR*) and the integration site in genomes without these phages (*attB*). Mean prophage integrase length was 387 $\pm$ 48 AA (Tab. A.5). Blastp comparisons of integrase type % ID and QC can be found in Tab. A.4, Appendix A. Integrase types and subtypes predominantly clustered with their respective prophages in the whole-prophage phylogeny (Fig. 2.3). Major prophage groups were located at insertion sites GBS2, GBS3, GBS4, GBS9 and GBS11. Phages with GBS*Int*2.2 (*n*=60) were more common than those with GBS*Int*2.1 (*n*=2). For completeness, one representative sequence of prophage type GBS*Int*11.3, which I had identified in analyses carried out outside the scope of this chapter, was added to the phylogeny of phages, and its integrase was added to the integrase phylogeny. Minor prophage groups (GBS1, GBS5, GBS8, GBS10, GBS12) branched out from within major clusters. For some prophages, a mismatch between their integrase type and the integrase type of the surrounding prophage cluster was observed (red branches in Fig. 2.3, Fig. A.2). This included all GBS6 prophages (GBS*Int*6.1 and GBS*Int*6.2), which were distributed across multiple branches of the prophage clades associated with GBS2 and GBS3 (Fig. A.2). GBS*Int*4 was associated with its own monophyletic phage clade and integration site, GBS4, but was also found on branches of the prophage clades associated with GBS2, GBS3 and GBS11. Likewise, GBS*Int*2.1 and GBS*Int*2.2, GBS*Int*3 and GBS*Int*11.1 were associated with their own clade and integration sites (GBS2,

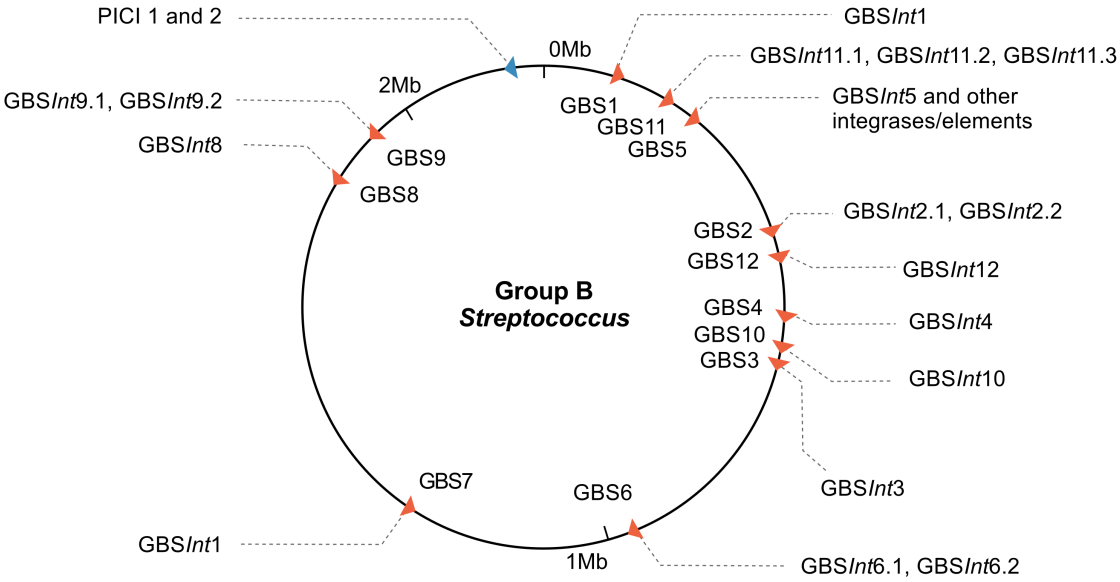


Figure 2.2: Map of the insertion sites of prophages and putative phage-inducible chromosomal islands (PICI) in group B *Streptococcus* (GBS). Twelve phage insertion sites (red arrows) and sixteen integrase types were identified. Phage insertion sites are indicated with "GBS" and a progressive number based on order of detection, while integrase type sub-number indicates the different subtypes of integrases found at the same insertion site based on less than 90% similarity in the amino acid sequence. GBS5 integration site corresponds to the *rpsI* gene, a site of integration in common with ICE (Ambroset et al., 2016; Brochet et al., 2008) and PICI-like elements (this work). The putative PICI insertion site is displayed in blue and is the same for both PICI integrases (1 and 2). Arrows show the direction of packaging.

GBS3 and GBS11, respectively), as well as being detected in other prophage clades, i.e. GBS2 and/or GBS3. The integrase phylogeny (Fig. A.3) showed defined clusters, with varying levels of diversity within clusters. Integrases located at the same insertion site generally formed monophyletic clades, with the exception of GBS*Int*11.3, which was more closely related to GBS*Int*3 than to GBS*Int*11.1 or GBS*Int*11.2.

Insertion site peculiarities and PICI identification

GBS*Int*5 was identified in GBS5 (*rpsI* gene) in genome QMA0323, where the full prophage is present, preceded and followed by other genes with signatures of an ICE (Fig. A.4). By contrast, in genome FSL_S3-026, integrase GBS*Int*5 is present as a singleton, i.e. not followed by a full prophage. Rather, it was found inside what was classified as a putative ICE

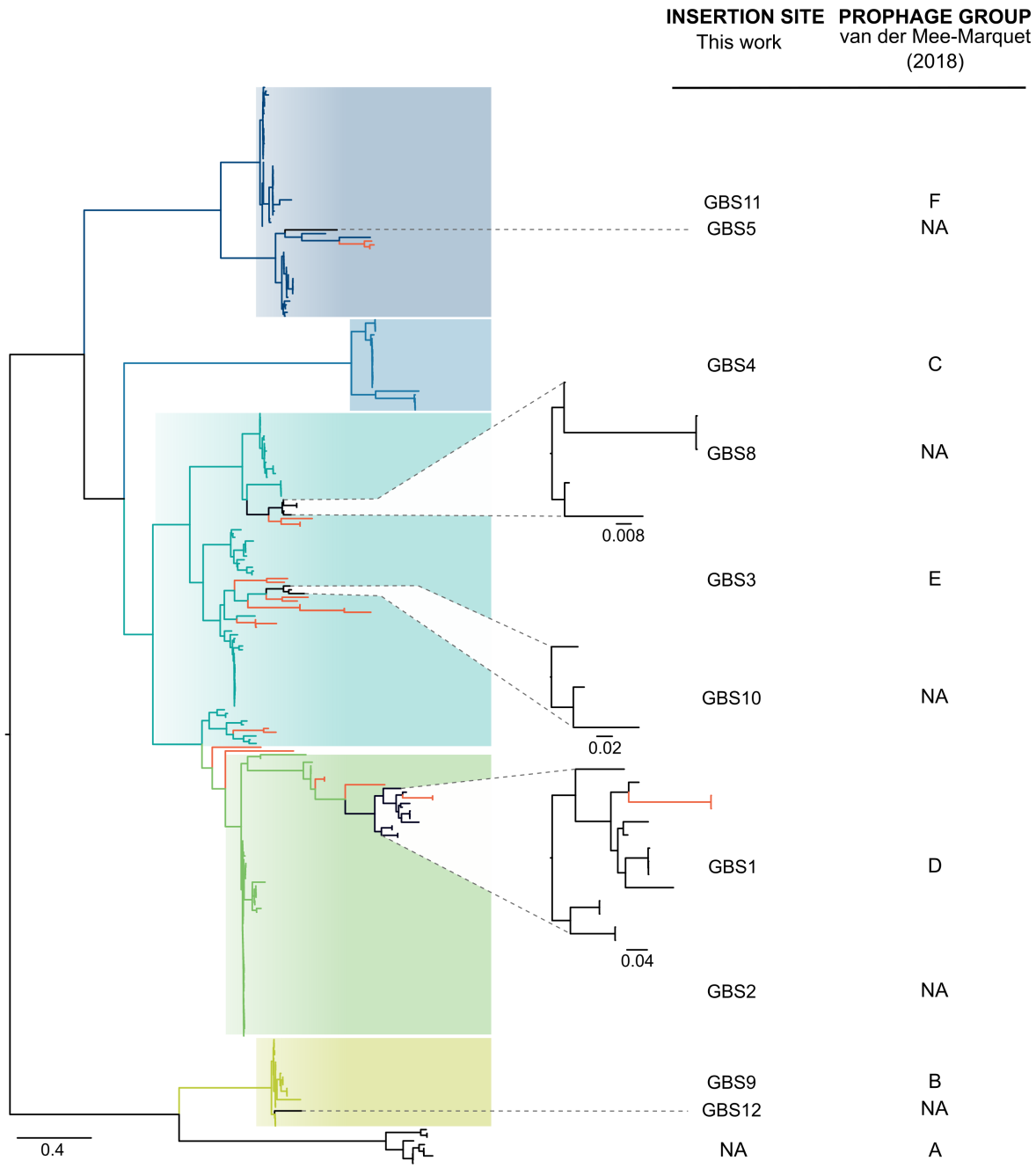


Figure 2.3: Approximate maximum-likelihood phylogenetic tree of 266 complete prophages identified in group B *Streptococcus* (GBS) in this study and 22 prophages identified by van der Mee-Marquet et al. (2018). In most cases, full-prophage phylogenetic clusters are concordant with insertion sites and their corresponding integrase types or subtypes (GBS2, GBS3, GBS4, GBS9, GBS11, blue to green branches), with smaller clusters (GBS1, GBS5, GBS8, GBS10, GBS12, black branches) embedded in the larger ones. Some mismatches between prophages and their integrase type or insertion site and major clusters were identified (red branches), which is suggestive of integrase switching events. Tree was rooted at midpoint. NA: not applicable.

(∼67,000bp) by ICEFinder (M. Liu et al., 2018). This larger ICE showed partial similarity with a region of ∼9,000bp found after the prophage in QMA0323.

The label GBS*Int*7 is not used because the site-specific integrase at insertion site GBS7 was identical to GBS*Int*1 at insertion site GBS1 (Fig. 2.1 and Tab. A.5). GBS*Int*1 at site GBS7 was only observed in this location when the GBS1 site was occupied by a prophage and it was uniquely observed in ST283 (n=6 and n=1 complete genomes from dataset 1 and 2, respectively), the only known hypervirulent GBS in human adults (Barkham et al., 2019). Interestingly, GBS*Int*1 at site GBS7 was only detected in closed genome sequences, and never in draft genomes.

At insertion site GBS11 (Tab. A.5), the full prophage immediately followed *gspF* (n=18 genomes), or it was separated from *gspF* by a few genes encoding small proteins (n=17 genomes). The latter included competence proteins, type II secretion system proteins, and hypothetical proteins (Fig. A.5). There was no clear correlation between the integrase subtype (GBS*Int*11.1, GBS*Int*11.2) and any of these GBS11 site variants, but there was correlation between prophage subcluster and integrase type (Fig. A.6). For 26 prophages with either GBS*Int*11.1 or GBS*Int*11.2, it was not possible to assess the integration site because the prophage was found at the end of a contig.

In addition to prophages, two putative PICI sequences (PICI1 and PICI2) were identified using manual screening (Fig. 2.2 and Fig. 2.4). Both integrases were 398 AA long and shared the same integration site (*rpsD* gene). Amino acid sequences for PICI1*Int* and PICI2*Int* can be found in Appendix A, subsection A.3. PICI2 was uniquely identified in dataset 2. PICI-like MGE were also detected in the integration site corresponding to the *rpsI* gene, i.e. in the same location as GBS*Int*5 (Fig. 2.5). However, it was not possible to classify these PICI-like elements with certainty, as they could have been fragments of other elements such as prophages or ICE (see discussion).

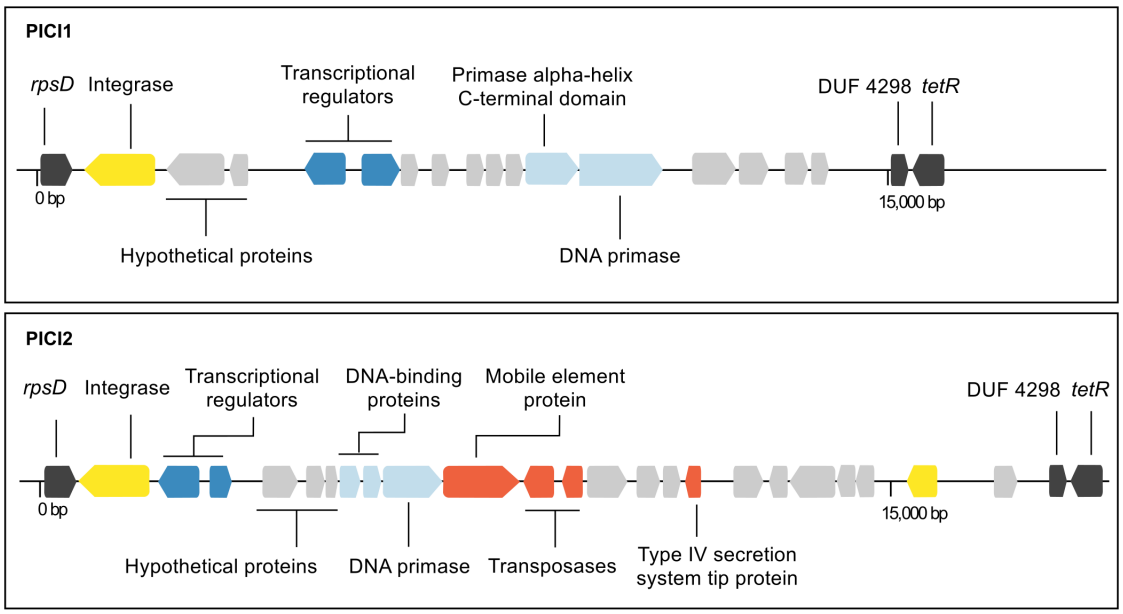


Figure 2.4: Annotated maps of genes in phage-inducible chromosomal island (PICI) 1 (genome 09mas018883) and PICI2 (genome ILRI005). The integration site is the same for both integrases (*rpsD* gene). Genes are colour-coded based on function (black: chromosomal genes; yellow: site-specific integrase; dark blue: lysogeny genes; light blue: replication genes; light grey: hypothetical; red: other genes).

2.3.2 Distribution of mobile genetic elements in a global GBS dataset

A total of 1,664 MGE (329 PICI, 696 complete and incomplete prophages, 10 plasmids, 248 ICE and 381 IME) were identified among 494 (98%) of the 503 isolates in dataset 2. Two or more MGE were seen in 434 isolates (86%) and three or more in 361 isolates (71%). The maximum number of MGE seen within a single isolate was seven; this was observed in a genome that originated from a bovine GBS isolate and belonged to ST67.

The distribution of MGE by major host species can be found in Tab. A.6. Bovine genomes (n=5) harboured a higher proportion of MGE compared to the other host groups.

Two hundred eighty-two complete prophages were detected in dataset 2. To create as complete an integrase database as possible, GBS genomes representing a wide variety of host species, countries and GBS clades were included in the analysis. However, the study was not designed to be an epidemiologically representative survey of prophage or integrase

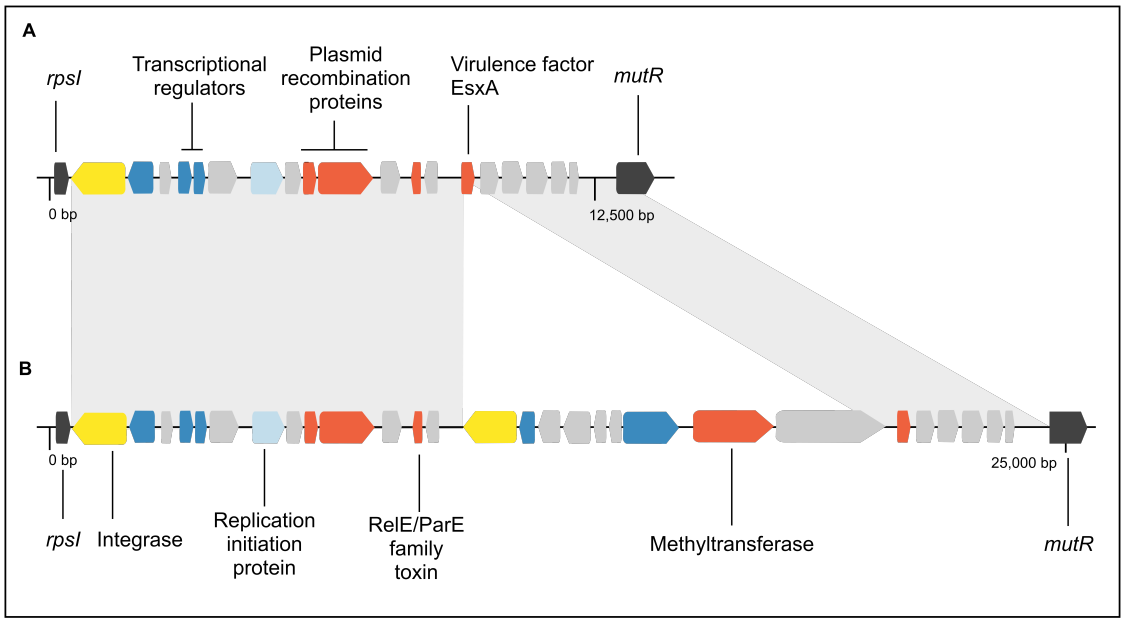


Figure 2.5: (A) Example of a putative phage-inducible chromosomal island (PICI) detected in the integration site *rpsI* (genome CU_GBS_98). This site is a hotspot for recombination of integrative conjugative elements (ICE) (Brochet al., 2008) and prophages (this work). The presence in this site of multiple site-specific integrase genes (B) is indicative of successive integration events (genome 09mas018883). Genes are colour-coded as in Fig. 2.4.

distributions, so calculation of prophage prevalences is not meaningful, but some qualitative observations about the association with genome origin can be made. Complete prophages were detected across isolates from most host species (Fig. A.7), including 47% of human GBS genomes (n=230 out of 486 isolates, with a total of 274 complete prophages) and three fish GBS genomes, but not in bovine and canine GBS genomes (n=5 and 1, respectively).

Prophages and their integrases were detected in most GBS clades, with the exception of certain clades represented by three or fewer isolates (CC22, CC67 and CC130, Tab. A.7), and the majority of integrases were detected in multiple clades (Fig. 2.6). The number of integrase types per CC ranged from 1 to 10 (Fig. 2.6). All major ST in dataset 2 (ST1, ST17, ST19, ST23, ST459) harboured at least four prophage types (Tab. A.8). Complete prophages were identified in GBS isolates from all continents except for South America (n=2 genomes). The number of discovered prophages tended to reflect the total number of genomes per continent, whereby more prophages were detected in continents with more genome sequences (Tab. A.9).

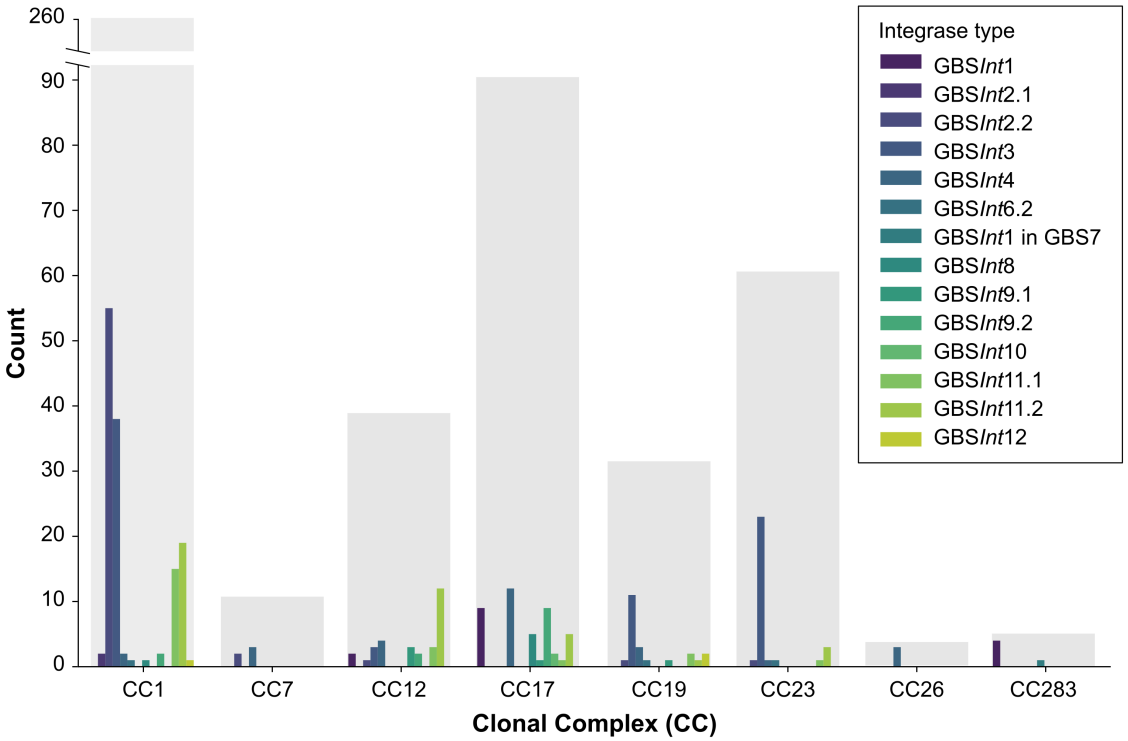


Figure 2.6: Distribution of complete prophages classified based on their integrase types (GBS*Int*1 to GBS*Int*12) in a publicly available dataset of 503 group B *Streptococcus* (GBS) genome sequences (dataset 2) comprising a global collection of isolates from seven hosts species. Results for major clonal complexes (CC) are shown. Grey bars show the total number of genomes per CC.

PICI were also detected across GBS from most host species, with PICI1 found in a total of 328 GBS genomes from humans, fish, cattle, a dog and a dolphin, and PICI2 found in a camel GBS genome from Kenya (ILRI005).

ICE and IME were well represented among the main CC (Fig. 2.7): one to three ICE were present among 141 CC1 isolates, 19 CC19, 18 CC17, 15 CC12, 10 CC23, 9 CC7 and 4 CC283. One to three IME were present among 185 CC1, 41 CC23, 33 CC12, 22 CC19, 10 CC7, 8 CC17 and 5 CC283.

Plasmid prevalence was low, with a total of 10 plasmids found in 9 genomes (n=8 of plasmid pGB2001, with two plasmids in one genome, pNEM316-1).

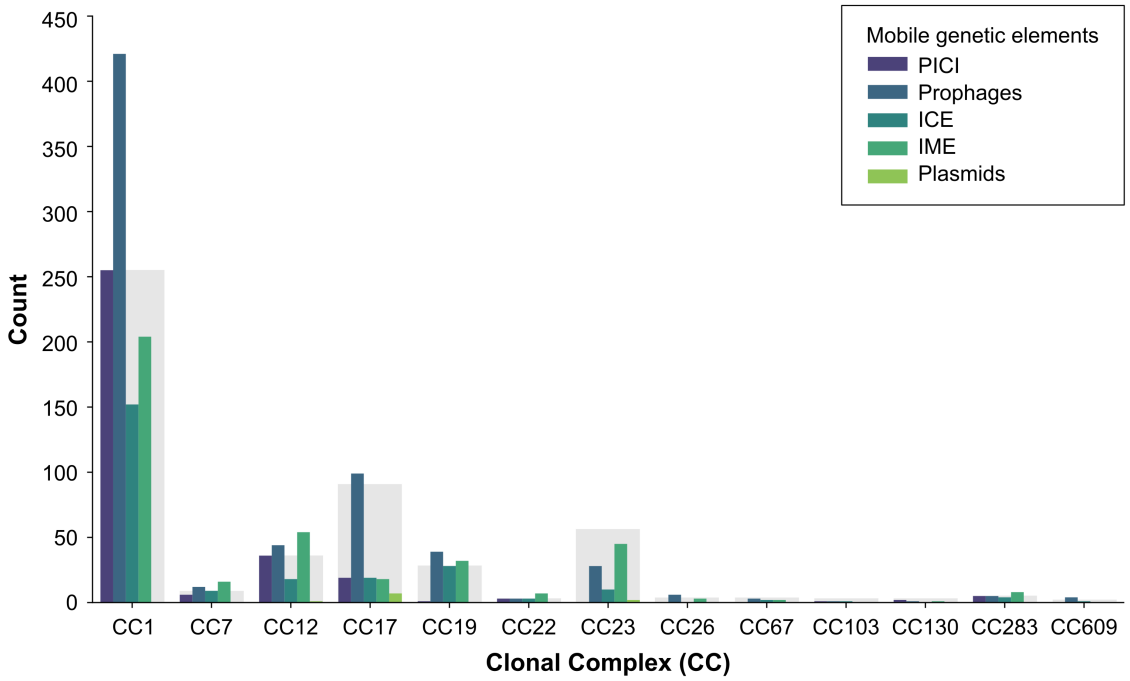


Figure 2.7: Distribution of mobile genetic elements (MGE) in a publicly available dataset of 503 group B *Streptococcus* (GBS) genomes from different host species among clonal complexes (CC). Grey bars show the total number of isolates per CC.

2.4 Discussion and conclusions

In this chapter, I explored the presence of various types of MGE within a wide dataset of complete and draft GBS genomes of both human and animal origin.

I describe the development of a typing scheme for GBS prophages based on site-specific integrase genes and insertion sites, similar to the scheme used for prophage typing of *S. aureus* (Jamrozy et al., 2017; Valentin-Domelier et al., 2011; Goerke et al., 2009). I have shown that this approach enables the rapid screening of large datasets of complete and draft genomes for the presence of GBS prophages, overcoming some of the limitations associated with existing phage detection programs, and enabling detection of phage content in fragmented genome assemblies. Additionally, blast-based searches of integrases can be performed by those with little computational experience, as blast is available as an online platform.

Phage integrase typing agreed with full-length prophage genome-based phylogenetic clusters, with a few exceptions. This is reminiscent of the relationship between the GBS

whole-genome phylogeny and capsular serotypes, where serotypes tend to match phylogenetic clusters but capsular switching may occur (Neemuchwala et al., 2016; Martins et al., 2010; Bellais et al., 2012). I propose that integrase switching may also occur, leading to mismatches between prophage genome phylogeny and integrase phylogeny, and conferring to the prophage the ability to integrate in a different location in the GBS genome. This genetic plasticity may impact on the function of the prophage, and on packaging of GBS genome content. There is growing evidence that prophages contribute to emergence, niche adaptation and spread of virulent GBS, especially in CC1 and CC17 (Renard et al., 2019; van der Mee-Marquet et al., 2018; Jamrozy et al., 2018). This may include transfer of prophage content between GBS from different host species, in agreement with the detection of prophage types and integrase types across GBS from different host species in my dataset. In addition, I discovered a potential contribution of prophages to the emergence of hypervirulent ST283, which has recently been recognised as a major cause of adult invasive disease in Southeast Asia (Barkham et al., 2019; Kalimuddin et al., 2017; Rajendram et al., 2016). Contradicting the dogma that phage integrase genes are site-specific (Frost et al., 2005), the integrase at insertion site GBS7 (5' end of *hylB*), was identical to the integrase at GBS1. Prophages in GBS7-*hylB* were only present when GBS1 was also occupied by a prophage, and they were unique to ST283. The virulence gene *hylB* codes for hyaluronate lyase, an enzyme that degrades extracellular matrix components and is believed to contribute significantly to invasion (Maisey et al., 2008; Herbert et al., 2004b). I hypothesise that this prophage could play a role in regulation of the transcription and expression of *hylB* and could contribute to the hypervirulence of ST283. However, GBS7 was only detected in closed genome sequences and never in draft assemblies, suggesting that these latter could actually hide false negatives due to the way assembly algorithms work on Illumina data (short reads). Its actual prevalence among ST283 remains unclear and should be further investigated with the use of long-read sequencing technologies. Further laboratory work on its functional role should also be undertaken.

My analysis of 572 GBS genomes extends previous work by van der Mee-Marquet and colleagues – who used whole genome sequences of 14 GBS isolates to identify prophages – by increasing the number of known prophages, insertion sites and integrase types. My

major full prophage clades matched their previously defined prophage groups (Prophage group B = GBS9, C = GBS4, D = GBS1, E = GBS3, F = GBS11) (van der Mee-Marquet et al., 2018), whilst other clusters, namely those located at GBS2, GBS5, GBS8, GBS10 and GBS12, are described here for the first time. As the GBS genome database expands, the typing scheme will need to be updated with emerging integrase types and subtypes. This is also illustrated by results for insertion site GBS11 (3' end of *gspF*), which is located within an operon involved with host competence (*com* operon). GBS11 had previously been classified as two separate insertion sites, F1 and F2, based on variations observed among three prophage genomes at this site (van der Mee-Marquet et al., 2018). Based on my analysis, the bifurcation of this group of prophages, which was also observed in my whole-prophage phylogeny, correlated with different integrase types (GBS*Int*11.1 and GBS*Int*11.2), rather than with different insertion sites. In many cases the insertion site for those integrases could not be confirmed because they were located at the edge of a contig. This suggests that sequence assembly tools struggle to assemble this region of the GBS genome, an issue that could be overcome by long read sequencing.

My typing scheme does not include type A prophages (van der Mee-Marquet et al., 2018) because they are defective rather than whole prophages, lacking the integrase gene. Although this could be considered a false negative result in my typing scheme, lack of an integrase gene renders type A prophages incapable of HGT so that they can only be spread through vertical transmission, limiting their contribution to the evolution of virulence or niche adaptation. Based on integrase typing, false positive results may also occur, as demonstrated for GBS*Int*5, which was found once as part of a full prophage (isolate QMA0323, piscine ST261; (Kawasaki et al., 2018)) and once as a singleton within a larger ICE (isolate FSL S3-026, bovine ST67; (Richards et al., 2011)). A blast search of genomes with more than 50 contigs showed the presence of GBS*Int*5 as a singleton within an ICE rather than as part of a full prophage in nine bovine GBS genomes from bovine-associated lineage CC67 (Richards et al., 2019). This phenomenon was only observed for GBS*Int*5, possibly because its insertion site, *rpsI* (30S ribosomal protein S9), is a hotspot for recombination of ICE in streptococcal species (Ambroset et al., 2016; Brochet et al., 2008). When this integrase is identified within a dataset, further analyses need to be performed to determine whether a full

prophage is present. The GBS5 insertion site also contained PICI-like elements. Because of these multiple integration events, PICI-like elements in this position could not be classified with certainty, as they could have been fragments of prophages or ICE.

PICI1 and PICI2, which are reported here for the first time, were integrated into *rpsD*, which encodes 30S ribosomal protein S4. This gene had previously been described as the site of integration of an *S. agalactiae* chromosomal island (SagCI) in 3 of 9 complete GBS genomes (S. V. Nguyen & McShan, 2014), but not as a PICI. GBS differs from other members of the group of pyogenic streptococci, including *S. pyogenes*, *S. canis*, *S. dysgalactiae* subsp. *equisimilis*, and *S. parauberis*, in that their chromosomal islands are primarily integrated in *mutL* rather than *rpsD*, which may affect their functional impact. The structure of both PICI1 and PICI2 includes typical PICI features such as transcriptional divergence, a size of around 15,000bp, and presence of a DNA primase (Martínez-Rubio et al., 2017), whereas the variable portion of their organisation and content resembles the structure of SpnCIST556 in *S. pneumoniae* and SpyCI6180 in *S. pyogenes*, respectively (Penadés & Christie, 2015). Streptococcal PICI may have roles in gene regulation (S. V. Nguyen & McShan, 2014) or gene transfer (Martínez-Rubio et al., 2017). The high prevalence of PICI1 across GBS genomes from different host species, geographic origins and clades suggests that its function warrants further investigation. By contrast, PICI2 was exclusively found in one isolate from CC609, which is a camel-specific clade from East Africa (Fischer et al., 2013) (see chapter 6). Further research and experimentation would be needed to understand if and how PICI play a role in GBS evolution and virulence.

ICEfinder was able to identify a high prevalence of ICE and IME in GBS, which was consistent with what has been found in the literature (Brochet et al., 2008). Of note, inconsistencies were found comparing the outputs of the GenBank vs fasta file searches: for a subset of genomes, the results were double checked with the local version of ICEFinder on GenBank files and with the online version on both GenBank and fasta files. Although the GenBank searches are suggested to be used by the program developers for more accuracy, none of the two clearly displayed a higher sensitivity and/or specificity; rather, in a lot of cases, they simply identified a different number of elements. The fasta search could possibly be impacted by suboptimal performance of the server annotation tool (CDSeasy) (J. Li et al.,

2017), although this is declared neither in the paper, nor in the online platform. Based on the recommendations of the program developers, and because of practicality when dealing with large datasets, it was decided to carry out ICE analysis with the local version of ICEfinder, which runs on GenBank files. It is therefore possible that a certain number of ICE/IME were missed, and these results represent only an estimate of their occurrence.

Plasmid prevalence in both datasets was very low (2% in both datasets) and it was limited to human isolates (CC17, CC12 and CC23). This could be a bias related to the fact that the plasmid search was based on a database of known plasmids, which derive from studies in human GBS isolates, where these elements have mainly been explored so far (see chapter 3). Nevertheless, as the majority of sequences in both datasets derived from humans (if excluding the largely clonal isolate collections from fish from Brazil in dataset 1), these elements do currently not appear to represent important drivers for GBS evolution and adaptation.

Overall, the total number of MGE within each CC was correlated with the number of genomes in each group. However, certain types of elements were associated with specific CC. As an example, ICE and IME were highly prevalent among CC1, CC7, CC12, CC19 and CC23, but uncommon in CC17. By contrast, CC17 together with CC1, CC12 and CC19 showed a high proportion of prophages compared to isolates from CC23. PICI1 was associated with CC1 (100%) and CC12, but much less with CC17. Plasmids were detected in CC17, CC23 and CC12 but not in CC1 or CC19. As mentioned above, preferential sequencing of clinical isolates as opposed to carriage isolates could have influenced the results of these analyses, and this is true particularly for prophages, which have been shown to be more prevalent in invasive isolates.

In summary, I assessed the occurrence of different MGE families in a large genomic dataset that comprised GBS isolates of human and animal origin. I propose a new typing scheme for rapid prophage identification in large datasets of GBS genomes based on site-specific integrase gene types and their relative insertion site. This method provides a practical way of identifying phage presence with a blast-based approach in full and draft genomes, overcoming detection issues related to genome fragmentation. It is also highly reproducible and it can be used by researchers with any level of computational experience. In addition,

I show for the first time that PICI are highly prevalent in GBS, and that the PICI family diversity in GBS is quite limited compared to other bacterial species. PICI1 appears to be ubiquitous among different CC and further investigation of its role in evolution is justified considering the importance of PICI in other gram-positive cocci. A high ICE/IME and low plasmid prevalence among GBS isolates originating from different hosts and geographical areas across the globe was also detected.

Chapter 3

Genomic explanations for the temporal shift in bovine group B *Streptococcus* subpopulations in Sweden

3.1 Introduction

Group B *Streptococcus* (GBS), or *Streptococcus agalactiae*, is the leading cause of human neonatal meningitis in high income countries (Seale et al., 2017) and causes invasive and non-invasive disease in adults with or without underlying medical conditions (Lyhs et al., 2016; Barkham et al., 2019). GBS is also a commensal of the lower gastrointestinal and urogenital tract of men and women, with an estimated carriage prevalence of 20 to 30% (Kwatra et al., 2016; van der Mee-Marquet et al., 2008). Additional colonisation sites include the skin and oropharynx (Cobo-Ángel et al., 2019; van der Mee-Marquet et al., 2008; Davies et al., 2005). Many animal species can be infected with GBS, and major economic impacts are recognised in the global dairy and aquaculture industries. Emergence of GBS in animal production systems occurred concurrently with changes in husbandry practices, such as use of milking machines, or the intensification of commercial aquaculture (Barkham et al., 2019; Richards et al., 2019; Mweu et al., 2012).

Within dairy herds, GBS generally spreads through contagious transmission, with infected animals acting as the main source of the pathogen and spread of bacteria during milking, e.g. via milking machines, udder cloths or milkers' hands (Zadoks et al., 2011). In the 1950s and 1960s, mastitis control programs were implemented to limit the impact of GBS on milk production (Nielsen & Emanuelson, 2013). Such programs focused on identification and antimicrobial treatment of infected cattle and prevention of GBS transmission during milking, and led to near-elimination of bovine GBS in Canada (Riekerink et al., 2010), the UK (Zadoks & Fitzpatrick, 2009) and northern Europe (Sampimon et al., 2009; Piepers et al., 2007; Pitkälä et al., 2004), with elimination ('reduction to zero of the incidence of disease or infection in a defined geographical area' (Heymann, 2006)) achieved by most farms in those areas. The success of GBS mastitis control programs, which predate genetic typing of bacterial isolates by several decades, was attributed to the perception that GBS is an 'obligate intramammary pathogen of dairy cattle' (Mweu et al., 2012), despite its prevalence in humans. In the UK (Bisharat et al., 2004), the USA (Richards et al., 2019), and Portugal (Almeida et al., 2016), a single bovine-adapted lineage of GBS, clonal complex (CC) 61/67, predominates in cattle. This observation, combined with the absence of CC61/67 among human GBS collections, has fuelled the perception that this GBS lineage is bovine-specific (Richards et al., 2019; Almeida et al., 2016; Bisharat et al., 2004). This is in contrast with host-generalist lineages that have been reported in both humans and cattle, such as CC1 (Sørensen et al., 2019; Richards et al., 2019).

In recent years, re-emergence of GBS in dairy herds has been documented in several Nordic countries (Lyhs et al., 2016; Jørgensen et al., 2016; Mweu et al., 2012). Pathogen (re-)emergence may be attributable to changes in the host, the environment, the pathogen or an interaction of those factors. Although host genetics and production levels have changed dramatically in the past 50 years, with production levels being linked to clinical mastitis (Heringstad et al., 2000), there is no evidence of a particular association between host selection and GBS disease. Environmental changes that may contribute to GBS emergence include changes in herd size, ownership structure and management, milking machines and housing systems (Katholm et al., 2012). For example, the transition from small tie-stall barns to large free-stall barns in Norway may have contributed to the oro-faecal transmission cycle

that was recently proposed for bovine GBS, with GBS being isolated from the animal intestinal tract and from the farm environment and water sources (Jørgensen et al., 2016). Presence in sources other than the infected mammary gland could explain why GBS detection has also been observed in dairy herds that did not acquire new cows (Mweu et al., 2014, 2012). Such sources include bovine faeces (Cobo-Ángel et al., 2018; Manning et al., 2010) as well as people, with growing evidence for human-bovine interspecies transmission (Cobo-Ángel et al., 2019; Sørensen et al., 2019; Lyhs et al., 2016). It is not clear, however, why approaches that were adequate for control of GBS in other decades or countries would fail in northern Europe, unless pathogen evolution has changed the paradigm on which these programs were built, necessitating the use of additional or alternative approaches. Pathogen evolution may be driven by small genetic changes, gene loss or gene acquisition. Pseudogenisation of the capsular operon is thought to have contributed to host restriction of the bovine-specific lineage CC61/67 (Almeida et al., 2016), while acquisition of a mobile genetic element (MGE) carrying genes for lactose fermentation (Lac.2) is believed to confer a fitness advantage and adaptation to the bovine mammary gland (Richards et al., 2013, 2011). MGE carrying advantageous genes had a significant impact in shaping the GBS human population as well: the introduction and extensive usage in medial clinical practice of tetracycline in the 1940s is thought to have selected a few human-adapted clones that had acquired integrative conjugative elements (Tn*916* and Tn*5801*) carrying a tetracycline resistance (TcR) gene, *tet*(M).

It was hypothesised that the bovine-specific GBS lineage CC61/67 was eliminated from the Swedish dairy cattle population through dedicated mastitis control programs , with subsequent emergence of GBS from other lineages, possibly as a result of host-species jumping, as described for *Staphylococcus aureus* mastitis in cattle (Weinert et al., 2012) and as suggested for GBS in fishes (Barkham et al., 2019). To test this, I investigated GBS isolates collected from bovine milk in Sweden over a period of six decades, focusing on shifts in population composition and on the detection of genetic markers of host adaptation that might provide insight into a potential reverse zoonotic origin of newly emerged GBS lineages in cattle.

3.2 Materials and methods

All supplementary material for this chapter, including tables and figures, can be found in Appendix B (these are indicated with the letter B in front of the sequential number).

3.2.1 Isolate selection

Historical (1953-1978; *n*=45) and contemporary (1997-2012; *n*=77) bovine GBS isolates were obtained from the National Veterinary Institute (SVA; Tab. B.1). No isolates were available from 1979 through 1996 (inclusive). Isolates originated from bovine milk samples from 107 farms and had been submitted to SVA for diagnostic testing. In Europe, GBS isolates from a dairy farm generally belong to a single strain or sequence type (ST) (Jørgensen et al., 2016; Zadoks et al., 2011). Therefore, one isolate per farm per year was selected for sequencing, with one exception (Farm 107, Table B.1).

3.2.2 Short read sequencing

GBS culture and DNA extraction were performed by personnel at the Moredun Research Institute, whilst sequencing was carried out at the Wellcome Sanger Institute under the supervision of Prof Ruth Zadoks and Prof Mark Holmes, respectively. Archived isolates were plated on sheep blood agar (E&O Laboratories, Bonnybridge, UK) and grown overnight at 37°C to confirm viability and purity. One colony of each isolate was inoculated into Todd-Hewitt broth (Oxoid - Thermo Fisher Scientific, Waltham, Massachusetts, US) and incubated aerobically at 37°C overnight. DNA was extracted with the GenElute Bacterial Genomic DNA Kit (Sigma-Aldrich, St. Louis, Missouri, US) as per the manufacturer's instructions. Library preparation was carried out with the Nextera XT DNA Sample Preparation Kit and Miseq Reagent Kit V2 Library Preparation Kit (Illumina Inc., San Diego, California, US) and DNA was sequenced with Illumina MiSeq technology. My work started with analysis of the sequence data as described below.

Paired-end raw reads were trimmed for low-quality bases and filtered for PCR duplicates with ConDeTri v2.3 (Smeds & Künstner, 2011) and *de novo* assembly was performed with SPAdes v3.11.1 (Bankevich et al., 2012) (the complete assembly pipeline can be found

in Appendix B, subsection B.2.1). Quality control of the assemblies generated from Illumina data (*n*=122) was carried out with QUAST v5.0.2 (Gurevich et al., 2013). Results for the total length of the genome, total number of contigs, N50 and GC content were plotted with the Python Seaborn library (Waskom, 2021) (Fig. B.1) and low-quality genomes were filtered with a custom bash pipeline (the script can be found in Appendix B, subsection B.2.2). Dataset mean values for genome length, total number of contigs and N50 were 2,126,345 bp, 58 and 492,052 bp, respectively. Two genome assemblies were excluded from subsequent analyses: the first had a high GC content compared to the dataset average (isolate GC = 36.92%, dataset mean plus twice standard deviation = 35.43% ± 0.32). The sequence was checked with KmerFinder v3.1 (Larsen et al., 2014) and was identified as belonging to a different bacterial species, *Enterococcus thailandicus*. The second genome had low quality scores for total number of contigs (*n*=1,837), N50 (1,992 bp) and genome length (2,751,323 bp), which are indicative of possible contamination. Therefore, only 120 high-quality genome assemblies were selected for subsequent analyses. After this filter was applied, quality control results were plotted a second time (Fig. B.2). A bi-modal distribution of the total number of contigs can be observed (Fig. B.2B): the second curve represents genomes that are more fragmented compared to the rest of the dataset and the majority of these sequences belong to CC61/67 (mean contig number = 112, compared to mean contig number = 35 for other genomes). Genome fragmentation was attributed to presence of a relatively high number of mobile genetic elements (MGE) and insertion sequences (IS) in this lineage (Richards et al., 2019, 2011).

3.2.3 Long read sequencing

To obtain closed circular genomes, Oxford Nanopore MinION sequencing (Jain et al., 2016; Mikheyev & Tin, 2014) was applied to a subset of isolates (*n*=22, indicated in Tab. B.1). Within each lineage, isolates were selected to maximise the diversity in terms of ST, antimicrobial resistance determinants and presence/absence of integrative conjugative elements (ICE) based on analysis of the short read sequencing data. Two libraries, each consisting of 11 samples and a negative control, were prepared with the Rapid Barcoding Kit (SQK-RBK004 - Oxford Nanopore Technologies) and sequenced for 2 to 5 hours, generating an average of 1.73 Gb per run, with an estimated mean sequence coverage of 83.9x. Base

calling and demultiplexing were carried out with guppy v3.3.0 (Wick et al., 2019), Filtlong v0.2.0 (`https://github.com/rrwick/Filtlong`) was used to filter out the lowest quality reads until only 500 Mbp remained, and Unicycler v0.4.8 (Wick et al., 2017) was used to generate high-quality hybrid assemblies of raw Nanopore and Illumina data (with default settings). Negative controls always generated empty read files.

3.2.4 Core genome analysis

A core genome alignment was obtained with Parsnp v1.2 (Treangen et al., 2014). RAxML-NG v0.9.0 (Kozlov et al., 2019) was used to infer a maximum-likelihood tree under a general time-reversible (GTR)+G model, which was inspected and annotated using iTOL (Letunic & Bork, 2006). Nucleotide sequences were annotated with Prokka v1.13.7 (Seemann, 2014). To investigate unresolved relationships between isolates that could be caused by recombination in the core genes, SplitsTree v4.15.1 (Huson, 1998) was used (Fig. 3.1).

Multilocus sequence typing (MLST) profiles were identified with SRST2 v0.2.0 (Inouye et al., 2014) and capsular serotyping was conducted *in silico* following the method described by the Centers for Disease Control and Prevention (CDC) (Metcalf et al., 2017). Briefly, blastn was used to search genome assemblies for the presence of serotype-specific short sequences extracted from the capsular serotype operon of selected reference genomes. With this approach, a perfect identity match is required for serotype VII and IX, whereas a minimum identity (ID) of 96% is suggested for serotypes Ia, Ib, and II through VI. I first validated this method on a database of publicly available GBS genomes (Da Cunha et al., 2014), comprising human and animal sequences. Whole genome sequence (WGS) serotyping results matched perfectly with phenotypic serotyping results from Da Cunha et al., 2014. Although most genomes had only one best match, two best matches were observed in a few cases. For these, the sequences were re-analysed using an *in silico* serotyping method that is based on the alignment of longer serotype-specific capsular operon sequences (A. E. Sheppard et al., 2016). A lower ID threshold was observed for most serotype Ia isolates in my study compared to the CDC study (Metcalf et al., 2017), as the majority of serotype Ia nucleotide sequences had a 94% ID match; this could be due to an inter-species difference between serotype Ia in humans (CDC study) and bovines. These isolates were also confirmed as

serotype Ia with the second method (A. E. Sheppard et al., 2016).

3.2.5 Analysis of accessory genome content

Antimicrobial resistance genes were detected with ResFinder v3.2 (Zankari et al., 2012). Presence of the lactose operon (Lac.2) (Sørensen et al., 2019; Richards et al., 2011; Zeng et al., 2010), which is a marker of bovine host adaptation, was assessed with blastn (query coverage QC>90%, and ID>95%), searching for alleles Lac.2a, Lac.2b and Lac.2c (Sørensen et al., 2019). Detection of ICE Tn*916* and Tn*5801*, which carry the TcR gene typical of human-associated GBS lineages, *tet*(M) (Da Cunha et al., 2014), was also conducted with blastn searches (QC>80% and ID>95%), using reference sequence *S. agalactiae* 2603V/R, ICESag2603VR-1 (length = 18,031 bp) and *S. agalactiae* COH1, AAJR01000021.1 (selected region from 14,055 to 34,289; length = 20,235 bp), respectively.

Lac.2 variants and ICE sequences were extracted from the genomes for phylogenetic analysis with ARIBA v2.14.4 (Hunt et al., 2017), with custom-built databases. For the ICE, two genomes that were *tet*(M)-positive did not lead to an extracted sequence from ARIBA; therefore, an area of 20,000 bp surrounding *tet*(M) was manually selected and blastn was used to determine the ICE family with ICEfinder (M. Liu et al., 2018). The ICE was identified as a Tn*5801*-like element, which diverged from the Tn*5801* reference in the presence of two additional genes (Fig. B.3). Manual extraction of amino acid sequences was carried out from annotation files for the Lac.2 integrases genes, when possible, and for the *tet*(M) gene. Alignments of the nucleotide sequences of the ICE and the Lac.2 variants, and of the amino acid sequences of the *tet*(M) and the Lac.2 integrase genes, were carried out with MAFFT v7.407 (Katoh & Standley, 2013) and Neighbor-Joining trees were built within Geneious software (Kearse et al., 2012) with a Jukes Cantor model (default settings) (Fig. 3.2 and Fig. B.3).

Figures were edited using Inkscape (`www.inkscape.org`).

3.3 Results

3.3.1 Analyses of clonal complexes and phylogenetic clusters show partial lineage replacement between historical and contemporary isolates

Six major lineages were identified from the phylogenetic network visualised in SplitsTree (Fig. 3.1A). Lineages were described using CC nomenclature: CC1, CC12[1], CC23, CC61/67, CC103/314, CC297. Bovine-specific lineage CC61/67 was exclusively detected among historical isolates collected before 1970, as was minor lineage CC297 (Fig. 3.1B); by contrast, two other lineages (CC1 and CC103/314) were only detected among contemporary isolates. CC23 and CC12 strains were found among historical as well as contemporary isolates.

[1]In the interest of simplification, in this chapter CC7 (here comprising ST6, ST724, ST1512 and ST1513) and CC12 (here comprising ST8, ST10, ST12) are grouped together and referred to only as CC12, but in other chapters they are separated.

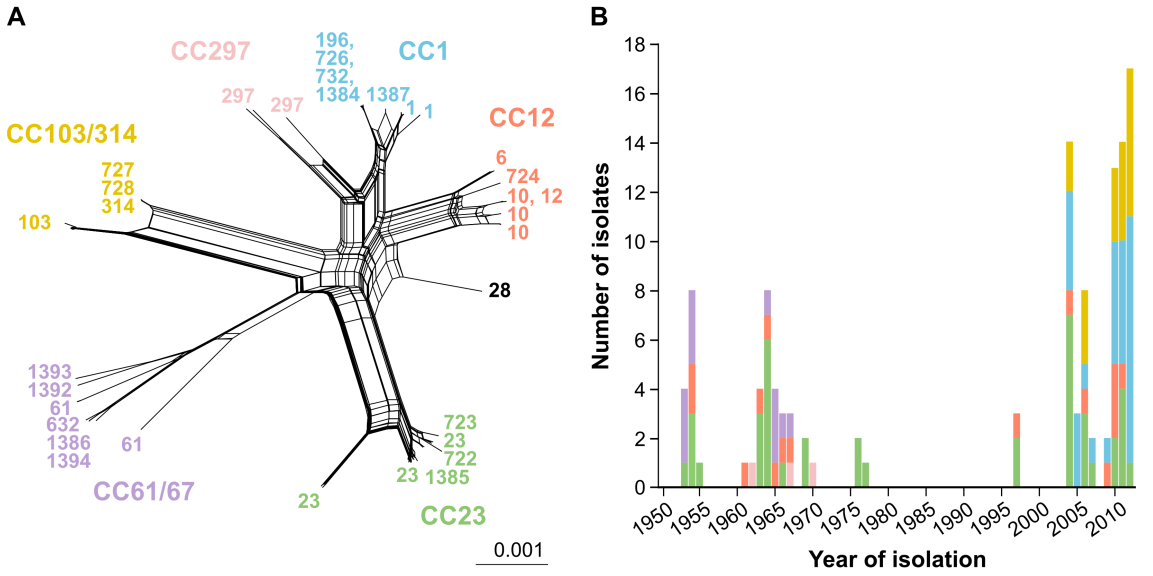


Figure 3.1: Population diversity of group B *Streptococcus* (GBS) in Swedish dairy cattle over six decades. A) Network phylogeny of 120 GBS isolates. Evidence of recombination is represented by the parallelograms which display relationships between six major clusters (CC1, CC12, CC23, CC61/67, CC103/314 and CC297). B) Bar chart displaying the year of isolation and relative abundance of the different clades over time (colours indicate clades in panel A).

Within CC23, the dominant variant during both periods was ST23, with three new variants appearing after 2006: ST722, ST723, and ST1385. Within CC12, ST6 and ST12 were detected until 1967, whereas ST8, ST10 and ST724 were detected from 1997. Overall, the most common ST were ST23 (*n*=34, 28.3%), ST1 (*n*=15, 12.5%) and ST103 (*n*=12, 10%), followed by ST10 (*n*=6), ST196 (*n*=5), ST6 (*n*=4) and ST314 (*n*=4). Eighteen new types were identified and submitted for MLST assignment at pubMLST (`https://pubmlst.org`): ST1384 to ST1387, ST1392 to ST1394 and ST1507 to ST1517. For the complete list of ST please refer to Tab. B.1.

Major serotypes identified *in silico* were: III (*n*=39, 32.5%), Ia (*n*=28, 23.3%), V (*n*=17, 14.2%), IV (*n*=13, 10.8%), Ib (*n*=10, 8.3%) and II (*n*=10, 8.3%). For three genomes, the serotype could not be determined. Serotype IV, V and all the nontypeable results were only detected among contemporary isolates, whereas all the other serotypes were present in both groups (Tab. B.1).

3.3.2 Lac.2 is highly prevalent among bovine GBS and has multiple integration sites indicative of its mobility

The vast majority of genomes (*n*=118/120) encoded for at least one of the Lac.2 variants. The first and second form ever described (Lac.2a, Lac.2c) (Richards et al., 2011) comprise nine and ten genes, respectively. Lac.2a (*lacRABCDFEGX*) begins with the *lacR* oriented in the opposite direction to the remainder of the operon, whereas in Lac.2c (*lacRABCDTFEGX*) all the genes follow the same orientation. Lac.2b (Sørensen et al., 2019) also has ten genes oriented as in Lac.2c, but it encodes an additional hypothetical protein downstream the *lacG* gene. A new variant of the Lac.2 operon was identified in this work, in addition to the three previously described. To be consistent with the current nomenclature, I named this Lac.2d (Fig. 3.2). This variant shows the same genes and orientations as Lac.2a, except that, similar to Lac.2b, it also encodes for a hypothetical protein downstream *lacG*. Thirty-five genomes (29%) encoded the Lac.2a variant, thirty-eight (32%) the Lac.2b, forty-six (38%) the Lac.2c and ten (8%) the Lac.2d. Two isolates had two copies of the operon: both carried a copy of the Lac.2c variant, one a Lac.2a and one a Lac.2b. One isolate (MRI Z2-342) carried only a partial form of the Lac.2c with no *lacRA* and in one isolate, MRI Z2-172, none of the four

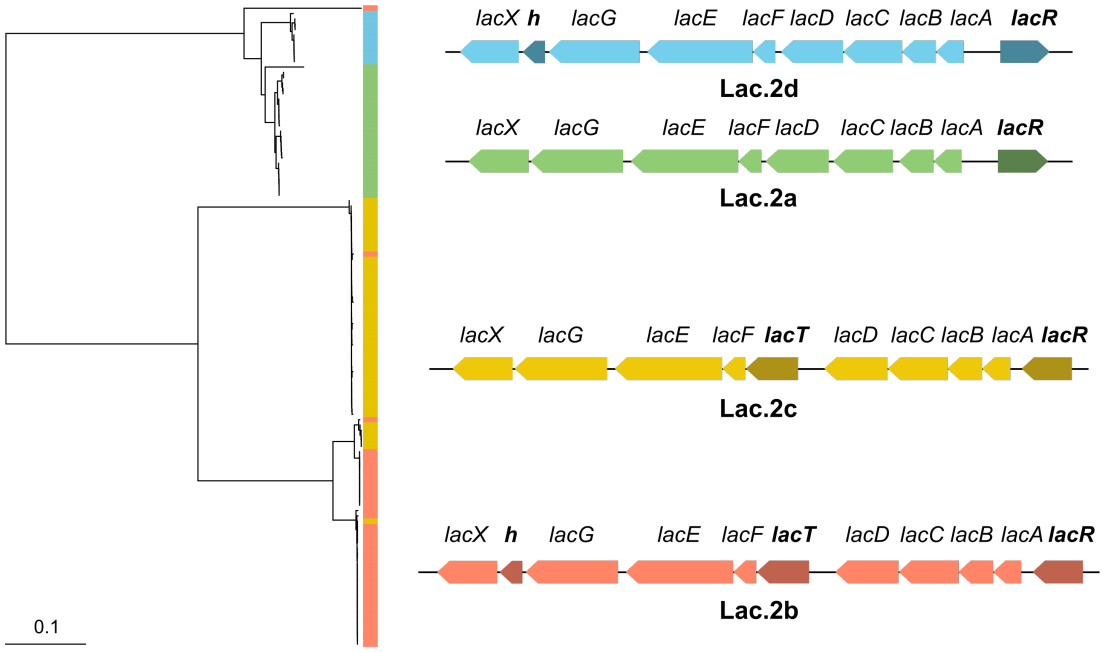


Figure 3.2: Neighbor-Joining phylogenetic tree of four lactose operon variants (Lac.2a, Lac.2b, Lac.2c, Lac.2d), and their genetic organisation (right). Phylogenetic clades and Lac.2 variants do not always match, which indicates that gene losses/acquisitions and rearrangements can be similar between these variants, even if they are phylogenetically distant. Tree was rooted at midpoint.

variants were detected. Considering that this isolate was subjected to long-read sequencing, resulting in a closed genome, this indicates true absence.

The differences in the organisation and number of genes of these variants do not perfectly match the clusters identified on their phylogenetic tree[2], and closely related Lac.2 sequences can in fact belong to two different variants (Fig. 3.2). This was especially observed within clusters Lac.2b and Lac.2c: one Lac.2b variant was observed within the Lac.2c cluster, and a small cluster of Lac.2c was observed within the larger Lac.2b group. In most cases, different variants were identified within the same ST (Tab. 3.1). Only ST196, ST314 and most ST23 isolates encoded for the same variant, Lac.2b, Lac.2b and Lac.2c, respectively.

The amino acid sequences of the integrase located next to the Lac.2 operon (Richards et al., 2011) were extracted from annotation files for phylogenetic analysis. The site of integration was also registered for each of the Lac.2 integrases. It was not possible to recover

[2]Phylogenetic clusters reflect single nucleotide polymorphisms (SNP), i.e. differences at the nucleotide level of a genetic sequence, which does not necessarily correlate with gene presence or orientation.

Table 3.1: Distribution among major sequence types (ST) identified in this work of four genotypic variants of the lactose operon (Lac.2). Lac.2 variants are defined based on their genetic organisation: orientation of the *lacR* gene, and the presence or absence of the *lacT* gene and one hypothetical protein upstream the *lacG* gene.

Lac.2	**Number of isolates**							
genotype	ST1	ST6	ST10	ST23	ST103	ST196	ST314	Other
Lac.2a	2	2	4	-	2	-	-	17
Lac.2b	8	-	-	2	9	4	4	11
Lac.2c	2	1	1	31	1	-	-	6
Lac.2d	3	1	-	1	-	-	-	4

the integrase sequence for 10 genomes, as in these sequences the Lac.2 was found at the edge of a contig, and both integrase and integration site were truncated. For 13 genomes, the integrase sequence was intact, but the site of integration was truncated. Five integration sites were identified and mapped on the complete genomes generated with hybrid Illumina-MinION assembly (Fig. 3.3). The most common integration site was the *N-6 DNA methylase* gene (*n*=58), followed by a *yxdL* gene (*n*=16), a multi-copy gene, and *rbgA* (*n*=9). Less common integration sites were the *lacD* gene from the Lac.1 operon (Richards et al., 2011) (*n*=4) and the *gcvT* gene (*n*=2). Eight Lac.2 integrases were found next to different hypothetical genes. The Lac.2 variants were associated with the integrase phylogenetic clusters in most cases, apart from a few exceptions of Lac.2b and Lac.2c (Fig. 3.3). The integration sites *gcvT* and *yxdL* were interspersed within clusters of *N-6 DNA methylase* genes, showing how very similar integrases can insert in different sites, consistent with results from chapter 2.

3.3.3 Human-associated tetracycline resistant ICE are found in newly-introduced lineages in the bovine population

None of the historical strains harboured any TcR genes, whereas 50% (*n*=37) of the post-90s isolates carried the *tet*(M) gene, which was particularly prevalent among CC1 and ST314 strains (Fig. 3.4). For 78% of these (*n*=29), the *tet*(M) was carried by a Tn*916* element, whereas for 16% (*n*=6) it was located on a Tn*5801*. Two isolates carried the *tet*(M) on a

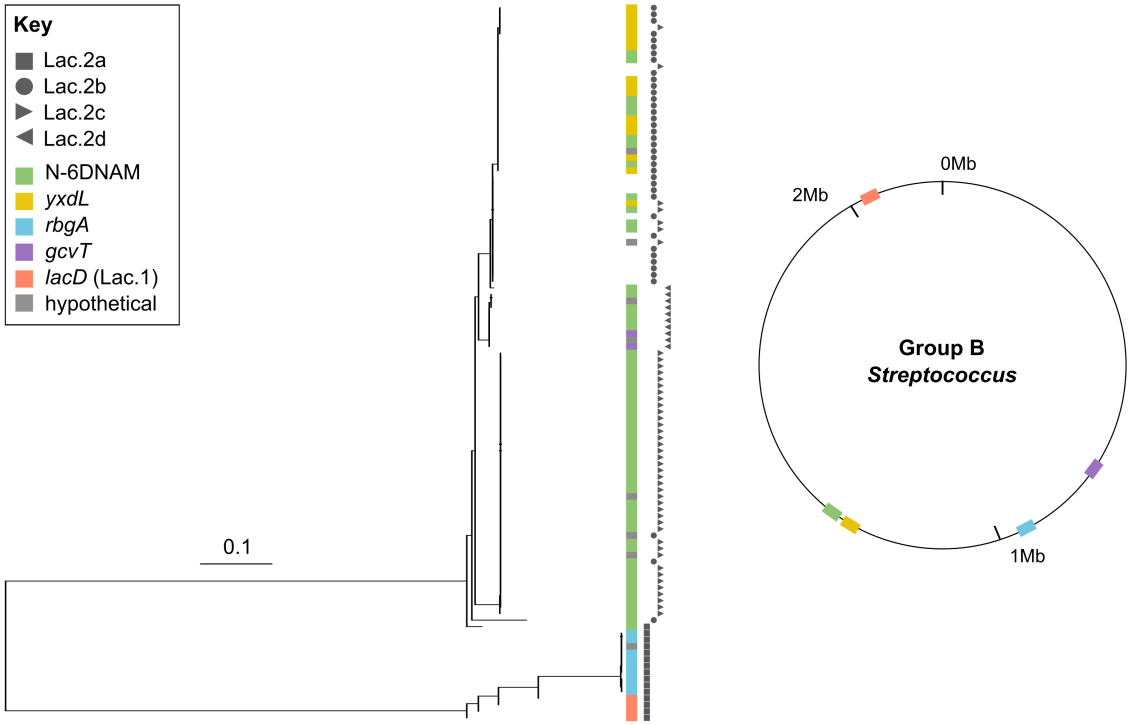


Figure 3.3: Neighbor-Joining phylogenetic tree of the lactose operon (Lac.2) integrase amino acid sequences, with their relative integration site (coloured strip) and Lac.2 variant (symbols). For 13 genomes it was not possible to determine the site of integration, as the integrase was found at the edge of a contig (blank strip). Integration sites have been mapped on an example group B *Streptococcus* genome (right). The *yxdL* gene was found in multiple copies within the same genome, however the Lac.2 was only detected next to the copy present in the region around 1.25 Mbp. Tree was rooted at midpoint.

Tn*5801*-like element, which diverges from Tn*5801* because of the substitution of one hypothetical gene with a different hypothetical gene and one IS256 family transposon (Fig. B.3). Tn*5801* and Tn*5801*-like elements shared the same site of integration (5' end of the *guaA* gene) (León-Sampedro et al., 2016; Da Cunha et al., 2014) and clustered closely in the phylogenetic tree (Fig. B.3), but separate from Tn*916*. The phylogenetic tree of the *tet*(M) gene differed from the whole-ICE phylogeny, with *tet*(M) from Tn*5801* elements clustering within the *tet*(M) from Tn*916* elements (Fig. B.3). TcR in CC1 was predominantly associated with Tn*916* (only one Tn*5801*-like element) whereas TcR in CC103/314 was exclusively associated with Tn*5801* in ST314. TcR in other clades was rare and could be associated with any of these ICE (Fig. 3.4). Other antimicrobial resistance determinants included *tet*(K) (n=3) and *tet*(A) (n=1) and genes for macrolide (*erm*(B) n=1, *lsa*(C) n=5), lincosamide (*lnu*(A)

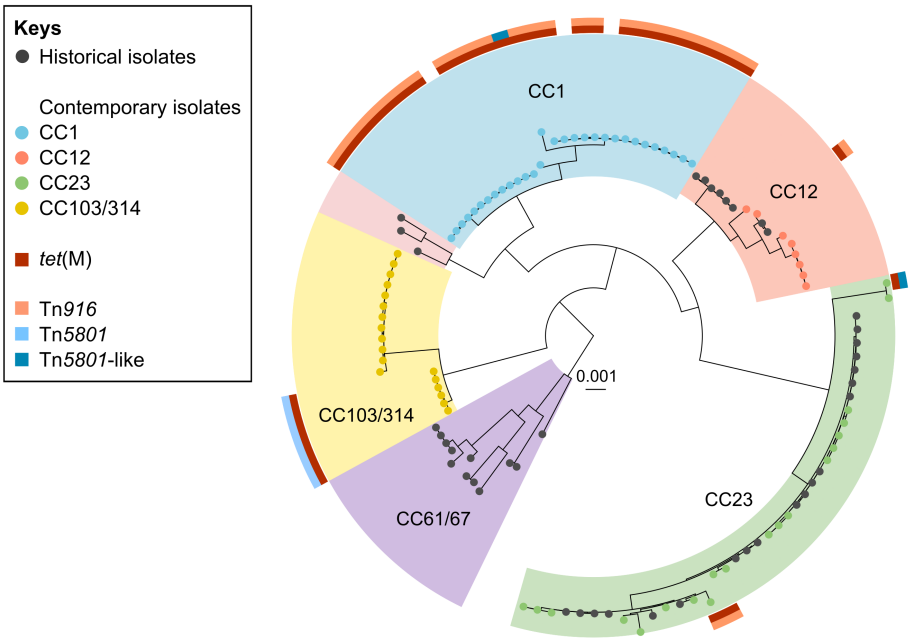


Figure 3.4: Maximum-likelihood core genome phylogeny of 120 group B *Streptococcus* (GBS) showing presence of *tet*(M) and the integrative conjugative element carrying the tetracycline resistance (TcR) gene. For each sequence, the isolation period and CC are displayed based on leaf colour (black = historical, various colours = contemporary) and clade colour range. Association between contemporary isolates, in particular clonal complex (CC) 1 and sequence type (ST) 314, and TcR, *tet*(M) is shown. *Tet*(M) was mostly carried by Tn*916* among CC1 strains and Tn*5801* among ST314 strains. Tree was rooted at midpoint.

n=1), aminoglycosides (*str* *n*=1) and chloramphenicol resistance (*cat*(pCC221) *n*=2).

3.3.4 First plasmids detected in animal GBS show high similarity with plasmids of human pathogenic streptococci

Unicycler was able to resolve 20 complete genomes. Nineteen of these were generated with a hybrid Illumina-Nanopore reads assembly, whereas for one of them (MRI-Z2-182) only a long-read-only assembly was able to generate a full genome (a hybrid assembly generated a sequence of *n*=8 contigs). I was not able to resolve a complete sequence for isolates MRI-Z2-332 and MRI-Z2-340 with either hybrid or long-read-only assembly. Four hybrid genome assemblies were found to have two or more circular sequences, one of which was the circular chromosome and the others were circularised MGE. Four plasmids and one integrative element were identified among four complete hybrid assemblies, belonging to

three different lineages (isolates MRI-Z2-299 and MRI-Z2-265 from CC61/67, MRI-Z2-174 from CC103/314, and MRI-Z2-336 from CC12). One of the plasmids comprised a single replication gene, and was therefore excluded from further investigations. None of the plasmids encoded for antimicrobial resistance genes.

Both isolates in the CC61/67 lineage (MRI-Z2-299 and MRI-Z2-265) carried the same plasmid (pZ2-265, length = 3,617 bp, accession MW118669, Fig. 3.5A), which was found to be similar to plasmid pA996 (Bergman et al., 2014) previously described in *Streptococcus*

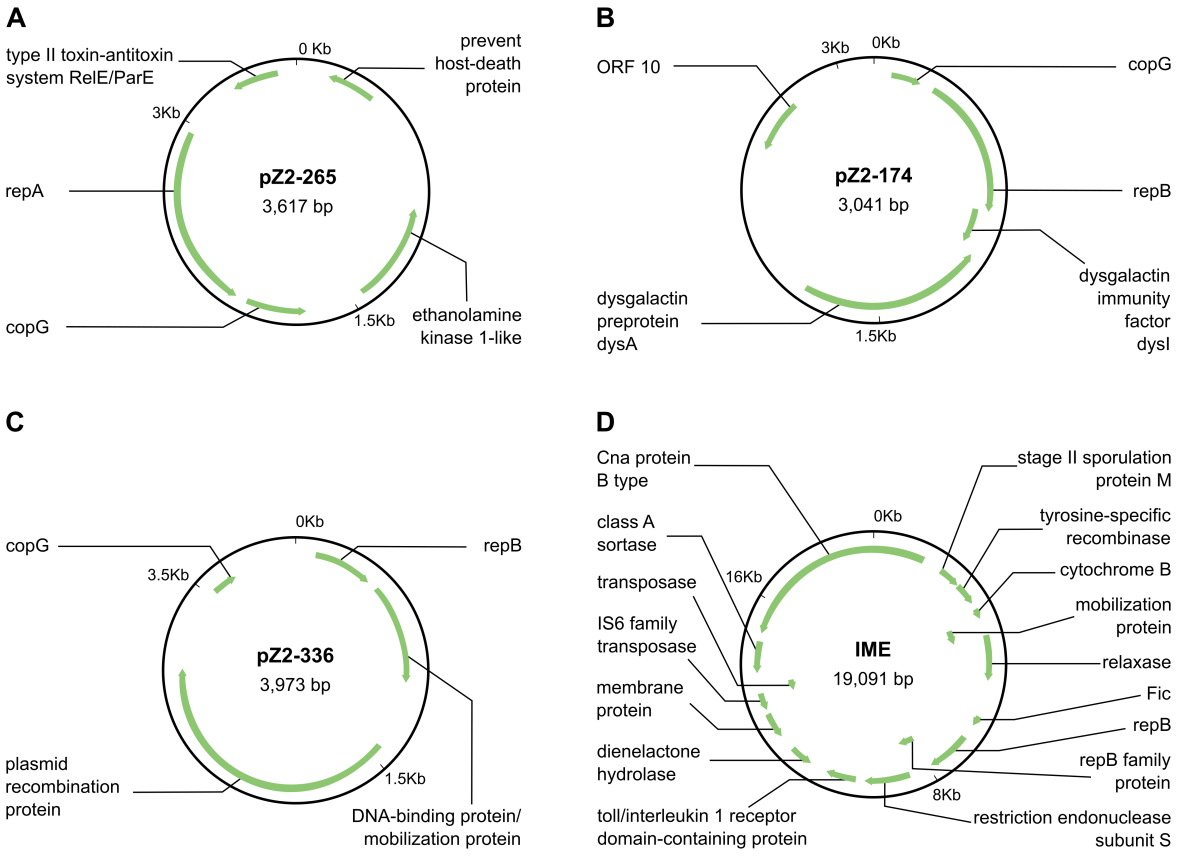


Figure 3.5: Hybrid Illumina-MinION assemblies of bovine group B *Streptococcus* (GBS) genomes revealed the presence of plasmids and integrative mobilisable elements (IME). A) Plasmid pZ2-265 has 99.28% sequence similarity to plasmid pA996 from *Streptococcus pyogenes* (KC895877.1). B) Plasmid pZ2-174 shows 98.85% sequence similarity to pW2580 from *Streptococcus dysgalactiae* subsp. *equisimilis* (AY907345.1). C) pZ2-336 did not show significant similarity with known plasmids, whilst a second circular element in the same genome assembly (D) could either belong to a novel unclassified mobile genetic element family or be an IME.

pyogenes, or group A *Streptococcus*, GAS (QC 100%, ID 99.28%). I used blastn to search for this plasmid in Illumina-only assemblies in my dataset and found it in three more sequences belonging to the same bovine-associated lineage (MRI-Z2-290, MRI-Z2-267 and MRI-Z2-289). The same approach was used to screen 88 publicly available ST61 bovine GBS isolates (Almeida et al., 2016), and 86 of these gave significant long hits, confirming that this plasmid is widely prevalent in CC61/67 and in particular in ST61 GBS strains, even among contemporary isolates. Plasmid pZ2-265 encodes a toxin/antitoxin system, comprising a toxin of the RelE/ParE superfamily, and a prevent-host-death antitoxin (*phd*), which represses transcription of the toxin and prevents host death by binding and neutralising the toxin (Smith, 1996).

A second plasmid (pZ2-174, length = 3,041 bp, accession MW118668, Fig. 3.5B), was found in MRI-Z2-174, an ST314 isolate. This plasmid showed significant similarity with a *Streptococcus dysgalactiae* subsp. *equisimilis* plasmid, pW2580 (QC 99%, ID 98.85%), which encodes the dysgalactin gene (*dysA*), a bacteriocin directed primarily against GAS (Heng et al., 2006), and its immunity factor (*dysI*) (Swe et al., 2010).

The third plasmid (pZ2-336, length = 3,973 bp, accession MW118670, Fig. 3.5C) was identified in MRI-Z2-336, an ST8 contemporary isolate, and it did not show significant similarity with any known plasmids. It encoded genes for plasmid mobilisation and recombination but no genes involved in bacterial protection or toxicity. In the same genome assembly, a second circular element was detected (length = 19,091 bp, accession MW118671, Fig. 3.5D). This element showed features of ICE/IME (tyrosine recombinase/integrase, relaxase), plasmids (plasmid mobilisation protein, plasmid replication initiation protein *repB*) and IS (IS6 family transposase). Additionally, it encoded genes with functions of cell adhesion (Cna protein B-type domain superfamily) and virulence factor expression (class A sortase). ICEFinder (M. Liu et al., 2018) identified a segment of this element (length = 11,068bp) as a putative integrative mobilizable element (IME). Hence, this newly described element could either belong to a novel unclassified MGE family or it could be an actual IME.

3.4 Discussion and conclusions

I investigated the structure of the GBS population isolated from dairy cows in Sweden in order to better understand its evolution over a period of six decades. The prevalence of GBS within dairy herds in this country declined and fell below the detection threshold during a 20-year period, from the late 1970s to the late 1990s (Fig. 3.1). This was achieved mainly thanks to the introduction of mastitis control programs in the 1960s (Nielsen & Emanuelson, 2013). From the late 1990s, GBS started to be isolated again at increasing rates in dairy farms in Northern Europe and it is now described as a reemerging pathogen in dairy cattle in several countries (Lyhs et al., 2016).

To test the hypothesis that the near-elimination and re-emergence of GBS in the Swedish dairy cattle population was associated with lineage replacement, I inferred ST and analysed phylogenetic clusters generated from the genomes of 120 GBS isolates from bovine milk, including 44 historical isolates collected from 1953 to 1978 and 76 contemporary isolates collected from 1997 to 2012. GBS detection in milk was exceedingly rare in the intervening period, and no stored isolates were available for typing. The historical isolates comprised three major lineages: the bovine-adapted lineage CC61/67, which was only detected up to 1967, and two host-generalist lineages, CC23 and CC12. These latter two were found among both historical and contemporary isolates. Major lineages among contemporary isolates were CC1, CC103/314 and CC23. In particular, CC1 and CC103/314 were only detected from 2002 onward.

Lineage CC61/67 was first recognised in the UK (Bisharat et al., 2004) and it is widespread in cattle in the USA (Richards et al., 2019) and Portugal (Almeida et al., 2016). With the exception of three recent cases in China (L. Li et al., 2018), CC61/67 has never been reported in people. Its absence from humans may be due to pseudogenisation of the operon that encodes the polysaccharide capsule, an important virulence factor in human but not bovine GBS infections (Almeida et al., 2016). Without alternative host species, elimination of CC61/67 from the cattle population would mean that no reservoir is left, precluding re-emergence and explaining its absence among contemporary isolates. Not much is known about the origin or fate of CC297, which is a rare type in humans as well as animals.

In contrast, the newly-introduced contemporary lineage CC1 is common among human isolates, including in Sweden (Lyhs et al., 2016; Luan et al., 2005). It has recently been recognised as a common cause of bovine mastitis in northern Europe (Lyhs et al., 2016; Jørgensen et al., 2016; Zadoks et al., 2011) and elsewhere (Cobo-Ángel et al., 2019). The other contemporary-only lineage CC103/314 is recognised as a human pathogen in Asia, including Thailand (Boonyayatra et al., 2020), Taiwan (Hsu et al., 2019), and China (Wu et al., 2019). In cattle, it is found across multiple continents, with reports of CC103/314 as a common lineage among bovine isolates from China (Y. Yang et al., 2013), Colombia (Cobo-Ángel et al., 2019), Denmark (Zadoks et al., 2011), Norway (Jørgensen et al., 2016), Finland (Lyhs et al., 2016) and Sweden (Fig. 3.4). Re-emergence of GBS may be due to cessation of control activities once near-elimination was achieved (Heymann, 2006). In Northern Europe, changes in animal husbandry and transmission patterns may have contributed to GBS re-emergence (Lyhs et al., 2016; Jørgensen et al., 2016), and the lineage replacement shows that re-introduction of GBS must also have occurred.

A large proportion of isolates belonging to the two newly-introduced lineages CC1 and CC103/314 carried the TcR gene *tet*(M), which in CC103/314 was exclusively associated with Tn*5801* (Fig. 3.4 and Fig B.3). TcR is rare among bovine isolates but very common among human isolates (Richards et al., 2019). Indeed, the human GBS population is dominated by a few TcR lineages that expanded after the introduction and extensive usage of tetracycline in human medical practice in the 1940s (Da Cunha et al., 2014). I interpret the presence of TcR in newly emerged bovine GBS lineages as an indication that those lineages have a human origin. In human GBS, it is estimated that CC1 acquired Tn*916* with TcR around 1935 (Da Cunha et al., 2014). Tn*5801* carrying TcR was acquired by human GBS around 1920 for CC17 and around 1950 for CC23, with no year reported for CC10 (Da Cunha et al., 2014). Since their acquisition, TcR determinants have persisted in the human GBS population even in the absence of selective pressure, presumably as a result of low fitness cost (Da Cunha et al., 2014). In contrast to CC1 and ST314, ST103 isolates were *tet*(M)-negative and did not harbour Tn*916* or Tn*5801*. ST103 was first reported in a guinea pig, a cat and a bovine isolate (Brochet et al., 2006). Studies focusing on GBS from dairy cattle identified ST103 in Denmark (Zadoks et al., 2011) and China (Y. Yang et al., 2013),

and later on in Brazil (Carvalho-Castro et al., 2017). Following these, reports of ST103 from cases of carriage or diseased human patients were published (Boonyayatra et al., 2020; Hsu et al., 2019; Wu et al., 2019; de Aguiar et al., 2016). At present, it is hard to reconstruct the evolutionary history of ST103 as well as its geographical origin, as limited genomic data on ST103 has been produced so far. However, considering the abilities of these isolates to survive well in water sources and in the farm environment (Jørgensen et al., 2016; Zadoks et al., 2011), probably also thanks to biofilm production (Pang et al., 2017), it is likely that the success of ST103 in the Nordic countries was at least partially due to its spread through the newly described environmental transmission cycle.

Two lineages, CC23 and CC12, were identified among both historical and contemporary bovine isolates (Fig. 3.1 and Fig. 3.4). CC23 is a common cause of bovine mastitis in northern Europe, whilst CC12 is less prevalent (Lyhs et al., 2016; Zadoks et al., 2011). Both lineages are considered to be multi-host, as they also affect humans and terrestrial and aquatic animals, including homeothermic species and poikilothermic species, e.g. seals and crocodiles, respectively, for CC23, or dolphins and fishes, respectively, for CC12 (Richards et al., 2019; Leal et al., 2019; Delannoy et al., 2016). Within CC12 historical and contemporary strains mostly cluster separately, in sub-clades that correspond to different ST. It could be argued that ST6 and ST12 were eradicated by the end of the 1970s and that new ST, such as ST10, have been recently introduced from a different source. However, considering the small number of isolates available from this clade compared to CC23, the only other lineage detected in both time periods, it is hard to tell with certainty whether ST10 was not circulating within the cattle population between the 1950s and the 1970s, and conversely if ST6 and ST12 are definitely not circulating at present. Multiple serotypes are associated with both CC12 and CC23 (Richards et al., 2019). For CC23, serotype Ia is primarily found in humans and serotype III in cattle (Sørensen et al., 2019; Lyhs et al., 2016). In this study, isolates from CC23 mostly belonged to serotype III although a few serotype Ia isolates were detected in both eras. In general, detection of host-generalist lineages among historical and contemporary isolates could reflect ongoing low-level transmission in cattle during the interim period, as suggested by the dominance of serotype III in CC23 and by the low genetic diversity between historical and contemporary isolates (Fig. 3.4). Alternatively or addition-

ally, it could be due to sporadic reverse zoonotic transmission, as suggested in studies from Colombia (Cobo-Ángel et al., 2019), Denmark (Sørensen et al., 2019; Mweu et al., 2012), and the USA (Dogan et al., 2005), and compatible with occasional detection of CC23 isolates with the predominantly human-associated serotype Ia.

The vast majority of the isolates in this study showed features of adaptation to the bovine niche, represented by the lactose-fermenting genes encoded by Lac.2 (Tab. B.1) (Richards et al., 2013, 2011), which corresponded to phenotypic lactose fermentation (as described in Lyhs et al., 2016); this finding includes newly-introduced lineages that are thought to have a human origin (CC1 and ST314). Phylogenetic analysis showed that closely related Lac.2 sequences can belong to different variants and multiple Lac.2 variants were identified within most ST (Tab. 3.1). The heterogeneous distribution of Lac.2 and the diversity of integration sites illustrates the high genome plasticity of GBS (Richards et al., 2019), which facilitates acquisition of accessory genome content and migration between host species.

The application of long-read sequencing technologies (Oxford Nanopore MinION), allowed me to discover novel plasmids in GBS, which are rarely reported in human GBS and which had never been described in animal GBS prior to this study (Richards et al., 2019). Two of these plasmids showed high sequence similarity with plasmids from human-pathogenic streptococci: GAS and *S. dysgalactiae* subsp. *equisimilis*, which co-exists with GBS in the human oropharynx (Davies et al., 2005). pZ2-174 may provide a survival advantage to GBS when competing for the same niche with GAS, thanks to its anti-GAS bacteriocin. Exchange of plasmids or other mobile genetic elements between GAS, GBS and *S. dysgalactiae* subsp. *equisimilis* is possible in the human oropharynx (Davies et al., 2005) and could potentially be followed by human to bovine transmission of GBS, as documented in epidemiological and evolutionary studies (Richards et al., 2019; Dogan et al., 2005). Finding two plasmids previously associated with other human streptococcal species in bovine GBS isolates suggests that reverse zoonotic events (i.e. human-to-bovine spill-over) have occurred more than once.

3.4.1 Final remarks

Although often described as an obligate intramammary pathogen of dairy cattle in the veterinary literature, GBS is a multi-host pathogen and a host-species jumper with diverse habitats on- and off-farm (Richards et al., 2019; Jørgensen et al., 2016; Lyhs et al., 2016). Evolutionary evidence shows that human-to-bovine jumps are twice as likely as migration in the opposite direction (Richards et al., 2019). Here, I provide evidence that elimination of a major bovine-adapted lineage (CC61/67) in Swedish dairy cattle was followed by emergence of new lineages that carry evolutionary evidence of human origin in the form of TcR markers (Da Cunha et al., 2014), suggesting introduction of human lineages into the cattle population through reverse zoonotic transmission. Subsequently, these new lineages likely established themselves in cattle with the acquisition of the lactose operon (Lac.2) (Richards et al., 2011), which represents the most important marker of the bovine-specific GBS accessory genome known to date. This sequence of events is supported by the fact that TcR is largely retained even in the absence of selective pressure (Da Cunha et al., 2014), such as in the Swedish dairy industry where antibiotic usage is low. The lactose operon does not appear to be retained outside of the bovine host (Sørensen et al., 2019; Lyhs et al., 2016). Thus, TcR and Lac.2 provide historical, or long-term, and recent, or short-term, 'records' of host adaptation, respectively.

Due to the unique historical nature of this isolate collection, direct comparison with genomic sequences of human isolates from the same area and era was not possible. Such comparisons, however, are not necessary for evolutionary analysis, whereby host species jumps have commonly been inferred based on sequence data of isolates derived from different host species without known interactions or epidemiological relatedness (Shepheard et al., 2013; Weinert et al., 2012; Lowder et al., 2009). For the emergence of GBS in farmed species, several routes of transmission from humans to animals can be envisaged, including, in the case of cattle, the handling and milking of cows, which may lead to direct human-to-animal transmission (Cobo-Ángel et al., 2019; Sørensen et al., 2019; Dogan et al., 2005). Changes in animal husbandry systems combined with pathogen evolution are the likely explanation for the re-emergence of GBS, which has been observed in several countries in Europe (Lyhs et al., 2016; Jørgensen et al., 2016; Katholm et al., 2012).

Of the two emerging lineages in cattle, CC1 is known to co-circulate in the human and bovine populations in northern Europe (Lyhs et al., 2016). By contrast, CC103/314 is common in dairy cattle on multiple continents but rare in humans, with the exception of Asia. Despite its low prevalence in humans, CC103/314 may have emerged in cattle due to a spill-over event, with subsequent amplification in modern dairy systems. There is precedent for such a chain of events, as there is reasonable evidence that GBS ST283, which is rare among human GBS isolates, emerged in aquaculture during its intensification in Asia as the result of spill-over from humans, with acquisition of fish-associated MGE facilitating this process (Barkham et al., 2019; Delannoy et al., 2016). Host switching exposes GBS to different selective pressures and sources of accessory genome content (Richards et al., 2019), including plasmids, as demonstrated for GBS, GAS and *S. dysgalactiae* subsp. *equisimilis* in the human oropharynx (Davies et al., 2005), and other MGE, as demonstrated for Lac.2 in GBS, *Streptococcus uberis* and *Streptococcus dysgalactiae* subsp. *dysgalactiae* in the bovine udder (Richards et al., 2011). As farming systems, host contact structures, and selective pressures change, new strains and transmission routes of GBS may continue to emerge through zoonotic and reverse zoonotic transmission, potentially erasing the success of decades of disease control efforts or creating new threats to animal and public health. Control of GBS and other multi-host pathogens will require ongoing monitoring of pathogen diversity across host species and adaptive management in response to changing selective pressures and emergence of new pathogen strains. In addition, this study highlights the importance of having strong biobanking systems in place, with rational and systematic archiving of bacterial isolates together with their metadata, both in large (e.g. the Biological Resource Center of the Institut Pasteur) and in smaller institutions, such as SVA. Biobanking is now recognised as a critical resource to the study and to our understanding of bacterial populations and their evolution, of disease dynamics and pathogenesis, and of public health threats such as AMR (Harris et al., 2012). In light of the much advocated One Health approach, which strives to "achieving optimal health outcomes recognising the interconnection between people, animals, plants, and their shared environment" (source: CDC), it is therefore essential that policy makers and stakeholders are made aware of the importance of funding biobanking systems not only in human medicine, but also in veterinary medicine.

Chapter 4

Host-specialist and generalist lineages of group B *Streptococcus* are associated with distinct patterns of core and accessory genome content and variable levels of genome plasticity

4.1 Introduction

Group B *Streptococcus* (GBS) is a multi-host pathogen which primarily affects three major host groups: humans, bovines and fish. The global GBS population is complex and comprises several lineages which can vary in their ability to cause disease in these hosts. Whilst some lineages are known for being exclusively or predominantly limited to one host group (host-specialists), others are often reported in more than one (host-generalists).

The host range of a bacterial lineage can be influenced by variation in the core genes, which is the subset of genes that are present in all isolates of a given dataset[1], or differences

[1]In principle, the core genome represents all essential genes of a given bacterial species; this will be the

in the accessory genes, the set of genes that are variably present in a group of isolates, which includes mobile genetic elements (MGE). Variation in the core genes can arise from mutations leading to single nucleotide polymorphisms (SNP), from insertions/deletions, or from homologous recombination. Homologous recombination, which is considered an important evolutionary driving force in streptococcal species (Lefébure & Stanhope, 2007; Brochet et al., 2006), is a genetic rearrangement in which DNA is exchanged between two similar or identical sequences, whereas non-homologous recombination (or 'illegitimate recombination') takes place between DNA segments that do not share sequence similarity (gene acquisition) (Frost et al., 2005). Homologous recombination causing core gene variation does not only depend on natural transformation and acquisition of external 'naked' DNA, but it can be directly affected by MGE through horizontal gene transfer (HGT). As an example, MGE in *Staphylococcus aureus* have been shown to drive recombination hotspots in the core genome (Everitt et al., 2014). Additionally, MGE can sometimes carry copies of certain chromosomal genes belonging to the donor cell, and instigate homologous recombination in the receiver cell. This has been observed in *S. aureus* bacteriophages, which can package part of the host cell core genes and transfer them to a new cell, where they swap with the original chromosomal genes thanks to specialised, generalised and lateral transduction (Chen et al., 2018). Therefore, detection of extensive homologous recombination in the core genes, especially when involving different clonal complexes (CC)/lineages, can actually be an indication of HGT (Murray et al., 2017; Everitt et al., 2014), and more generally of high genome plasticity. Host-shifts have not only been linked to homologous recombination events affecting core genes, as shown in *S. aureus* (Murray et al., 2017; Spoor et al., 2015), but they can also be caused by acquisition of new useful genes, often carried by MGE, which confer a fitness advantage within a specific environment (S. K. Sheppard et al., 2018). This mechanism has been shown in several bacterial species, like *S. aureus* (Richardson et al., 2018; Guinane et al., 2010; Viana et al., 2010; Lowder et al., 2009) and *Campylobacter jejuni* (S. K. Sheppard et al., 2013). These accessory genes can range from specific metabolic pathways, which promote the utilisation of a substrate that is particularly abundant in a certain niche[2], as shown

case if the dataset analysed is sufficiently large and comprehensive.

[2]A niche can be defined as a certain biological activity space in which an organism exists in a particular habitat (Wetzel, 2001). In this chapter, the term 'niche' can refer to a particular host, tissue tropism, or both (e.g. bovine-adapted isolates are not only adapted to cattle, which represents the host niche, but within cattle

in cattle for vitamin B_5 biosynthesis genes in *C. jejuni* (S. K. Sheppard et al., 2013) and as demonstrated for Lac.2 lactose-fermenting genes in GBS (Richards et al., 2013, 2011), to virulence factors such as *scpB* in human GBS (Gleich-Theurer et al., 2009) and host-specific coagulase genes in ruminants and equine species in *S. aureus* (Guinane et al., 2010; Viana et al., 2010). As the acquisition and loss of accessory genes through non-homologous recombination of MGE can influence not only host-species range and niche-adaptation, but also virulence and antimicrobial resistance (AMR) profiles (S. K. Sheppard et al., 2018), analysis of accessory genome content provides precious information towards understanding bacterial population dynamics and evolution, which should not be discarded as was often standard practice in past bacterial genomic studies (McNally et al., 2016).

Host-specialist lineages are particularly well adapted to a certain host/niche, and they usually carry host-associated accessory genes, as shown in *S. aureus* (Richardson et al., 2018). Host-specificity can have variable nuances, from host-predilection (i.e. almost complete association with one host, although occasional isolation from other hosts is possible), such as for CC17, primarily isolated from humans (Seale et al., 2017; Manning et al., 2009), to host-restriction (i.e. exclusive association with one host species/group), as observed for CC552 in fish/poikilotherm species (Kawasaki et al., 2018; Barony et al., 2017; Rosinski-Chupin et al., 2013). Niche-restriction is often a consequence of genome downsizing through gene loss and/or pseudogenisation (i.e. reductive evolution), as observed in CC552 (Richards et al., 2019; Rosinski-Chupin et al., 2013) and it is usually linked to ancient host specialisation events. This phenomenon has been shown in other bacterial species, such as *Salmonella enterica* subsp. *enterica* serovar Gallinarum and Pullorum in poultry (Langridge et al., 2015), *S. enterica* subsp. *enterica* serovar Typhi in humans (Parkhill et al., 2001) and *S. aureus* lineage CC133 in ruminants (Guinane et al., 2010). Reductive evolution is likely responsible for the inability of these lineages to escape their preferred host, as they lost useful genes for successful colonisation and survival in other hosts (i.e. they have a narrow gene pool) (S. K. Sheppard et al., 2018). Pseudogenisation can affect host range even when it involves a limited number of genes. As an example, CC61/67 in GBS, which is known to be almost exclusively associated with dairy cattle[3] (Richards et al., 2019; Almeida et al., 2016; Bisharat

they are also adapted to the mammary gland epithelium, which represents the tissue/organ niche).

[3]With the exception of three recent human cases in China (L. Li et al., 2018), CC61/67 has never been

et al., 2004), shows pseudogenisation of the capsular operon (*cps*) genes (Almeida et al., 2016). As the capsule is an important virulence factor in humans, but less so in cattle, this mechanism has been associated with the almost complete absence of CC61/67 from the human population. However, the underlying genetic mechanisms that explain host-restriction in other GBS host-specialists, such as the camel-specific sequence types (ST) (ST609, ST616, ST617 and others), are still unclear (see chapter 6). Whilst host-specialist lineages are highly adapted to particular hosts, host-generalists are more versatile, and they can affect a range of hosts. The existence of such lineages, within bacterial species that comprise host-specialists as well, has been observed in *S. aureus* (Richardson et al., 2018) and *C. jejuni* (S. K. Sheppard et al., 2014). Examples in GBS are CC283 in humans and fish (Barkham et al., 2019; Ong et al., 2018; Kalimuddin et al., 2017), CC1 and CC12 in humans and cattle (Lyhs et al., 2016; Seale et al., 2017), and CC7 in cattle and fish (Delannoy et al., 2016). In *S. aureus*, the ability of host-generalist lineages to infect multiple host species has been attributed to the presence of particular combinations of accessory genes/MGE which confer a more generalist host tropism (Richardson et al., 2018). The genetic phenomena that could influence the ability of host-generalists to adapt to multiple host groups, which include MGE acquisition and homologous recombination, have not yet been fully investigated in GBS.

As described above, accessory genes carried by MGE can promote the survival of a bacterial cell in a specific environment, resulting in niche-adaptation, but MGE are often associated with fitness costs (Dahlberg & Chao, 2003). In addition, some MGE, such as virulent bacteriophages, are detrimental to bacterial cells. Therefore, bacteria evolved several mechanisms to protect themselves from invading MGE, such as CRISPR (clustered regularly interspaced short palindromic repeats) and RMS (restriction modification systems). These are both DNA defence mechanisms that cleave alien DNA which integrates in the host chromosome, and RMS in particular can be considered rudimentary immune systems (Rodic et al., 2017). Notably, RMS have been recognised as major drivers in shaping bacterial populations, especially in the maintenance of heterogeneity. They are variably present among distinct lineages within bacterial species (lineage-specific RMS), as shown for *Neisseria meningitidis* (Budroni et al., 2011), *S. aureus* (Lindsay, 2010) and *Staphylococcus pseudintermedius*

reported in other host species.

(Brooks et al., 2020), and they can be carried by MGE, as a way of promoting MGE survival (Sánchez-Busó et al., 2019). They are also involved in adaptation to different environmental conditions (Ershova et al., 2015), thus exerting a role in host-adaptation as well. RMS have been classified based on their gene composition (presence of a DNA-methyltransferase gene M, an endonuclease gene R, and a DNA recognition protein S) and mode of action (Ershova et al., 2015). RMS recognise unmethylated (non-self) DNA and cleave it from the chromosome, thus preventing permanent integration of horizontally acquired genes (Rodic et al., 2017; Ershova et al., 2015). However, RMS have also been recognised to induce recombination and genomic rearrangements when creating double stranded DNA breaks (Asakura et al., 2011; Rocha, 2004; Handa et al., 2001). As RMS directly impact on MGE acquisition and recombination, they could be playing a role in the variable levels of host-specificity observed among GBS lineages, on homologous recombination, and on population structure.

Knowledge about the host range of lineages comprised in a bacterial species, i.e. whether they are host-generalist or host-specialists, and about their genetic potential to adapt rapidly to multiple hosts, is important in particular as a matter of public health. If a bacterial lineage shows signatures of host-restriction to an animal species, such as GBS CC552 in fish, it is highly unlikely to have zoonotic potential, thus posing a low threat to human health. On the other hand, if isolates from a lineage can easily be transferred between hosts, such as GBS CC1 in humans and cattle (Cobo-Ángel et al., 2019) and CC283 in humans and fish (Barkham et al., 2019), these represent a higher threat to human health, as they could be acquired not only from human-to-human transmission, but also from animal-to-human. In addition, the impact of human-to-animal transmission, or reverse zoonotic transmission, should not be underestimated (as described in chapter 3).

The identification/classification of distinct lineages/subpopulations of genetically related isolates, which are then characterised based on various attributes (e.g. host-range or tissue tropism, virulence and AMR genes, among others), can be challenging. It can sometimes be hard to clearly delineate a subpopulation, and to assign a certain isolate with intermediate genetic characteristics to one lineage or the other, especially considering the impact of homologous recombination on similarity between isolates. That being said, several tools exist nowadays to tackle this issue. Since the advent of next generation sequencing (NGS),

a large amount of genetic information is available, and bacterial populations can be studied based on both their core and accessory genes. Most NGS genotyping methods are based on core genes, the simplest and most widespread one being the seven-gene multilocus sequence typing (MLST). MLST assigns isolates to 'absolute' categories, or ST, based on an allelic profile of seven housekeeping genes (Maiden, 2006), either obtained from sequenced PCR fragments or from extraction from NGS data. A clustering algorithm named eBURST (Feil et al., 2004) was specifically designed to infer evolutionary relationships among bacterial isolates based on MLST data, identifying groups of similar isolates that belong to the same CC[4]. Some studies in the past reconstructed phylogenetic trees from MLST data. However, inferring phylogenetic relationships purely based on MLST trees can be misleading, as MLST represents only a minimal proportion of the whole core genome. As shown in GBS by Sørensen et al., 2010, phylogenetic trees resulting from different sets of housekeeping genes can lead to disparate inferences about phylogenetic relationships between isolates. However, as more isolates are sequenced, the boundaries between CC have started to disappear in GBS, with most CC being connected to others through single locus variants (SLV), tending towards a single broad CC. It is therefore evident that identification of lineages based on these methods has limitations, and, with the advancements of NGS, methods based on the whole repertoire of core genes within a set of isolates have become the gold standard. In particular, core genome phylogenetic trees can reconstruct relationships between isolates with a higher resolution compared to MLST trees. However, identifying groups of similar isolates (subpopulations) even within a phylogeny based on all core genes can sometimes be challenging (Tonkin-Hill et al., 2019), as boundaries between phylogenetic clusters might be blurry. To this end, several genetic clustering algorithms were developed, such as BAPS (Bayesian Analysis of Population Structure) (Corander et al., 2008; Corander & Marttinen, 2006; Corander et al., 2003), hierBAPS (Cheng et al., 2013) and fastbaps (Tonkin-Hill et al., 2019). A limitation of these methods is that they do not assign isolates to a consistent nomenclature of subpopulations, which makes comparison of populations detected in different studies unfeasible, unless these are coupled with information on CC/ST.

[4] A clonal complex (CC) groups together a cluster of biologically meaningful sequence types (ST) that have diverged recently from a founding (ancestral) genotype (or allelic profile).

In addition to core genes, bacterial subpopulations can be defined based on their set of accessory genes. Various tools for pangenome analysis and identification of presence/absence of accessory genes exist, such as roary (A. J. Page et al., 2015) and panaroo (Tonkin-Hill et al., 2020). Presence/absence matrices can then be used to calculate distances between isolates in terms of their accessory genome content, or to run statistical analyses, named genome-wide association studies (GWAS), to identify genes associated with a particular phenotype of interest. In order for these statistical tests to be unbiased, but also for obtaining a representative picture of a bacterial population, it is important to operate a rigorous evaluation of which isolates are included in the analyses. When studying population-wide genomic phenomena, it is crucial to include a diversity of isolates, not only in terms of genetically determined characteristics such as ST and serotype, but also in terms of host, country and year of origin. It is also important not to over-represent lineages that are preferentially characterised in the medical literature. For example, in GBS there is a bias towards the sequencing of invasive human isolates (primarily neonatal); these are consequently the most abundant sequences in databases, but they are not necessarily the most abundant isolates existing in nature. As an example, human carriage of GBS in healthy individuals is estimated around 20-40% (Seale et al., 2017), but most scientific literature focuses on neonatal invasive disease, making these isolates the most well-represented in public databases. Rigorous sequence selection is not always implemented within large genomic studies, which often tend to select isolates more on the basis of availability than on other criteria; this was the case in Richards et al., 2019, the most comprehensive comparative genomics study of GBS until now. In my work, a particular focus was dedicated to dataset curation in order to select a representative subsample of the sequenced GBS isolates available to date and to reduce sampling bias.

The aim of this study was to gain insight into the genetic phenomena that might be playing a role in shaping GBS lineages, through the investigation of the GBS population structure. In particular, I was interested in identifying the underlying genetic profiles and patterns that could be impacting on GBS ecology and on the different levels of host-specificity observed in its various lineages. To this end, an evaluation of the population structure based on the whole repertoire of core and accessory genes was carried out in this chapter, including specific investigations into homologous recombination and RMS. To detect specific host-

associated genes, GWAS were carried out in the following chapter (chapter 5).

4.2 Materials and methods

All supplementary material for this chapter, including tables and figures, can be found in Appendix C (these are indicated with the letter C in front of the sequential number).

4.2.1 Dataset curation

A total of 1,913 GBS genomes were collected for this project. Sequence data was in part generated at the Wellcome Sanger Institute (WSI) (Prof Mark Holmes), in part at the Genome Institute of Singapore (GIS) (Prof Swaine Chen)[5] and in part obtained from public repositories in the form of either assembled genomes or, when possible, raw reads. The latter and the GIS data were assembled with SPAdes v3.14.0 (Bankevich et al., 2012), whereas WSI data were assembled with velvet v1.2.10 (Zerbino & Birney, 2008) as part of the WSI pipeline. As much metadata as possible were gathered, with specific attention to the following features: country of origin (for large countries such as Canada or the U.S.A., province or state was also registered when the information was available), host-species, origin/clinical manifestation (e.g. carriage vs invasive), year of isolation and farm of origin (this was applicable to most bovine and fish isolates, but only to a few human isolates for which herds-persons had been sampled together with their livestock).

As the focus of my PhD project was to work on host-associated accessory genome content, I aimed at maximising the total number of animal genomes relative to the human GBS, which represented the majority of the assemblies. Therefore, I applied a filtering algorithm

[5]New data generated at the GIS derived from isolates from Prof Swaine Chen, Dr Nguyen Ngoc Phuoc (Hue University) and Dr Wanna Sirimanapong (Mahidol University). New data generated at Wellcome Sanger Institute (WSI) comprised isolates from Dr Jørgen Katholm (Knowledge Centre for Agriculture, Denmark), Dr Ulrike Lyhs (University of Helsinki, Finland), Prof Karin Persson-Waller (National Veterinary Institute, Sweden), Dr Katrina Bosward (University of Sydney), Dr Derek Brown (NHS Greater Glasgow and Clyde, University of Glasgow), Prof Andrew Smith (University of Glasgow) and Dr Nguyen Ngoc Phuoc. All new sequence data that are part of this chapter will be made publicly available upon publication in peer-reviewed journals.

as follows (a diagram summarising dataset curation and subsequent analyses described in this chapter can be found in Appendix C, Fig. C.1). First, capsular serotype was determined for all assemblies with a blast approach (Metcalf et al., 2017) and ST was extracted with SRST2 v0.2.0 (Inouye et al., 2014). New ST were found among both published and WSI generated data (*n*=237). Human GBS genomes with new ST (*n*=207) were excluded from the analysis, whereas novel ST from animal genomes (*n*=30) were submitted for ST assignment at pubMLST (`https://pubmlst.org`). Second, Pandas v1.1.3 (McKinney, 2011) was used to select unique combinations of country (down to the province/state level), host-species, origin/clinical manifestation, year of isolation, farm of origin (when applicable), ST and serotype with the drop_duplicates() method. This was done to maximise the genomic diversity in the dataset while avoiding the introduction of multiple assemblies of clonal origin. Some publicly available data included in my work had been published in the context of studies that focused specifically on one serotype or one ST from the same area (e.g. studies in Canada selecting for ST1 isolates such as Flores et al., 2015, or for serotype IV such as Teatero, McGeer, et al., 2015 and Teatero, Athey, et al., 2015). Incorporating all of the genomes from these studies would not have added significant information from a genetic variation standpoint, while it would likely have increased analysis run-time. After this filter was applied, 874 genomes were retained for further analysis.

A genome assembly quality control was run with QUAST v5.0.2 (Gurevich et al., 2013). Reference ranges for the GC content (%) and total genome length (bp) were calculated as their mean $\pm$ twice the standard deviation (2SD), and for the total number of contigs as its mean + 2SD (reference range GC%: 34.32-36.46; reference range total length: 1,492,199-2,681,139; reference range total number of contigs: $<$271) (with the script described in Appendix B, subsection B.2.2). Genomes that were outside at least one of these ranges were explored further (*n*=24), and later omitted from the analysis (Fig. C.2 and Fig. C.3). Nine of these low-quality assemblies showed recognisable contamination with other bacterial species (e.g. *Staphylococcus* spp., *Legionella* spp., *Enterococcus* spp., *Proteus* spp.; Tab. C.1) when run through KmerFinder v3.0.2 (Clausen et al., 2018; Larsen et al., 2014; Hasman et al., 2014). In addition, four low-quality entries had not produced assemblies of sufficient length ($<$6,500 bp) from the published raw reads (SRR2068051, SRR2451878,

SRR2068045, SRR8052453). The quality control filter lead to the selection of 850 high-quality genome assemblies. Of these, 96 were published as assembled genomes, 254 had been assembled with velvet and 500 with SPAdes.

The final dataset comprised genomes from nine host groups/sample types: human (*n*=420), bovine (*n*=277), fish (*n*=101), food market fish samples (*n*=26), camel (*n*=9), dog (*n*=6), sea mammals (*n*=6), frog (*n*=4) and goat (*n*=1) (Fig. C.4). Isolates originated from 37 countries (Fig. C.5) across six continents (Fig. C.6). They comprised 10 serotypes, of which the most well-represented was serotype III (*n*=235) (Fig. C.7 and Fig. C.8), and a few non typeable (NT) isolates. The filtered dataset included 154 ST, with the most common being ST1 (*n*=95), ST23 (*n*=81) and ST283 (*n*=77). Year of isolation spanned from 1953 to 2019. A complete list of isolate names and associated metadata can be found in Appendix C, Tab. C.2.

4.2.2 Core genome analysis

I used Prokka v1.14.5 to generate annotation files (gff) within the WSI pipeline; these files represented the input for panaroo v1.2.0 (Tonkin-Hill et al., 2020), which was used to create a core genome alignment. From this alignment, I extracted SNP sites with the script snp_sites v2.5.1, available within the WSI server. Maximum-likelihood (ML) phylogeny was reconstructed with IQTREE v1.6.10 (L. T. Nguyen et al., 2015), with a general time-reversible (GTR) model, from the core SNP alignment file created with snp_sites (Fig. C.1).

For identification of clusters of genetically similar isolates, I used fastbaps v1.0.4 (fast hierarchical Bayesian analysis of population structure) (Tonkin-Hill et al., 2019) within RStudio v1.3.1093 (Allaire, 2012), R v4.0.3 (R Core Team, 2013) (for all the commands used, see Appendix C, subsection C.2). I selected fastbaps as it is computationally less demanding and more efficient (i.e. it scales well to large datasets) compared to other existing clustering algorithms such as BAPS (Corander et al., 2008; Corander & Marttinen, 2006; Corander et al., 2003) and hierBAPS (Cheng et al., 2013). Input files for fastbaps were the core SNP alignment from snp_sites and the ML phylogenetic tree from IQTREE. The results of the clustering algorithm were exported and visualised within iTol (Letunic & Bork, 2006).

For the classification of lineages/fastbaps populations into either host-specialists or host-generalists, it is important to note that since the dataset curation process was based on the selection of unique combinations, the proportion of infrequent isolation events was artificially increased. As an example, in the CC103/314 lineage the human isolates represent a good proportion of the total in my dataset (Fig. 4.1), however, CC103/314 primarily occurs in dairy cattle (Sørensen et al., 2019; Cobo-Ángel et al., 2019; Jørgensen et al., 2016; Lyhs et al., 2016; Y. Yang et al., 2013), with only a few records of isolation from humans (Boonyayatra et al., 2020; Sørensen et al., 2019; Hsu et al., 2019; Wu et al., 2019). Consequently, a classification of the lineages purely based on proportions of the various hosts within each lineage using this dataset would have been inaccurate and misleading. Therefore, the classification was based on previous knowledge on the frequency of occurrence of ST/CC/lineages in the various host species. As CC103/314 is primarily bovine-associated (host-predilection) it was classified as a host-specialist in this work. Similarly, for the two subpopulations of CC23 (Richards et al., 2019) each sub-lineage shows host-predilection towards either humans or cattle, and they were therefore classified as host-specialists. Other host-specialists include CC17, CC22, CC26, CC452, which are all human-specialists (Seale et al., 2017; Campisi et al., 2016; Bisharat et al., 2004), CC61/67, a bovine-specialist (Richards et al., 2019; Almeida et al., 2016; Bisharat et al., 2004), CC609, which I found to be a camel-specific lineage (see Results section, and chapter 6), and the host-restricted CC552 in fish (poikilotherm species) (Kawasaki et al., 2018; Barony et al., 2017; Rosinski-Chupin et al., 2013). The only lineage that was not categorised before the analyses as either specialist or generalist was CC130, a lineage that is rarely but primarily isolated from humans, with one recent record from dairy cattle (Sørensen et al., 2019). Limited information and genomic data are available for this lineage, which made its categorisation challenging. Considering the genomic characteristics of this lineage detected during this study (e.g. core and accessory genome similarities with host-generalists) I classified CC130 as a generalist.

To detect homologous recombination, a core genome alignment file was generated using snippy v4.4.5 (`https://github.com/tseemann/snippy`). Gubbins v2.4.1 with default parameters was used to identify areas of high SNP density, which are likely to correspond to homologous recombination events. I selected snippy to create an alignment file

instead of using the one generated with panaroo as snippy is routinely used to generate the input file for gubbins, whereas alignments generated within programs for pangenome analysis (e.g. roary) are fundamentally incompatible with it (`https://sanger-pathogens.github.io/Roary/`).

4.2.3 Analysis of accessory genome content

To calculate pairwise distances of accessory gene content of isolates (using the Jaccard similarity index), the gene presence/absence matrix generated with panaroo was processed with GraPPLE (Graphical Processing for Pangenome Linked Exploration) (downloaded on 27 April 2021, `https://github.com/JDHarlingLee/GraPPLE`). The resulting file was visualised with Graphia v2.2 (`https://graphia.app`). For a complete list of commands used for GraPPLE and Graphia visualisation refer to section C.2, Appendix C.

All genomes were screened for the presence of RMS with blastn (Camacho et al., 2009), with thresholds for query coverage (QC) and identity (ID) both set at 100%. These strict thresholds were chosen as some RMS included in the database are almost identical, and lower thresholds would have led to false positive results. Reference sequences were obtained from the REBASE database (R. J. Roberts et al., 2005) for the following genes: type II DNA methyltransferases, type II restriction enzymes (RE), type I M subunit genes, type I R subunit, type I S subunit, type III M subunit, type III R subunit, type C, type N and type V genes.

All figures were edited using Inkscape (`www.inkscape.org`).

4.3 Results

4.3.1 Core genome population structure

Fastbaps identified 18 populations (clusters), which largely corresponded to known CC (Fig. 4.1) and mostly aligned with the topology of the core genome phylogeny. Two populations (10 and 18) comprised isolates that clustered in relatively distant parts of the phylogenetic tree, whereas all other populations were coherent with recognisable phylogenetic groups.

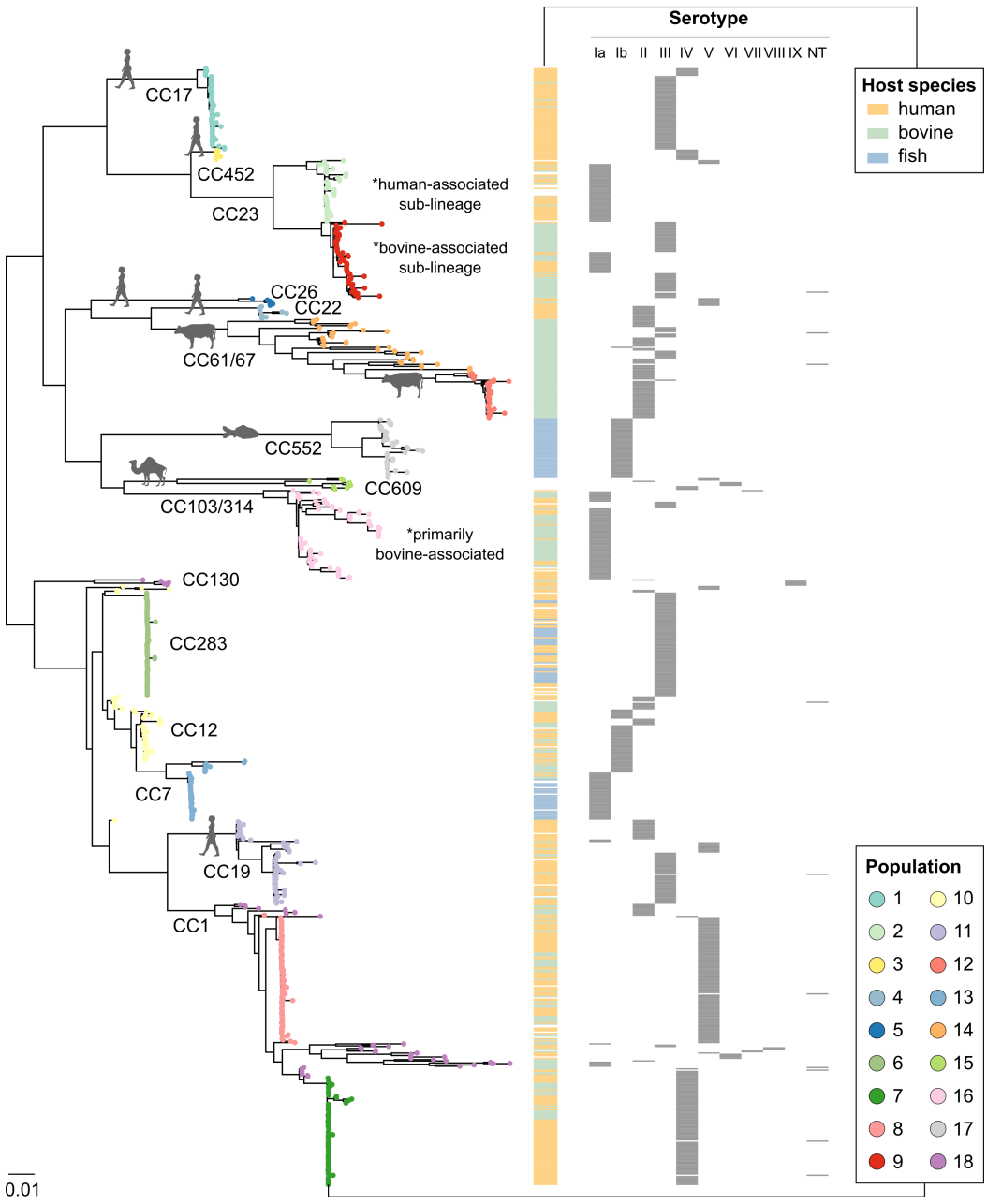


Figure 4.1: Maximum-likelihood phylogenetic tree of 850 group B *Streptococcus* (GBS) isolates. Leaves show the 18 BAPS clusters identified in this study (legend named Population). The coloured outer strip shows the species of isolation (human, bovine and fish), whilst grey blocks show the serotype (NT: non typeable). Host-specialist lineages have been indicated with host icons on their branches. CC23 shows two distinct sublineages: the first is primarily associated with humans (mostly serotype Ia), the second with cattle (mostly serotype III). CC103/314 is considered bovine-associated, as there are few records of isolation from humans (the proportion of human vs bovine genomes was artificially increased by my dataset curation method). Tree was rooted at midpoint.

Population 10 corresponds to CC12, and comprised ST that are normally assigned to this CC (ST8, ST9, ST10, ST12, ST41, ST590, ST652) and one genome assigned to ST7 (serotype V). Population 18 belonged to two CC, CC130 (ST130 and ST104) and CC1 (mostly ST1 and ST2); these two CC appear more separated in the phylogenetic tree compared to isolates in population 10.

Multiple BAPS clusters were comprised within CC23 (populations 2 and 9), CC61/67 (populations 12 and 14) and CC1 (populations 7, 8 and 18). For CC23, which largely comprised ST23 isolates, the two populations (here also referred to as sublineages) correspond to a bifurcation in the phylogenetic tree (Fig. 4.1). The first one (population 2) is mostly associated with human serotype Ia isolates, whilst the second (population 9) is more bovine-associated and primarily belongs to serotype III, except for a subclade of serotype Ia which mostly corresponds to human isolates. For CC61/67 (mostly serotype II), population 12 (a population showing little core genome diversity and mostly comprising ST61 isolates) has likely arisen from population 14, as observed in the phylogenetic tree (Fig. 4.1). All genomes in CC61/67 belong to bovine isolates. CC1 comprises three populations, of which two (population 7 and 8) show little diversity at the core genome level: population 7 mostly comprises serotype IV ST196 and ST459, whereas population 8 is largely dominated by serotype V ST1. By contrast, population 18 shows long branches and a plethora of different serotypes (Ia, II-VIII, NT) and ST (n=14).

Interestingly, when mapping known host-specialist and generalist lineages on the BAPS populations in the phylogenetic tree (rooted at midpoint), these appear divided into two distinct groups corresponding to the two halves of the tree, with the exception of CC19, which is a host-specialist (primarily affecting the human host) that clusters with host-generalist lineages (Fig. 4.1).

4.3.2 Population structure based on accessory genes

On the accessory genes distance network, distinct clusters were observed. Host-species and BAPS populations/CC were overlaid on the network to explore associations with these metadata (Fig. 4.2).

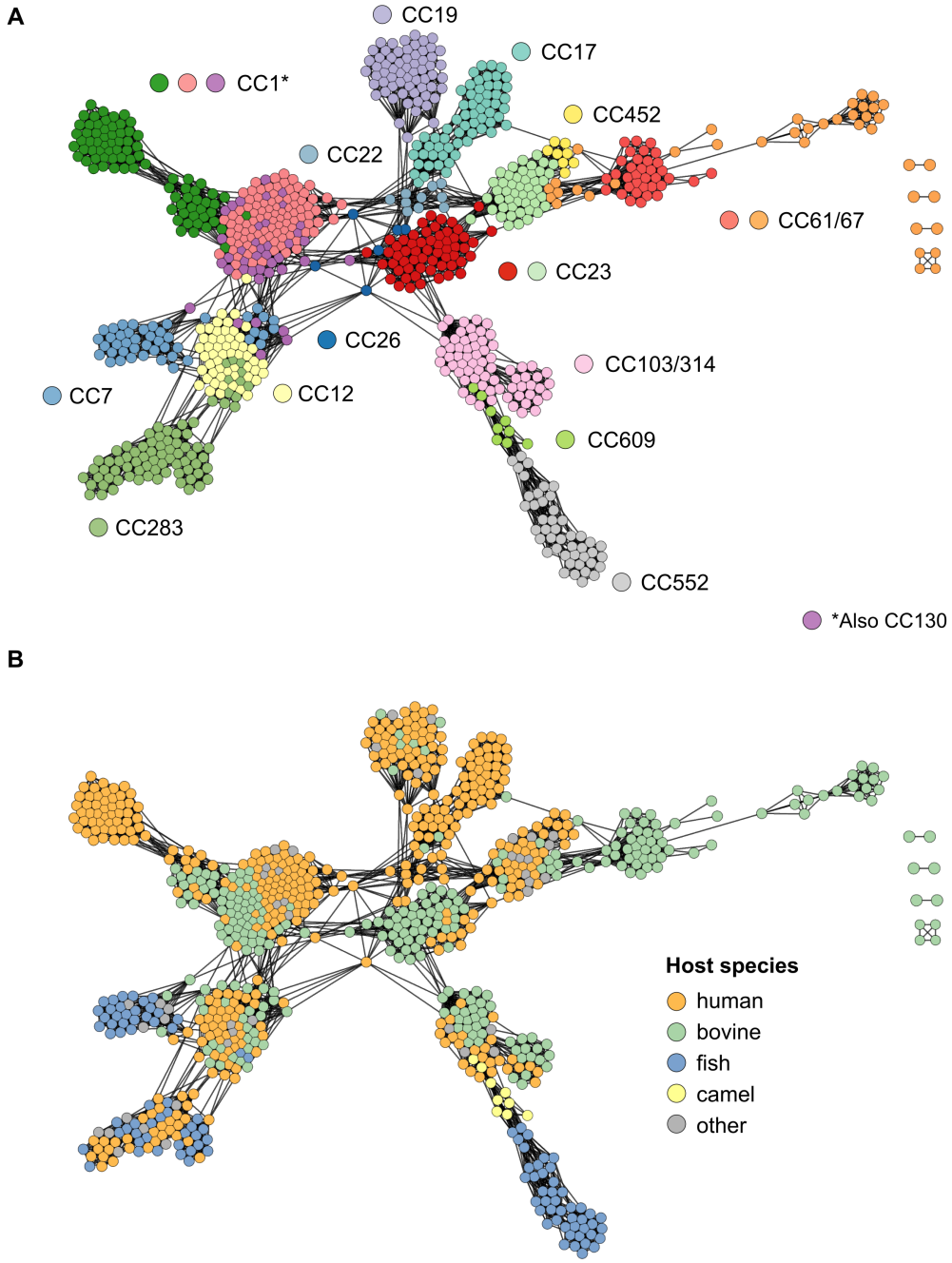


Figure 4.2: Network graph of accessory gene distances between 850 group B *Streptococcus* (GBS) isolates. Accessory gene clusters largely agree with clades from the core genome phylogeny/BAPS clusters/clonal complexes (CC) (colours in panel A are the same as those used in Fig. 4.1), and consequently with their lineage-associated host species. The division between host-specialist (right hand side of network) and host-generalist (left hand side of network) lineages is more evident when looking at accessory genes compared to core genes (Fig. 4.1, Fig. C.9), with CC19 clustering closer to CC17 and other host-specialists.

A high concordance between BAPS populations/CC and accessory gene clusters can be observed (Fig. 4.2A), with a few exceptions. As an example, accessory gene content of some CC7 and CC283 isolates is more similar to that of most CC12 than to their own CC. Similarly, some CC61/67 isolates have an accessory gene profile that is closer to that of CC23 (the human-associated sublineage). As accessory gene clusters mostly align with the core genome phylogenetic lineages, the host-association observed in some of these clusters is a direct reflection of that of the core genome populations (Fig. 4.2B). As an example, three main clusters are associated with fish, and these all correspond to lineages that are known to occur in fish (CC7, CC283, CC552). Similarly, clusters corresponding to CC17 and CC19 are primarily associated with the human host, as these lineages are more common among this species. Of note, within the host-generalist CC1, population 7 (dark green, Fig. 4.2) is split in two clusters based on accessory genes. One of the two is uniquely associated with the human host (Fig. 4.2B) and corresponds to serotype IV ST459 (and its SLV), whereas the other, which is closer to ST1 isolates, is associated with both human and cattle and corresponds to serotype IV ST196.

Additionally, the accessory gene repertoire provides a clearer distinction between host-specialist and host-generalist lineages compared to the core genes (Fig. C.9). In fact, in the accessory gene network, CC19, the only host-specialist lineage (host-predilection for humans) that clustered with host-generalists in the core genome tree (Fig. 4.1), is found on the right hand side of the network, which only comprises host-specialists (Fig. 4.2). By contrast, all host-generalists are found on the left hand side of the network (Fig. C.9). Isolates belonging to CC26, a human-specific lineage, show quite diverse accessory gene profiles which connect the two halves of the network.

4.3.3 Recombination predictions

Large blocks of shared recombination can be observed among all host-generalists, whilst others are limited either to the wide CC1 lineage (populations 7, 8 and 18), or to CC7, CC12 and CC283 together (populations 13, 10 and 6, respectively) (Fig. 4.3). CC19, a lineage that shows host-predilection towards the human host although it clusters within the host-generalist group in the core genome phylogeny, does not share all recombination blocks with

the host-generalists; instead it shows several smaller areas of recombination that are unique to its lineage.

Recombination is almost absent from the majority of host-specialists (upper part of the phylogenetic tree Fig. 4.3), in particular CC17 (population 1), the bovine-associated sublineage of CC23 (population 9) and CC552 (population 17). Smaller blocks of recombination that are limited within lineages can be observed for CC103/314 (population 16) and the human-associated sublineage of CC23 (population 2). Small but numerous blocks of recombination that are also limited within the lineages were detected for CC61/67 (population 12 and 14) and CC19 (population 11).

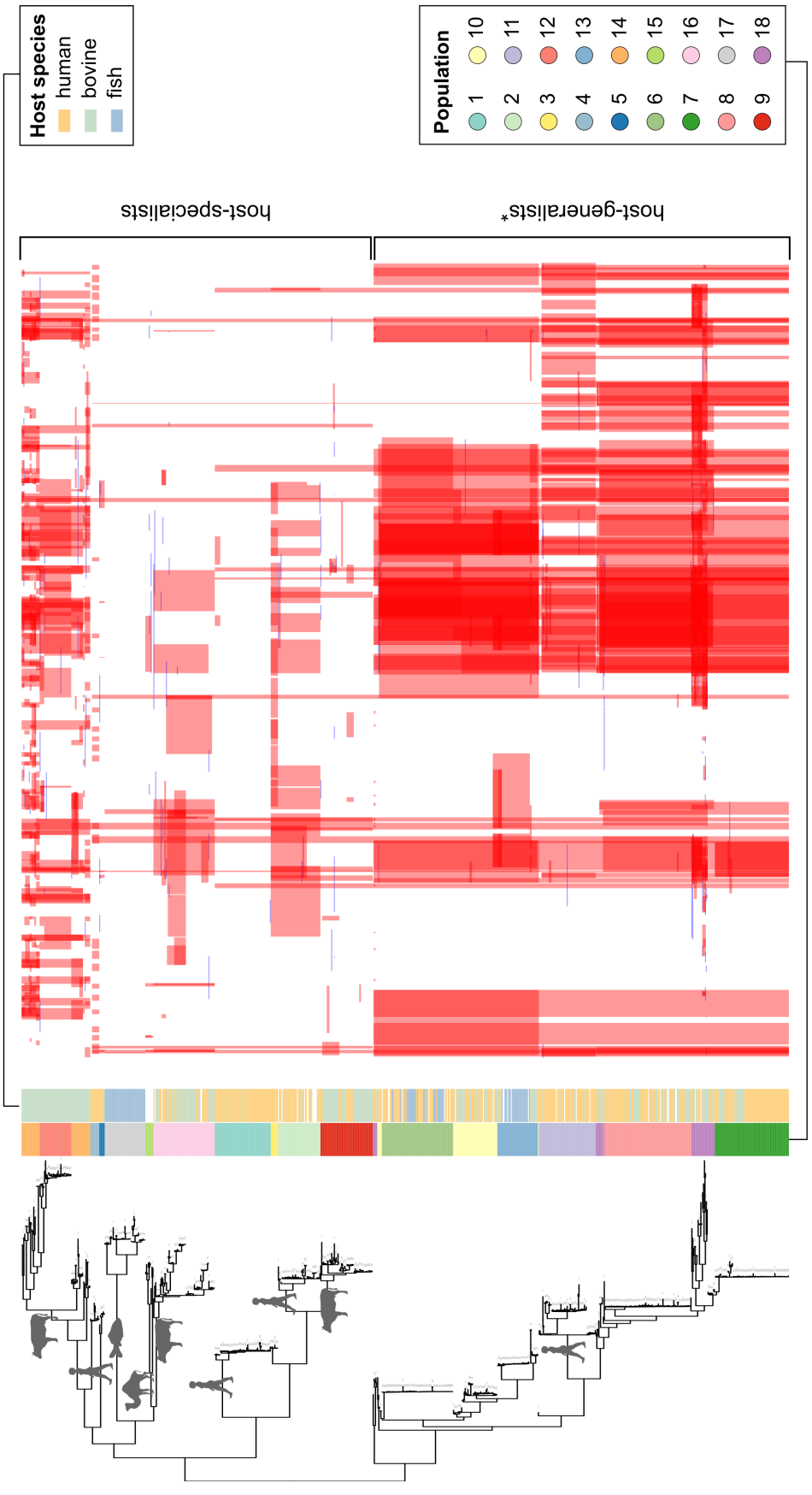


Figure 4.3: Maximum-likelihood phylogenetic tree of 850 group B *Streptococcus* (GBS) genomes. Outer strips show the 18 BAPS populations identified in this study with fastbaps and the species of isolation for the three major host groups (humans, bovines and fishes). Areas of recombination in the core genome detected with gubbins are shown as red blocks (recombination shared between two or more isolates) and blue blocks (recombination found in one isolate only) across the GBS genome (x-axis). Host icons have been indicated on host-specialist branches. Tree was rooted at midpoint.

4.3.4 Detection of restriction modification systems (RMS)

With a few exceptions, type II DNA methyltransferases were detected among almost all genomes (Fig. 4.4), the majority of which encoded between 1 and 6 of these enzymes. Of note, type II DNA methyltransferases were completely absent from population 17 (the CC552 poikilothermic lineage), whose genomes also did not encode for any other type of RMS (Fig. 4.4). Type II RE were also found across much of the dataset (mostly between 1 and 2 per genome), except for population 7 (CC1 sublineage ST459), population 17 (CC552), most genomes in population 12/14 (CC61/67) and population 16 (CC103/314), and some assemblies in population 11 (CC19).

Other types of RMS were limited to certain lineages. For type I RMS, three patterns were observed: isolates either carried a complete system (I S, I M and I R genes), or coded for only two genes (I S and I M genes), or lacked them completely (Fig. 4.4). The first pattern is observed among some host-specialist lineages: population 2 (human-associated sub-lineage of CC23), the majority of isolates in population 16 (CC103/314) and a few genomes from population 12 and 14 (CC61/67). The second pattern was observed among all host-generalist lineages (e.g. CC1, CC7, CC12, C130 and CC283), population 15 (CC609) and a few isolates from population 16 (CC103/314). The third pattern, which means the absence of type I RMS genes, was observed in the following lineages: population 11 (CC19), population 1 (CC17), population 9 (bovine-associated sublineage of CC23), population 17 (CC552), populations 5 and 4 (CC26 and CC22, respectively), and some isolates of population 12 and 14 (CC61/67).

Most genomes among the host-generalist lineages encoded two type I S genes (in particular S.Sag01173ORF8650P fragment and S.Sag1000ORFDP fragment), whilst the majority of host-specialists lacked these genes (Fig. 4.4). Host-specialist lineages which encoded type I S genes were: population 15 (camel lineage CC609), in which the two genes mentioned above were detected, population 16 (CC103/314), encoding for a different gene (S.Sag009ORFCP), a few genomes in population 12/14 (CC61/67), and all genomes in population 2 (CC23 human-associated sublineage). In these latter two clusters, between 1 and 16, and between 1 and 24 type I S genes were detected per genome, respectively.

A similar distribution was observed for type I M genes, whereby all host-generalists encoded for these enzymes (the majority had two per genome, with M.Sag01173ORF8650P and M.Sag7736ORF1965P fragment being the most common), while they were largely absent from host-specialists. As above, the only exceptions were population 2 (CC23 human-

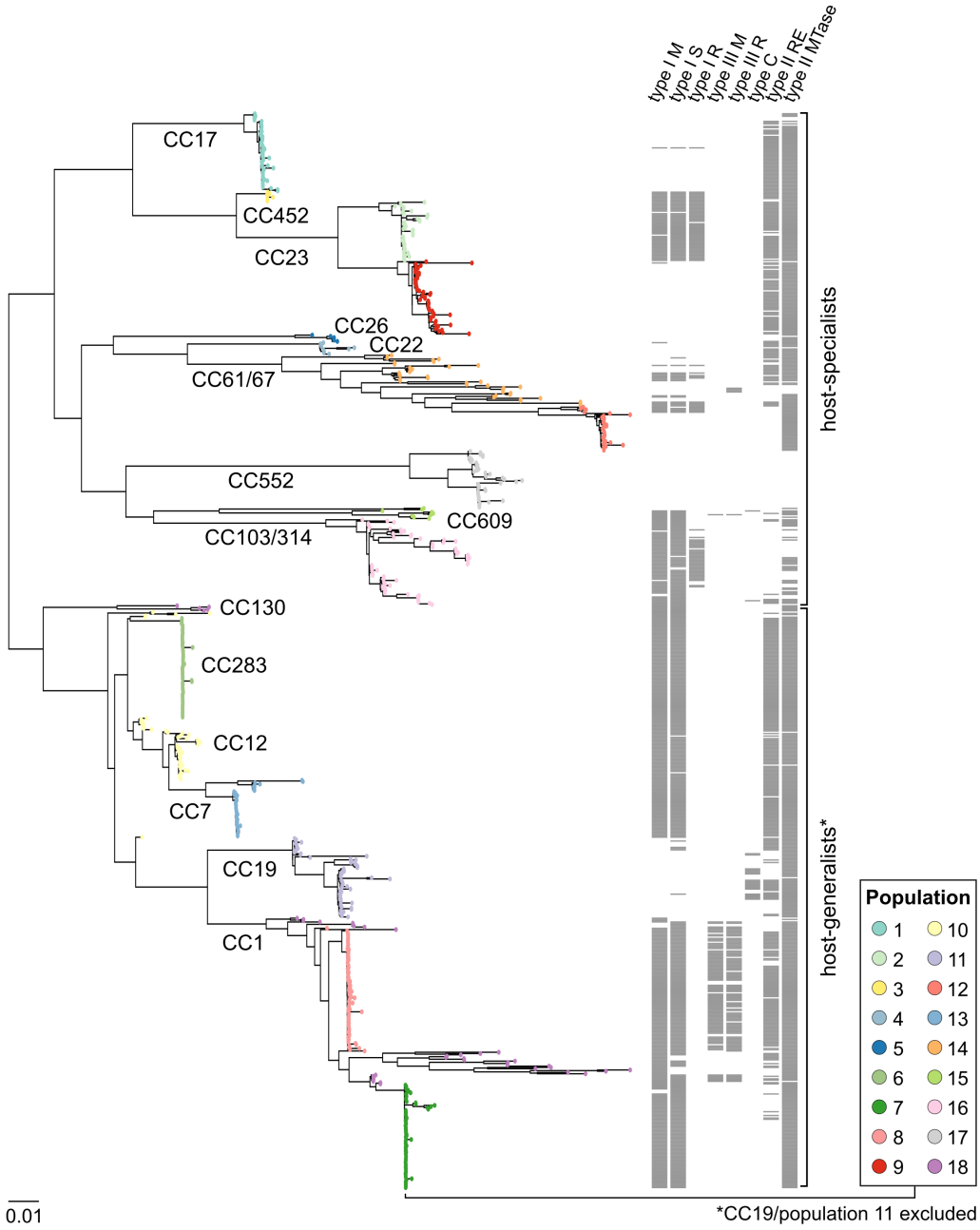


Figure 4.4: Distribution of restriction modification systems (RMS) among the selected dataset of 850 group B *Streptococcus* (GBS) genomes included in this study. Grey and white blocks indicate presence and absence of the different types of RMS, respectively. Methyltransferases are indicated with the acronym MTase. Tree was rooted at midpoint.

associated sublineage) and 16 (CC103/314) (M.Sag512ORF4620P and M.Sag009ORFCP, respectively), and a few genomes in population 12/14 (CC61/67) (M.Sag13813ORF1944P).

RMS limited to specific lineages were also detected (Fig. 4.4). Population 2 (CC23 human-associated sublineage) encoded for one type I R gene (Sag512ORF4620P), as did some genomes in population 16 (CC103/314, Sag009ORFCP). Population 8 (ST1) mostly carried two type III R (Sag37ORF8325P and SagBS13ORF1500P) and one type III M gene (M.Sag37ORF8325P). Finally, type C genes were mostly limited to a small subset of population 11 (CC19) (C.EsaVE80ORF6930P).

No type V and type N genes were detected in this dataset.

4.4 Discussion and conclusions

With the present work, I aimed at finding genomic explanations for the different levels of host-specificity observed across GBS lineages through investigations into the structure of the global GBS population in terms of core and accessory genes, homologous recombination and RMS systems.

4.4.1 Core and accessory genome population structure

With regards to core genome clusters, most of the populations identified with fastbasps in my work (n=18, Fig. 4.1) matched one CC and formed independent lineages (Tab. 4.1), as in Richards et al., 2019, the largest comparative genomic study carried out in GBS so far, in which BAPS v6 (Cheng et al., 2013) was used for population clustering. In contrast to the findings of Richards et al., 2019, the following CC also formed independent lineages in my analyses (Tab. 4.1): CC609 (population 15), CC103/314 (population 16), CC7 (population 13), CC12 (population 10) and CC283 (population 6). In addition, in my results CC7 and CC283 appear as distinct clusters within CC12 in the core genome phylogenetic tree, whereas in Richards et al., 2019, CC283 clustered within CC7. Some CC comprised more than one population: CC23 (population 2 and 9) (in contrast to Richards et al., 2019, in which ST23 isolates are all part of a single population, Tab. 4.1), CC61/67 (population

Table 4.1: Comparison between group B *Streptococcus* (GBS) lineages (clonal complexes, CC) and populations identified with clustering algorithms in two studies (fastbaps was used in this study, BAPS in the study from Richards et al., 2019).

Clonal complex (CC)	Host species	Fastbaps population (this study)	BAPS population (Richards et al., 2019)
CC1	human, bovine	7, 8, 18	1, 11
CC7	human, bovine, fish	13	4
CC12	human, bovine	10	4
CC17	human	1	7
CC19	human	11	3
CC22	human	4	6
CC23	human bovine	2 9	8
CC26	human	5	9
CC61/67	bovine	12, 14	5
CC103/314	bovine	16	10
CC130	human, bovine	18	10
CC283	human, fish	6	4
CC452[a]	human	3	2
CC552[b]	fish	17	12
CC609	camel	15	10

[a]CC452 is considered part of CC23 in Richards et al., 2019; [b]Named CC260 in Richards et al., 2019.

12 and 14) and CC1 (population 7, 8, 18). CC61/67 corresponded to a unique population in Richards et al., 2019; my finding of two subpopulations is likely a result of the rigorous dataset selection process carried out for this work, which, for CC61/67 alone, included a higher number of isolates (n=77 vs n=32), from a higher number of countries (n=10 vs n=6), and from a wide temporal range (year 1953-2014, not reported for most genomes in

the collection from Richards et al.). CC1 separated into only two lineages in Richards et al., 2019; however, a high degree of uncertainty in the BAPS assignment can be observed for a subclade showing long branches (Richards et al., 2019, Fig. 1). Interestingly, population 18 comprises isolates belonging to ST/CC that are found in relatively distant parts of the tree (CC1 and CC130), highlighting possible issues with the categorisation of these isolates. This is highly likely a consequence of the high level of recombination observed among the host-generalists group, which could be blurring the boundaries between lineages. For CC609 (population 15), the camel-specific lineage, previous work based on MLST trees had detected two sub-populations, one of which (ST609 and ST614) clustered close to ST23 and ST17, whilst the other (ST616, ST617 and others) formed an independent distant clade (Fischer et al., 2013). Based on my phylogenetic analysis of the core genome, hence not only limited to the seven MLST genes, I show how these ST form a monophyletic clade (Fig. 4.1). As the geographical area of origin of available camel genomes is restricted to the Horn of Africa (as described in chapter 6), it is not possible to rule out the existence of other camel lineages that are shared with the human host in other parts of the globe (e.g. Australia, India and the Middle East). In addition, as the human population which lives in close contact with these animals (e.g. pastoralist communities) has never been sampled for GBS carriage, the possibility of this monophyletic clade being shared with humans cannot be excluded, although further genomic analyses suggest this clade is specific to camels (chapter 6).

Of note, the core genome phylogeny almost perfectly partitioned lineages that had been categorised as host-generalists from those previously known to be host-specialists (Fig. 4.1). An even more accurate distinction between these two categories can be observed when analysing the accessory gene distances (Fig. C.9). In particular, the accessory gene set tends to be more homogeneous within the host-generalist group compared to the host-specialists. In contrast, the accessory gene content for many of the individual host-specialist lineages diverged significantly from other lineages of the group (e.g. CC552, CC609, CC61/67) (Fig. 4.2). Strikingly, similar patterns were observed in their core genome, as indicated by the deep branches in the core phylogenetic tree for host-specialists compared with those for most of the host-generalist lineages (Fig. 4.1). Among the host-generalists, deep branching was less commonly observed in the core phylogeny, and was limited mainly to a subclade of CC1

(part of population 18). This suggests that a certain set of accessory genes and/or a higher genome homogeneity due to extensive recombination within the group result in a more generalist host tropism, conferring the ability to infect multiple host species (CC1, CC7, CC12, CC130, CC283). Segregation of host-generalists and host-specialists based on both core and accessory genes is highly indicative of these two groups evolving independently from each other, suggesting the existence of a barrier to genetic exchange between the two. This is further supported by the absence of shared homologous recombination between host-generalists and specialists. The greater recombination observed in the core genome of host-generalists across lineages likely corresponds to an overall higher genome plasticity, which could be responsible for their higher level of adaptability to multiple hosts, and suggests they might be more subject to HGT compared to specialists. However, it is still unclear which biological mechanisms are responsible for the differences observed in genome plasticity across GBS lineages. On the one hand, genetic competence, which is the ability of a bacterium to uptake external DNA, could be expressed at variable levels in the different GBS lineages, as shown in *S. pseudintermedius* (Brooks et al., 2020) and *Listeria monocytogenes* (Rabinovich et al., 2012; Loessner et al., 2000), in which lineage-specific bacteriophages are responsible for disruption of genetic competence through integration within the competence genes *comGA* and *comK*, respectively. Alternatively, it could be regulated by particular environmental/niche conditions, such as in the case of nutrient depletion inducing competence in *Haemophilus influenzae* and inhibiting it in *Streptococcus pneumoniae* (Solomon & Grossman, 1996). GBS is not known to be naturally transformable, in contrast to *S. pneumoniae* (Straume et al., 2015) and *N. meningitidis* (Alexander et al., 2004), although it does carry genes for competence (e.g. *comX*, the site of integration of prophage GBS1, as described in chapter 2). On the other hand, distinct lineages could have different 'spectra of sensitivity' towards various types of MGE based on characteristics such as the capsular serotype or the variable presence of DNA defence mechanisms (e.g. CRISPR and RMS). An example of this is the different levels of susceptibility to bacteriophage infection observed across capsular serotypes, with only certain types of phages being able to infect specific serotypes, as described in *Klebsiella pneumoniae* (Haudiquet et al., 2021; J. A. M. de Sousa et al., 2020). Similarly, plasmid conjugation can be influenced by the presence of the capsule, with increased conjugation in isolates in which the capsule is inactivated (Haudiquet et al., 2021).

The variable presence across lineages of a diverse set of defence mechanisms against the integration of MGE, such as RMS, could also be playing a role in the different 'spectra of MGE sensitivity' observed across GBS lineages. A higher susceptibility to e.g. bacteriophages caused by differences in RMS could not only result in the acquisition of new genes (e.g. phage-related virulence, toxin and AMR genes), but it could also lead to a greater degree of homologous recombination in the core genes through specialised, generalised and lateral transduction (Chen et al., 2018). I identified lineage-associated patterns of presence/absence of RMS, in particular differences in type I RMS between host-specialists and host-generalists; the majority of the latter coded for the M (DNA-methyltransferase) and S (DNA recognition protein) genes but never for the R gene (restriction enzyme). Most host-specialists, and in particular the three lineages in which recombination was virtually absent (CC17, CC23 bovine-associated sublineage and CC552), completely lack type I RMS, whereas host-specialist lineages that carried type I M and I S genes also carried the I R gene (CC23 human-associated sublineage, CC103/314 and some CC61/67). The function of the S and M genes is recognition of the restriction site and protection of the DNA from cleavage through methylation, respectively, whereas the R gene cleaves non-self DNA. In the absence of a type I R gene, as observed among host-generalists, incoming external DNA and MGE would not be removed from the chromosome, and their DNA would be protected through methylation. This could explain the higher rates of integration/recombination observed in type I R gene-negative genomes (host-generalists), leading to an increased ability to adapt to different environmental conditions, hosts and niches due to access to a larger gene pool. On the other hand, host-specialists that coded for a complete type I RMS would be less subjected to such DNA exchanges. Host-specialists that lacked these systems could actually be characterised by an overall inability to uptake external DNA (competence) and accept MGE, which is supported by the absence of recombination in these lineages. This is consistent with the fact that RMS are less abundant in non-transformable bacterial species that endure fewer DNA exchanges (Sánchez-Busó et al., 2019; P. H. Oliveira et al., 2016).

In addition to RMS, I also compared the full set of accessory genes across isolates, and how this relates to the population structure and host species of GBS. To my knowledge, no previous studies have evaluated accessory gene distances in a global GBS collection to

find correlation with either lineage or host species. In *S. aureus*, similar analysis of accessory gene distances suggested the existence of a host-specific gene pool necessary for host-adaptation (Richardson et al., 2018). In Richardson et al., 2018, isolates with <50% shared accessory gene content were removed from the analyses. In addition, the paper does not show how lineages/CC map onto the accessory gene distances network graph. I argue that in the work by Richardson et al., 2018, no clear clustering is associated with a particular host species, apart for a few exceptions (birds, horses and pigs) that however also represent single lineages in the core genome phylogeny, similar to my GBS network (Fig. 4.2). My results show that, in GBS, the accessory gene content of an isolate is first correlated with the lineage/CC, and only secondarily with the host group. In fact, a remarkable agreement between core and accessory genome content can be observed (Fig. 4.1 and Fig. 4.2); such a strong correlation suggests the existence of genetic phenomena that limit the interaction of a certain set of accessory genes to its core lineage, especially among host specialists in which recombination is either absent or limited within lineage. In addition, the association of certain accessory gene clusters with a host species is a reflection of the host species associated with the lineage (e.g. CC552 fish). If the accessory gene repertoire was influenced more by the host species than by the lineage, the isolates would form distinct clusters based on host species, independent of the lineage of origin (e.g. one human cluster, with isolates belonging to all human-associated lineages such as CC1, CC17, CC19, CC283, CC12 and others; one bovine, with with isolates belonging to all bovine-associated lineages such as CC61/67, CC103/314 and others, etc.). This points to the fact that a limited number of niche-associated accessory genes, rather than a large set, might be playing a role in host-adaptation in GBS (as investigated in chapter 5), together with other genomic phenomena such as genome reduction leading to host-restriction (Rosinski-Chupin et al., 2013). Investigation of pseudogenisation in lineages other than CC552 and CC61/67, together with evaluation of GBS host-jumps on a timed-scaled phylogeny (as described in chapter 7), would be helpful to gain more insight into the evolutionary history of GBS lineages.

4.4.2 Methods discussion

The dataset curation process, which forms an integral part of the study design, was focused on two main objectives: i) including the broadest diversity of GBS genomes available to date, not only in terms of genetically determined characteristics such as ST and serotype, but also in terms of host, country and year of origin, gathering as much sequenced data and metadata as possible; ii) reducing selection bias which causes over-representation of certain lineages/serotypes/ST/hosts. The majority of available GBS genomes to date are of invasive human isolates; however, this does not necessarily reflect the full diversity of GBS in nature (e.g. carriage isolates, isolates from other hosts species and from the environment). An example of bias caused by inclusion of genomes based on availability is that of Richardson et al., 2018; in their work on *S. aureus*, the authors state a bias towards genomes of human origin (60% vs 40% animal sources) and from Europe. In addition, some studies that sequenced substantial numbers of GBS isolates were focused on specific ST/serotypes in a limited geographical area, therefore they could have introduced a selection bias if used in their entirety for the purposes of my study (e.g. studies in Canada selecting for ST1 isolates such as Flores et al., 2015, or for serotype IV such as Teatero, McGeer, et al., 2015 and Teatero, Athey, et al., 2015). As the aim of my work was to assess the structure of the GBS population, the dataset selection process was targeted at balancing the proportion of human isolates relative to the animal isolates (49% vs 51%, respectively), all the while maintaining the highest diversity possible in terms of genetic profiles (e.g. serotype, ST) and origin of the isolates (e.g. host, country, year). Similar to Richardson et al., 2018, my dataset still presented a European bias after the curation (48% of all genomes). Additionally, part of my curation process was focused on quality statistics for the genome sequences included in this work. To my knowledge, building reference ranges for quality parameters such as GC content, genome length and total number of contigs is not standard practice. I showed how this step is actually very important to ensure the quality of the published sequenced data is of high standard. Among published data I detected some raw reads that were contaminated with reads from other bacterial species, and others that had a very small number of reads. One limitation of my dataset is that the curation process artificially increased the number of rare isolation events, such as the number of CC103/314 in humans relative to dairy cattle. This prevented me from making inferences based on prevalences within the clades. As

an example, the assignment of a lineage to either the host-specialist or generalist categories could not be based on the proportion of human vs animal isolates within that lineage but it had to be based on prior knowledge on which lineages occur more often in the various hosts. The curation process also determined the inclusion of a high number of bovine isolates from Sweden as a result of the availability of genomes from multiple farms and a wide temporal range (chapter 3).

I used fastbaps to assign the isolates to distinct populations as the delineation of boundaries between clades/lineages within a core genome phylogeny can often be challenging (Tonkin-Hill et al., 2019). In my analysis, fastbaps was efficient and rapid in the identification of GBS subpopulations, and it did not require significant computational resources (it was run on a MacBook Pro 2017, 8 GB of memory and 2.3 GHz processor), in contrast to hierBAPS (Cheng et al., 2013), which I tried to run for comparison and which was stopped for excessive run time. Another program that I considered using for population clustering is PopPUNK (Lees et al., 2019), as it gives the option of maintaining a stable nomenclature for the populations when adding new genomes to a pre-analysed dataset. However, PopPUNK is not recommended for clustering of GBS, due to poor performance in this bacterial species (Lees, personal communication). When I ran it on a subset of genomes from this dataset, the program often identified a very high number of clusters in subpopulations that show a high diversity, in particular CC61/67 (n=10 PopPUNK clusters instead of n=2 fastbaps populations).

4.4.3 Final remarks

In conclusion, I analysed the population structure of a large dataset of high-quality GBS genomes that were selected to provide a representative picture of the GBS population, including human and animal isolates. I show how host-generalist and host-specialist lineages largely evolve independently, and that accessory genome content is associated with lineage, rather than host species. Host-specialists exhibit lineage-specific RMS patterns and within-lineage recombination only, or absence of recombination. These genetic signatures, together with distant genetic relationships observed in the core and accessory genome analyses, in-

dicate that these lineages have a low potential for host-switching. Therefore, elimination efforts of these lineages from their preferred host are likely to lead to successful long-term eradication results, as observed for CC61/67 in dairy cattle in Sweden (chapter 3). In contrast to this group, host-generalists have an inherently higher genetic potential for host-jumps due to a higher genome plasticity, as indicated by the extensive shared recombination and lack of type I restriction enzymes. As a limited number of accessory genes are likely to be major drivers of host-adaptation, as shown by the network analysis and results of the next chapter (chapter 5), it is clear how host-generalist lineages pose a threat to GBS control programmes. Host-generalist lineages could rapidly adapt to a new host through uptake of few but crucial host-associated genes (e.g. Lac.2 in dairy cattle), potentially erasing elimination efforts, as shown for CC1 in dairy cattle in Sweden (chapter 3).

Chapter 5

A limited number of mobile genetic elements are associated with host adaptation in group B *Streptococcus*, GWAS reveals

5.1 Introduction

Group B *Streptococcus* (GBS) is a multi-host pathogen with a complex population structure. It comprises host-specialist lineages, which uniquely or primarily infect one species (e.g. clonal complex (CC) 17 in humans, CC61/67 in cattle and CC552 in fish), and host-generalists (e.g. CC1, CC7, CC283), which can variably affect the three major host groups: humans, dairy cattle and fish. GBS genomes show high plasticity and a mosaic structure, with high levels of recombination, particularly among host-generalist lineages (chapter 4), and acquisition of external genetic material and mobile genetic elements (MGE) (Richards et al., 2019; Brochet et al., 2006) (chapter 2). These are thought to have shaped the GBS global population, promoting its adaptation to different niches. The identification of host-associated MGE, such as the *scpB-lmb* composite transposon in humans (Franken et al., 2001), the lactose operon Lac.2 in dairy cattle (Richards et al., 2011), and locus 3 in fishes (Delannoy et al., 2016), are examples supporting the notion that MGE promote adaptation to different

niches[1].

Host-adaptation can be influenced by multiple forces, which can be linked to factors related to the pathogen (e.g. rates of genetic mutation, recombination and genetic exchange, ability to evade the host immune response and to utilise substrates), to the host (e.g. type and effectiveness of host immune response, availability of substrates, competition with the resident microbiota) and to the wider environment (e.g. ecological and geographical/spacial segregation) (S. K. Sheppard et al., 2018). The recent advances of next generation sequencing (NGS) technologies have helped in identifying the genetic mechanisms involved in bacterial host-adaptation. Genetic mutations as small as single nucleotide polymorphisms (SNP), in particular non-synonymous mutations[2], can affect bacterial host ranges, as shown for *Staphylococcus aureus* in rabbits (Viana et al., 2015) and *Salmonella enterica* subsp. *enterica* serovar Typhimurium in pigeons (Kingsley et al., 2013) and in cattle (Yue et al., 2015). These examples show how incredibly few genomic changes may be required to cause host jumps or shifts in host predilection. Host switching events can be followed by the evolution of niche-restricted lineages (host-specialists) through gene loss of function and reductive evolution[3]. This phenomenon is often associated with ancient host-specialisation events, and has been shown in several bacterial species, such as *S. enterica* subsp. *enterica* serovar Gallinarum and Pullorum in poultry (Langridge et al., 2015), *S. enterica* subsp. *enterica* serovar Typhi in humans (Parkhill et al., 2001), *Mycobacterium leprae* in humans (Cole et al., 2001), *S. aureus* lineage CC133 in ruminants (Guinane et al., 2010) and GBS lineage CC552 in fish (Rosinski-Chupin et al., 2013).

In addition to mutations, genetic exchanges/acquisitions play an important role in bac-

[1]A niche can be defined as a certain biological activity space in which an organism exists in a particular habitat (Wetzel, 2001). In this chapter, the expression niche can refer to a particular host, tissue tropism, or both (e.g. bovine-adapted isolates are not only adapted to cattle, which represents the host niche, but within cattle they are also adapted to the mammary gland epithelium, which represents tissue/organ niche).

[2]Non-synonymous mutations are nucleotide changes that result in the translation of a different amino-acid compared to the original sequence, or a stop codon, as opposed to synonymous mutations for which the translated amino-acid is identical to that of the original sequence.

[3]Reductive evolution is the process of genome downsizing through gene loss and/or conversion of genes to pseudogenes.

terial adaptation, through both homologous and non-homologous recombination[4], and acquisition of extra-chromosomal genetic elements, such as plasmids. Acquisition of genetic material through horizontal gene transfer (HGT), often from the host resident microbiota, promotes a rapid adaptation to the new niche. The uptake of large amounts of DNA in a single event can maximise the speed and effectiveness of bacterial remodelling in response to new environmental challenges thanks to gain of advantageous functions. These can range from acquisition of new metabolic pathways (substrate utilisation), as shown in cattle for vitamin B_5 biosynthesis genes in *Campylobacter jejuni* (S. K. Sheppard et al., 2013) and for Lac.2 lactose-fermenting genes in GBS (Richards et al., 2013, 2011), to interaction with the host immune system and increased invasiveness, as suggested for *scpB* in human GBS (Gleich-Theurer et al., 2009). The existence of a host- or niche-specific accessory gene pool, which confers a fitness advantage within a specific environment, has been shown for several bacterial species such as *S. aureus* (Richardson et al., 2018) and *C. jejuni* (S. K. Sheppard et al., 2013). However, isolates that have recently been exchanged between hosts do not always display genetic changes responsible for adaptation. These isolates might just be causing transient infections and never develop adaptation to the host (dead-end) (Nowrouzian et al., 2005; Hohwy et al., 2001); another possibility is that they might be in a transitional state, from non-adapted to niche-adapted, but not yet displaying signatures of host-adaptation, such as the acquisition of useful accessory genes specific to that niche. This has been shown for Lac.2 in GBS, in which host-specialist lineages show clear phenotype-genotype pairs (human-absent, bovine-present), whereas host-generalists show a 'grey area' whereby some isolates do not match these phenotype-genotype pairs (human-sometimes present, bovine-sometimes absent) (Lyhs et al., 2016). The rapid acquisition of host-adapted traits thanks to imported genes is often mediated by MGE, such as bacteriophages, plasmids, pathogenicity islands and integrative conjugative elements (ICE). Examples of this in *S. aureus* are pathogenicity islands (SaPIs, part of the PICI family, described in chapter 1), which encode for host-specific coagulase genes in ruminants and equine species (Guinane et al., 2010; Viana et al., 2010), and its avian-specific MGE repertoire in poultry isolates (Lowder et al.,

[4]Homologous recombination is a genetic rearrangement in which DNA is exchanged between two similar or identical sequences, whereas non-homologous recombination (or illegitimate recombination) takes place between DNA segments that do not share sequence similarity.

2009). In GBS, the acquisition of ICE conferring tetracycline resistance (Tn*916* and Tn*5801*) is thought to have selected for a few clones that spread globally in the human population from the 1960s (Da Cunha et al., 2014).

Comparative analyses of large sets of whole genome sequences can help in identifying underlying genotypes associated with host adaptation (S. K. Sheppard et al., 2018). First employed in human genetics, these genome-wide association studies (GWAS) only recently gained attention in the investigation of microbial genomes. GWAS apply statistical tools to detect genetic determinants associated with a phenotype of interest. Phenotypes can represent anything from antimicrobial resistance (Farhat et al., 2019), to duration of infection (Lees et al., 2017), from virulence (Laabei et al., 2014), to host species (S. K. Sheppard et al., 2013). Several programs are currently available to conduct microbial GWAS, which employ different statistical approaches (San et al., 2020). These tools can be based on SNP, *k*-mers[5], unitigs[6] or gene presence/absence, and they may or may not adjust for population structure. As explained in chapter 4, when running statistical analyses such as GWAS, it is important to select a representative sample of the population being studied. This is not always the case with large genomic studies, which often tend to select isolates purely based on availability. As an example, Gori et al., 2020, ran a GWAS to identify CC-associated genes in GBS; however, the dataset used by the authors only included isolates from five countries, with a considerable difference in the proportion of human isolates compared to animal isolates (96% vs 4%, respectively). In addition, for one of these countries, genomes originated from a study that focused specifically on sequence type (ST) 1 (Flores et al., 2015), introducing a selection bias. Hence, this dataset cannot be considered a good representation of the GBS population, and it likely influenced the authors' findings. This highlights the importance of rigorous choices when selecting the genomes to be included in a study, as was the case with the dataset used for my work (refer to chapter 4), especially in light of the types of analyses to be conducted.

Although some host-associated gene clusters have already been described in GBS (*scpB*

[5] *K*-mers are nucleotide subsequences (substrings) of variable length (*k*) contained within a biological sequence.

[6] Unitigs are defined as high-confidence contigs.

transposon in humans, Lac.2 in bovines, locus 3 in fishes), their detection was based on the comparison of a limited number of genomes available at the time (Delannoy et al., 2016; Richards et al., 2011). The aim of this study was to expand the genomic comparison to a vast, high-quality collection of genomes representative of the global GBS population, in order to identify a more comprehensive set of host-associated genes and MGE. This was attempted through two different GWAS approaches and assessed in light of prior knowledge on GBS host-associated genes.

5.2 Materials and methods

All supplementary material for this chapter, including tables and figures, can be found in Appendix D (these are indicated with the letter D in front of the sequential number).

5.2.1 Dataset curation

The dataset curation for this work and the rationale behind this process is described in chapter 4, section 4.2.

5.2.2 Genome-wide association study (GWAS)

Pyseer: host association computed with linear mixed model based on *k*-mers and unitigs

Prokka v1.14.5 (Seemann, 2014) was used to generate gff annotation files, within the Wellcome Sanger Institute (WSI) pipeline, which represented the input for panaroo v1.2.0 (Tonkin-Hill et al., 2020) (for a comprehensive list of commands, refer to Appendix D, section D.2). The core genome alignment created with panaroo was used to reconstruct a maximum-likelihood phylogenetic tree with IQTREE v1.6.10 (L. T. Nguyen et al., 2015), with a general time-reversible (GTR) model of nucleotide substitution. The phylogeny_distance.py script was used to extract the distance matrix from the phylogeny (all scripts included in the pyseer suite can be found at `https://github.com/johnlees/pyseer`). Fsm-lite v1.0 (`https://github.com/nvalimak/fsm-lite`) was used to count *k*-mers,

and unitig-counter v1.0.5 (`https://github.com/johnlees/unitig-counter`) to count unitigs.

Pyseer v1.3.3 (Lees et al., 2018; Jaillard et al., 2018) is a python implementation of *SEER* (Lees et al., 2016), an alignment-free method for GWAS that is based on non-redundant variable length *k*-mers (or unitigs) to represent variation across the pangenome; linear models with a control for population structure are then used to test for association. In my analysis, pyseer with a linear mixed model was run to find associations between the accessory genome and each of the three major GBS host groups (human, bovine, fish), one at a time vs all other genomes. Two runs, one on *k*-mers and the other on unitigs, were set up for each host species, which alternately represented the phenotype of interest (coded as binary feature, host species of interest = 1, vs other species = 0). For the fish phenotype, a strong lineage effect was found after running pyseer on *k*-mers with default settings (Fig. D.1A); this was likely due to the fact that only three lineages within the entire GBS population include genomes isolated from fish species, as described in chapter 4. Additionally, the number of genomes from this phenotype (*n*=101) was the minimum required for a genome-wide association study, whereas the number of genomes for the other two phenotypes exceeded this value by hundreds (human, *n*=420; bovine, *n*=277). The *k*-mer based analysis for fish suffered from a long run time ($>$3 weeks), so it was decided to continue only with unitigs for this phenotype, as they significantly decrease run time. To control for population structure in the fish unitig analysis as much as possible, different minimum MAF (minor allele frequency)[7] cut-offs were used, with a chosen final cut-off of 4% (Fig. D.1D).

To check whether population structure had been successfully controlled for in the human and bovine phenotypes, quantile-quantile (Q-Q) plots of the expected and observed negative logarithm of the association *p*-values, -$\log_{10}$(*p*-value), were created with qq_plot.py (Fig. D.2). The count_patterns.py script was used to calculate the significance thresholds for association (*p*-value 1.70×10^{-8} for *k*-mers and 2.26×10^{-7} for unitigs).

[7]Minor allele frequency (MAF) is the frequency at which the second most common allele occurs in a given population. Rare genetic variants are defined as those with a MAF $< 5\%$. GWAS are typically performed on genetic variants that have common minor allele frequencies (MAF $> 5\%$).

Only significant *k*-mers/unitigs, that is with *p*-values lower than the threshold, were kept for further steps. Annotation of significant *k*-mers/unitigs, which is the process of mapping the *k*-mers/unitigs to the genes they belong to, was performed with multiple reference and draft genome assemblies (n=6 for human, n=2 for bovine, n=6 for fish, for a complete list see Tab. D.1, Appendix D). I also tried subsequent annotations with all reference genomes in Tab. D.1 for each phenotype and with a pangenome (see discussion section 5.4.2). Tab-delimited files of significant genes and their statistics were created with summarise_annotations.py. All plots were obtained with matplotlib v3.3.2 (Barrett et al., 2005).

Scoary: the pan-GWAS approach

Annotation files obtained as described above (gff format) were used as the input for roary v3.13.0 (A. J. Page et al., 2015). The resulting presence/absence gene matrix, together with a trait file containing the different phenotypes in binary format (as above), was used to run a GWAS with scoary v1.6.16 (Brynildsrud et al., 2016), with default settings. To distinguish it from traditional GWAS, which is based on SNP, scoary's approach has been named pan-GWAS, from pangenome. Scoary is based on gene presence/absence, rather than on nucleotide sequences (SNP, *k*-mers, unitigs). Scoary assigns each variant[8] a null hypothesis (H0 = the gene is not associated with the phenotype of interest). Then, each variant undergoes a series of filters: a population-independent Fisher's exact test, subsequently a population-aware filter (pairwise comparison to find the maximum number of phylogenetically unrelated contrasting pairs, e.g. gene = 0 and trait = 0, vs gene = 1 and trait = 1), and finally a label-switching permutation analysis.

The output of scoary is a list of significant genes per trait, where the best scoring genes are reported first (i.e. those genes that were most associated either positively or negatively with the trait). For each gene, the following statistics are indicated, among others: the *p*-value (naïve, with Bonferroni and Benjamini–Hochberg corrections), number of genomes for phenotype 1 or 0 in which the gene is either present or absent (Tab. 5.1), sensitivity (SE), specificity (SP). Both Bonferroni and Benjamini–Hochberg corrections for *p*-values

[8]In the scoary paper (Brynildsrud et al., 2016) the word 'variant' is used throughout the text to indicate different alleles (gene variants); here, I use the term variant and allele interchangeably.

Table 5.1: Explanation of the nomenclature used by scoary in its output to list the number of genomes for phenotype 1 or 0 in which the gene is either present (+) or absent (-). Letters shown within cells (a, b, c, d) are indicated to facilitate description of formulae used to calculate statistics such as sensitivity (SE) and specificity (SP).

	Phenotype 1	**Phenotype 0**
Gene +	Number_pos_present_in (a)	Number_neg_present_in (b)
Gene -	Number_pos_not_present_in (c)	Number_neg_not_present_in (d)

aim at controlling the false positive rate; however, the Bonferroni correction is often considered too conservative, as it treats all input *p*-values equally, and this can generate a lot of false negatives (Diz et al., 2011; Holm, 1979). Benjamini–Hochberg calculates a threshold which discriminates between false and true positives, based on the ranking of the original *p*-values (Diz et al., 2011; Benjamini & Hochberg, 1995). After inspecting the results for my three phenotypes, I observed that the order in which genes were listed did not change based on the different corrections, hence, in the interest of simplicity, all subsequent scoary *p*-values described in this chapter are naïve. SE is calculated as the proportion of entries from phenotype 1 that have the gene (a/[a+c]), as per Tab. 5.1), whilst SP is calculated as the proportion of entries from phenotype 0 that do not have the gene (d/[b+d]). They express the probability that the phenotype 1 has the gene, and that phenotype 0 does not have the gene, respectively. In addition to these statistics, I calculated positive and negative predictive values (PPV and NPV, respectively), as the proportion of entries with a certain gene that belong to phenotype 1 (a/[a+b]) and as the proportion of entries without that gene that belong to phenotype 0 (d/[c+d]). They express the probability that an entry that is positive for a gene belongs to phenotype 1, and that an entry that is negative for a gene belongs to phenotype 0, respectively.

BLAST+

To verify results, I searched for the presence of high-scoring genes/MGE from both pyseer and scoary outputs in the whole dataset with local blast v2.6.0+. Amino acid sequences

of single genes carried by host-associated MGE were used for Lac.2 (as *lacEG* are genes specific to Lac.2, I used these two genes to detect the presence of the whole MGE; *n*=6 *lacE* alleles, *n*=7 *lacG* alleles), *scpB* transposon (*n*=6 *scpB* alleles, *n*=2 *lmb1* alleles), locus 3 (*n*=17 genes in total, five of which had two alleles), *pezAT* (*n*=3 *pezA* alleles, *n*=6 *pezT* alleles) and *cadDX* (*n*=6 *cadD*, *n*=3 *cadX* alleles). Genome assemblies were scanned with tblastn, and minimum thresholds for positivity were set at 90% sequence identity (ID) and 90% query coverage (QC). I chose to use this threshold to make sure to capture other possible alleles of the genes of interest, which might not have been represented by my database of amino acid sequences; as I was only interested in presence/absence of these elements, if a genome contained multiple alleles of the same gene, I counted it as one (this happened e.g. for *lacE*, for which some variants are shorter than others, but they match a segment of the longer alleles, hence the positivity to the longer variant is a false positive). For IS*Stin5* (*n*=7 unique alleles, identified from scoary) ID and QC were both set at 100%, as I was interested in calculating the total number of occurrences of each allele in the genome (paralogs).

For detection of the ICE identified by pyseer, I used blastn of a shared human-bovine region (12,497 bp, area straddling a LPxTG gene and an N-6 DNA methylase gene) of the ICE. Threshold for positivity was set at 90% ID and 80% QC.

5.3 Results

5.3.1 Pyseer

Pyseer identified the *scpB-lmb* transposon as significantly associated with the human phenotype. This element is enclosed between two IS*Sag2* insertion sequences, and usually comprises the following genes: IS*Lre2*, *scpB*, *lmb*, a pneumococcal histidine triad protein and an IS3 family transposase. This transposon had the highest scores in the unitig-based analysis, particularly for the *scpB* gene, whereas it was less evident in the *k*-mer based analysis (Fig. D.3 and Fig. D.4). The *scpB* -$\log_{10}$(*p*-value) for *k*-mers was 24.49, whereas the *scpB* -$\log_{10}$(*p*-value) for unitigs was 71.07. In addition to this MGE, genes belonging to an ICE had very high scores in the *k*-mer output, with lower scores being reported in the unitig output, but still appearing among the highest peaks in the Manhattan plot (Fig. 5.1). The

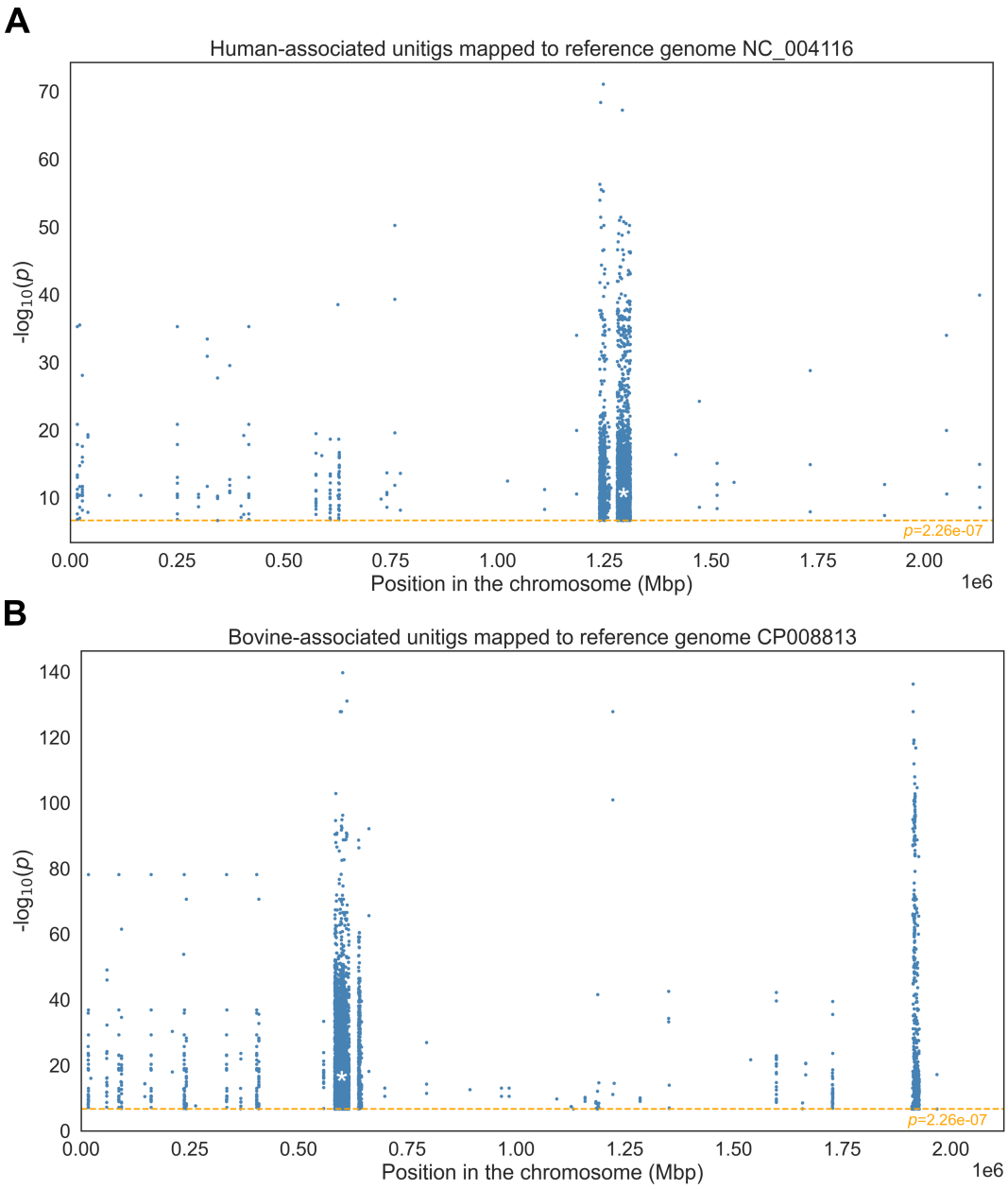


Figure 5.1: Manhattan plots of significant unitigs from the pyseer analyses of human and bovine group B *Streptococcus*. Significance thresholds (dashed lines) and corresponding *p*-values are indicated in orange. Reference genomes used to plot the data were the most representative closed genomes for each phenotype (human: NC_004116; bovine: CP008813). (A) The first large peak (around 1.25 Mbp) corresponds to the *scpB-lmb* transposon, whereas the large genomic island (GEI) marked with the white asterisk in both plots corresponds to an ICE whose genes were negatively associated with the human and positively associated with the bovine phenotype. (B) The linkage of the ICE to the *pezAT* gene cluster in cattle can be observed in the reference genome CP008813 (first peak after the ICE, separated by a blank space that corresponds to the insertion of Tn*916* within the island in this genome). The last high peak in this plot (around 1.9 Mbp) corresponds to the Lac.2 operon.

final pyseer output does not indicate the direction of effect, which means that discriminating between a positively- or negatively-associated gene requires further analysis; to assess direction of effect, I inspected the average beta-coefficient[9] in the annotated_unitigs.txt file. The ICE was found to be negatively associated with human phenotype, whereas it was positively associated with the bovine phenotype (see below). The tetracycline resistance (TcR) gene *tet*(M) was not identified by pyseer as significantly associated with the human host.

For the bovine phenotype, the highest-scoring gene cluster corresponded to Lac.2 (Fig. 5.1, Fig. D.5 and Fig. D.6), in particular its genes *lacEG*, and to the fructose operon, which is in linkage[10] with Lac.2. As mentioned above, the same ICE that was reported in the human output scored highly in the bovine phenotype (Fig. 5.1), but this time it was positively associated with the phenotype of interest. Additionally, in bovine genomes this ICE carried a toxin/antitoxin system (*pezAT*) (Fig. D.7), which was also significantly associated with the bovine host. A blastn of *pezAT* showed how these genes are almost uniquely associated with cattle (Fig. 5.2).

Results for the fish phenotype were unsatisfactory. Although population structure seemed to have been controlled for in the unitig analysis (Fig. D.1D), no evident peaks were present in Manhattan plots for the CC552 lineage. Two peaks were evident uniquely in CC283 genomes, which coincided with some genes corresponding to significant unitigs in the scatter plot (Fig. D.8); these corresponded to the *scpB-lmb* transposon, and to a genomic island (GEI) carrying a cadmium resistance two component system (*cadDX*). Blast confirmed *scpB* as being principally associated with human genomes (Fig. 5.2), and found uniquely in one of the three fish lineages, CC283. The genes *cadDX* were also found to be highly represented in the human phenotype, and much less so among fish genomes (Fig. 5.2). Known fish-associated locus 3 was not identified by pyseer.

[9]The beta coefficient is the regression coefficient which estimates the mean change in phenotype value per one unit change in genotype (average effect size).

[10]Linkage is described as the nonrandom association of alleles of different loci; it occurs when genes close to each other are consistently inherited together in a cluster, as often happens in MGE.

5.3.2 Scoary

For the human phenotype, scoary confirmed the *scpB-lmb* transposon as significantly associated with this host. Interestingly, the first entry for *scpB* was less significant than *lmb* (*p*-value 1.23×10^{-33} and 8.34×10^{-71}, respectively, Tab. 5.2). This was due to the fact that multiple variants of *scpB* exist (*n*=4, as identified by roary), and they were considered separately by scoary when computing significance. A single main *lmb* variant (*lmb1*) is common among human genomes (92% of all human genomes)[11]; this resulted in lower scores for *scpB* compared to *lmb* (Tab. 5.2). Additionally, several genes that were negatively associated with the human phenotype appeared among the highest-scoring genes (34 out of the first 50 variants). These comprised genes belonging to Lac.2.

Genes belonging to Lac.2, and in particular the most common *lacE* variant out of four alleles (69.3% of all bovine genomes), scored very highly in the bovine analysis (*p*-value 4.98×10^{-101}, Tab. 5.2). Similar to *lacE*, *lacG* also has four variants, however, the most prevalent one was found in fewer genomes compared to the most common *lacE* (44.0% of bovine genomes), and showed poorer epidemiological characteristics (Tab 5.2). In addition, several phage-related genes were reported among the first 50 genes; these genes belonged to three prophage types (GBS2, GBS3, GBS11) (Crestani et al., 2020) (chapter 2) and to type A prophages, which are permanently integrated/incomplete prophages (van der Mee-Marquet et al., 2018). Among the 50 highest-scoring genes, only two were negatively associated with the bovine phenotype.

Among the fifteen best-scoring genes (i.e. lowest *p*-value) for the fish phenotype, seven belonged to locus 3 (*p*-values between 1.82×10^{-53} and 2.10×10^{-51}, Tab. 5.2). Locus 3, which is described as fish-associated (Delannoy et al., 2016), was present in 100% of fish genomes (SE and NPV 100%) and in a minority of non-fish assemblies ($\sim$23%, mainly human CC12 and CC283, bovine CC12 and CC103/314). Most of these high-scoring genes are involved in carbohydrate metabolism (e.g. galactitol transporters, UDP-glucose 4-epimerase). Upon inspection of the fifty best-scoring genes, as was done for other host

[11] A different *lmb* variant, *lmb2*, is found uniquely in bovine genomes (22% of bovine genomes) (Fig. 5.2), whilst *lmb1* is found across host groups (human, bovine, fish, others).

Table 5.2: Scoary results are reported for some of the most significant genes of the three phenotypes analysed (human, bovine, fish). The best-scoring genes belonging to three host-associated mobile genetic elements (MGE) are indicated (*scpB-lmb* transposon for human, Lac.2 *lacEG* for bovine, locus 3 AraC for fish), as well as their variants, when applicable (these have been indicated with an underscore sign and a progressive number). For the fish phenotype, a hypothetical gene belonging to locus 5 and the best-scoring IS*Stin5* variant were chosen as examples to represent highly specific fish genes.

Phenotype	Gene	SE	SP	PPV	NPV	*p*-value
Human	*lmb1*	91.7	64.7	71.7	88.8	8.34×10^{-71}
	scpB_1	75.0	65.8	68.2	72.9	1.23×10^{-33}
	scpB_2	15.0	99.1	94.0	54.4	3.60×10^{-16}
	scpB_3	16.7	98.4	90.9	54.7	8.78×10^{-16}
	scpB_4	34.5	77.7	60.2	54.8	9.96×10^{-5}
Bovine	*lacE_1*	69.3	97.2	92.3	86.8	4.98×10^{-101}
	lacE_2	18.1	98.1	82.0	71.2	1.99×10^{-16}
	lacE_3	9.0	98.6	75.8	69.2	2.68×10^{-7}
	lacE_4	2.2	99.7	75.0	67.8	0.017
	lacG_1	44.0	96.7	86.5	78.1	9.94×10^{-49}
	lacG_2	35.7	99.1	95.2	76.1	1.10×10^{-47}
	lacG_3	18.1	98.1	82.0	71.2	1.99×10^{-16}
	lacG_4	2.2	99.7	75.0	67.8	0.017
Fish	AraC (locus3)	100	74.5	34.6	100	1.82×10^{-53}
	hypothetical (locus5)	44.6	100	100	93.0	6.51×10^{-47}
	IS*Stin5* (one variant)	38.6	100	100	92.4	4.11×10^{-40}

species, genes belonging to some of the fish-specific loci described by Delannoy et al., 2016, were recognised (locus 1, 2, 4, 5, 6, 8) up to position fifty-four (ten positively-associated genes in total); these were present in some fish genomes (∼45%) but in none of the non-fish

(100% SP and PPV, Tab. 5.2). Similarly, sixteen genes, mostly annotated as hypothetical, were found to belong to new fish-specific loci. Interestingly, among the few functionally annotated genes, I observed genes for surface and membrane proteins (fibrinogen-binding protein and its upstream intergenic region, a GlsB/YeaQ/YmgE family stress response membrane protein, an intergenic region of the capsular operon between *cpsK* and *cpsL*) and a small gene cluster which comprised a bacteriocin. Exploring further genes that were uniquely associated with fish, I identified forty-one entries annotated as IS3 family transposase IS*Stin5*, an insertion sequence (IS) from *Streptococcus iniae*. These were uniquely detected in CC552 (a lineage including only isolates from cold-blooded species) and they were found in multiple copies within the genome (between 1 and 33, see subsection 5.3.3). In GBS, IS*Stin5* is translated into two smaller proteins rather than one and IS*Stin5* gene variants showed similarity to either the first or the second half of the IS*Stin5* reference sequence (`https://isfinder.biotoul.fr`) (Siguier et al., 2006). Among the first 50 genes in the scoary output, twelve were negatively-associated with fish.

5.3.3 BLAST+

Local blast confirmed host-association for the three main MGE reported by the two GWAS tools (Fig. 5.2). Genomes that encoded *scpB* always carried *lmb1* too, and these genes were more prevalent in human GBS (92%) compared to GBS from other species. In cattle GBS, these two genes were present in 39% of genomes (mostly among CC1 and CC23), while in fish GBS they were found in 14% of genomes, all of which belonged to CC283, a shared fish-human lineage. Additionally, although neither of the GWAS programs identified TcR as highly associated with the human host, blast supports a higher prevalence in human isolates (Fig. 5.2).

Lac.2 was confirmed as highly prevalent among bovine GBS genomes (98%) and was detected in very few human GBS genomes (8%), with no detection in fish. Similarly, *pezAT*, which was present in 45% of cattle GBS genomes, was not found among fish GBS genomes, and it was observed in only 6% of human GBS genomes. In human GBS, Lac.2 and *pezAT* were not independent, as 3% of human GBS genomes were both *pezAT*+ and Lac.2+[12].

[12]If these two were independent the joint probability P(*pezAT* & Lac.2) = P(*pezAT*) x P(Lac.2) = 0.06 x

Among bovine GBS genomes, *pezAT* was particularly associated with lineage CC61/67 (37%) and lineage CC103/314 (92%). *PezAT* was carried by an ICE in dairy cattle, which, when present in human genomes, mostly did not encode for this toxin/antitoxin system; it was found in 23/49 ICE+ human GBS genomes, in 114/127 ICE+ cattle GBS genomes and in 0/19 ICE+ fish GBS genomes. The prevalence of the ICE among hosts was 10% in human, 46% in bovine and 19% in fish GBS genomes (Fig. 5.2); thirteen genomes (*n*=10 bovine, *n*=3 human) were *pezAT*+ but ICE-. It is important to underline that I tried to minimise false negatives selecting only a segment of the ICE for blast searches, as described in the materials

0.08 = 0.0048 (0.48%).

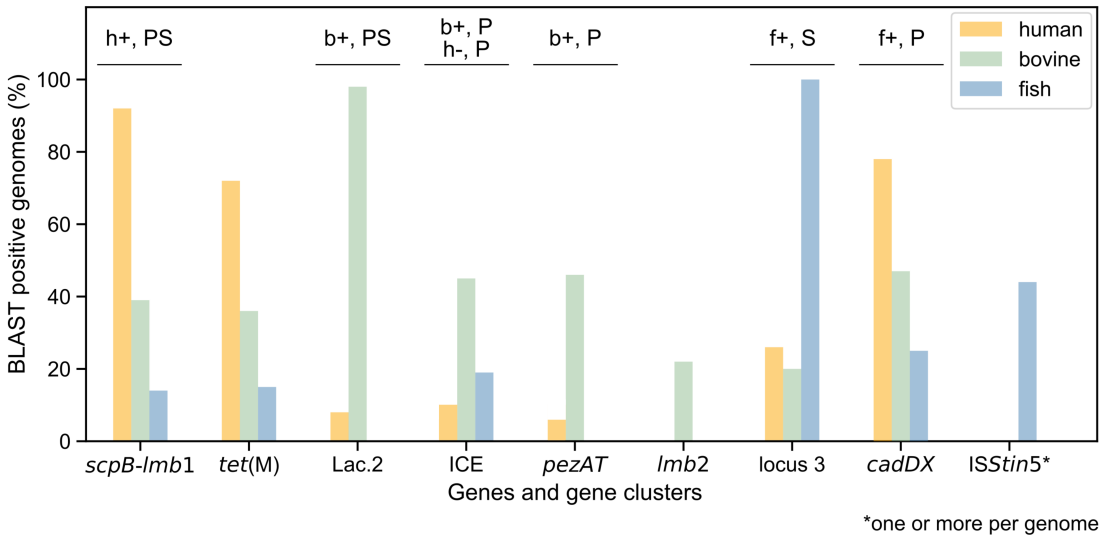


Figure 5.2: Frequency plot showing the distribution as detected by blast of single genes and gene clusters relevant to this chapter in a dataset of 850 group B *Streptococcus* (GBS) genomes among three major host groups. Genes that were detected as significantly associated with a host group are marked with the corresponding letter (h: human, b: bovine, f: fish), and the same has been done for the two programs for genome-wide association studies (GWAS) that identified the association (P: pyseer, S: scoary). Positive and negative associations are shown with the respective signs (+, -). Genes that were not identified as highly significant by either program, but that are uniquely found in one host species are shown (*lmb2*, IS*Stin5*). The *tet*(M) tetracycline resistance gene (TcR) is known to be more prevalent among human isolates; however, *tet*(M) was not reported as significantly associated with the human phenotype by either pyseer or scoary. The prevalences of *cadDX* genes, which pyseer identified as fish-associated, in the various host groups are shown to highlight the poor performance of this program for the fish phenotype.

and methods section. However, due to the high plasticity and modular structure of ICE, false negatives cannot be ruled out.

Locus 3 was found in 100% of fish GBS genomes, and in only 26% of human (in particular within CC283 and CC12) and 20% of bovine GBS sequences (CC7, CC12, CC103/314). IS*Stin5* variants were uniquely found in fish (44% of fish genomes), and they were solely associated with the CC552 lineage. The number of copies of IS*Stin5* (either first or second half of the reference sequence, as described above) varied between 1 and 33 per genome; complete genomes mostly showed a high number of IS*Stin5* (n=35 closed genomes had between 13 and 33 copies, n=1 had only one copy), whereas draft genomes showed fewer hits (n=8 draft genomes with 1-3 copies) (Fig. D.9).

5.4 Discussion and conclusions

5.4.1 Host-associated accessory genome content

The majority of significantly host-associated genes detected in my analyses matched previously known MGE (*scpB-lmb* transposon in humans, Lac.2 in bovines, locus 3 in fishes) that had been found on smaller datasets and with less sophisticated approaches (Delannoy et al., 2016; Richards et al., 2011; Franken et al., 2001). Considering the diversity of genomes included in my collection, which aimed at being representative of the global GBS population, and the different methods used, it is remarkable that no other genes scored as highly as these three MGE, which effectively acted as positive controls. This suggests that a limited number of genes act as major drivers of GBS host-adaptation, together with other forces such as genome reduction and pseudogenisation. The functional relevance of these genes in adaptation to the different host species has been described by other authors. Gleich-Theurer et al., 2009, demonstrated that *scpB*, a transposon-encoded virulence factor that is known for its higher prevalence among human isolates (Morach et al., 2018; Franken et al., 2001), is selectively induced by human serum but not by bovine serum. *ScpB* likely has no functional relevance in dairy cattle and it is therefore considered a human-associated virulence gene, in which it exerts multiple roles: it interacts with the host immune system (cleaving the C5a complement component) and it contributes to cellular adhesion and invasion (Gleich-Theurer

et al., 2009). All of these are mechanisms that, when acquired through HGT, can provide a rapid adaptation to the human host and an advantage over competing strains (S. K. Sheppard et al., 2018). Moreover, the fact that *scpB-lmb* in fish GBS isolates is found uniquely in a lineage that is shared with humans (CC283) could explain why this is the only lineage from fish that shows evidence of zoonotic transmission. In addition, the *scpB-lmb* transposon is shared with other human pathogenic streptococci, such as group A *Streptococcus* (GAS), *Streptococcus dysgalactiae* subsp. *equisimilis* and *Streptococcus canis*, and it has been associated with their ability to colonise or infect the human host (Franken et al., 2001). Richards et al., 2013, showed how Lac.2, which is carried by an MGE (Richards et al., 2011), is responsible for the metabolism of lactose; lactose-fermenting genes are thought to provide a fitness advantage to isolates living within the bovine mammary gland, which is rich in lactose. In fish, Delannoy et al., 2016, identified genes in locus 3 (a gene cluster located in a pathogenicity island, or PAI) which encoded products that are involved in galactose transport and metabolism, such as the alpha-galactosidase (*galA*) and genes for all enzymes of the Leloir pathway (*galK*, *galT*, *galE*). Galactose is present in high concentration in fish tissues like the brain (Tocher, 2003), and this may explain the tropism of GBS for the central nervous system in fishes. Lac.2 and locus 3 are perfect examples of how HGT represents a means for rapid host-adaptation and development of tissue-tropism, thanks to the acquisition of the ability to utilise available substrates in a specific niche (S. K. Sheppard et al., 2018).

Few genes other than the MGE described above were reported as significantly associated with any of the host species, and they were all associated with MGE (i.e. ICE, prophages, IS), supporting the importance of the mobilome in host-association of GBS. Among these, a variant of the *pezAT* toxin/antitoxin system was reported as highly associated with cattle by pyseer, and confirmed as primarily found in cattle by blast (Fig. 5.2). Interestingly, *pezAT* was particularly prevalent among lineage CC61/67 (37%) and lineage CC103/314 (92%), which are uniquely and primarily associated with cattle, respectively. *PezAT* was carried by an ICE that had been reported as significantly positively associated with cattle by pyseer, but not by scoary, and significantly negatively associated with human genomes, in which it carried *pezAT* only in half of the isolates (23/49 ICE+ human genomes, and 114/127 ICE+ cattle genomes). It must be noted that, since *pezAT* genes have a short sequence and their

translated protein sequence was searched on nucleotide genome assemblies, it is likely that the number of false negatives has been minimised. For the ICE, the results probably include a number of false negatives (as suggested by *pezAT*+ ICE- genomes), as the blast search for this element was based on a longer nucleotide sequence, and ICE are known to be modular and highly plastic. The ICE shows sequence homology with segments of two other streptococcal ICE: ICE*Sde*3396 (Davies et al., 2009) from *S. dysgalactiae* subsp. *equisimilis*, which does not encode for *pezAT*, and ICE*Slu*van from *Streptococcus lutetiensis* (previously *Streptococcus bovis* biotype II) (Bjørkeng et al., 2013) (Fig. D.7). Mosaic structures were observed for both these ICE, which showed similarities with sequences from different bacterial species (streptococci, *Faecalibacterium*, enterococci, and other gram-positives). Similar to ICE*Slu*van, which not only carries *pezAT*, but also a Tn*1549* element that encodes the vancomycin resistance gene cluster *vanB*, my bovine ICE carried Tn*916* (which did not show significant association with the bovine phenotype, as the rest of the ICE), which carries the *tet*(M) resistance gene, in the reference genome shown in Fig. D.7. This highlights how recombination events are key factors in the evolution of ICE in streptococci and the degree of variability that can exist even within the same bacterial species. The GBS *pezA* nucleotide sequence shows 100% ID and QC with that of *Streptococcus suis*, in which it is carried by different ICE (e.g. ICE*Ssu*YS388, ICE*Ssu*YS108, ICE*Ssu*YS34), and of *Enterococcus faecalis*, and over 93% ID with that of *S. canis*, *Enterococcus faecium*, *Streptococcus porcinus* and *Streptococcus uberis*. The gene *pezT* has analogous sequence similarities (100% ID and QC with *S. suis*, 96.74% ID with *E. faecalis*). The *pezAT* gene cassette has been described in *Streptococcus pneumoniae* (Khoo et al., 2007), *S. suis* (Holden et al., 2009), and GBS (Holden et al., 2009; Khoo et al., 2007), and consists of two genes, encoding an epsilon antitoxin and a zeta toxin, respectively. Traditionally, toxin/antitoxin addiction systems are found on plasmids and help ensure their maintenance in the bacterial population (Gerdes et al., 2005). Recently, they have been found in other MGE and have been shown to aid fixation of ICE in the chromosome, for example in *Vibrio cholerae* (Wozniak & Waldor, 2009). These systems also play a role in the stress response, helping bacteria to survive in a hostile environment (Christensen et al., 2003); *pezAT* specifically has been shown to reduce resistance to β-lactam antibiotics, currently the drug of first choice for GBS infections in humans and the major antibiotic family used for treatment of mastitis in cattle, and to reduce

genetic competence[13] in *S. pneumoniae* (Chan & Espinosa, 2016). Unfortunately, GWAS results alone are not able to provide information on the levels of expression of *pezAT* in GBS from dairy cattle (which would require wet-lab experiments with RT-PCR), on its role in promoting GBS growth within the mammary gland, or on its potential role in β-lactam activity against GBS (although resistance to β-lactams in GBS is mainly associated with structural changes in the *pbp* gene (Hayes et al., 2020; C. Li et al., 2020; van der Linden et al., 2020)). GWAS and blast can only give an indication of which genes could play a role in host-adaptation, due to a higher prevalence of a gene in one species compared to other species, but do not give information on their functional role. I believe it would be interesting to further investigate the role of *pezAT* in GBS from dairy cattle with wet-lab experiments.

Scoary identified as significantly associated with fish seven of the eight fish-specific (locus 1, 2, 4, 5, 6, 8) and fish-associated (locus 3) loci described by Delannoy et al., 2016. In particular, genes belonging to locus 3 scored the highest. Locus 3 is thought to exert the greatest role in GBS fish-adaptation, as indicated by experimental challenge studies in tilapia performed with locus 3 knock-out mutants generated at the Moredun Research Institute (Penicuik, Scotland) by Dr John Bell, after which GBS was attenuated in fish (Ruth Zadoks, personal communication). Fish-specific loci had been defined by Delannoy et al., 2016, as gene clusters associated uniquely with the CC552 lineage, which comprises only GBS isolates from cold-blooded species, whereas the fish-associated loci were described as being present in CC7, CC283 and CC552. Blast showed that locus 3 is highly prevalent among isolates of the human-fish shared lineage, CC283, and of CC12; it can also be observed, with a variable prevalence, in other lineages (CC1, CC23, CC130, CC103/314) (Fig. 5.3). In addition to these seven loci, other genes were identified as fish-associated by scoary. Unfortunately, apart for a few surface and membrane proteins, many of these genes were annotated as hypothetical genes, and it is not possible to fully understand their role in fish-adaptation. Another striking feature identified by scoary, uniquely in the CC552 lineage, was the presence of several copies of the insertion sequence IS*Stin5*, which, after further investigations, appeared to be consistently partitioned into two smaller ORFs (open reading

[13]Genetic competence is the ability of a cell to uptake naked extracellular DNA from its environment through transformation.

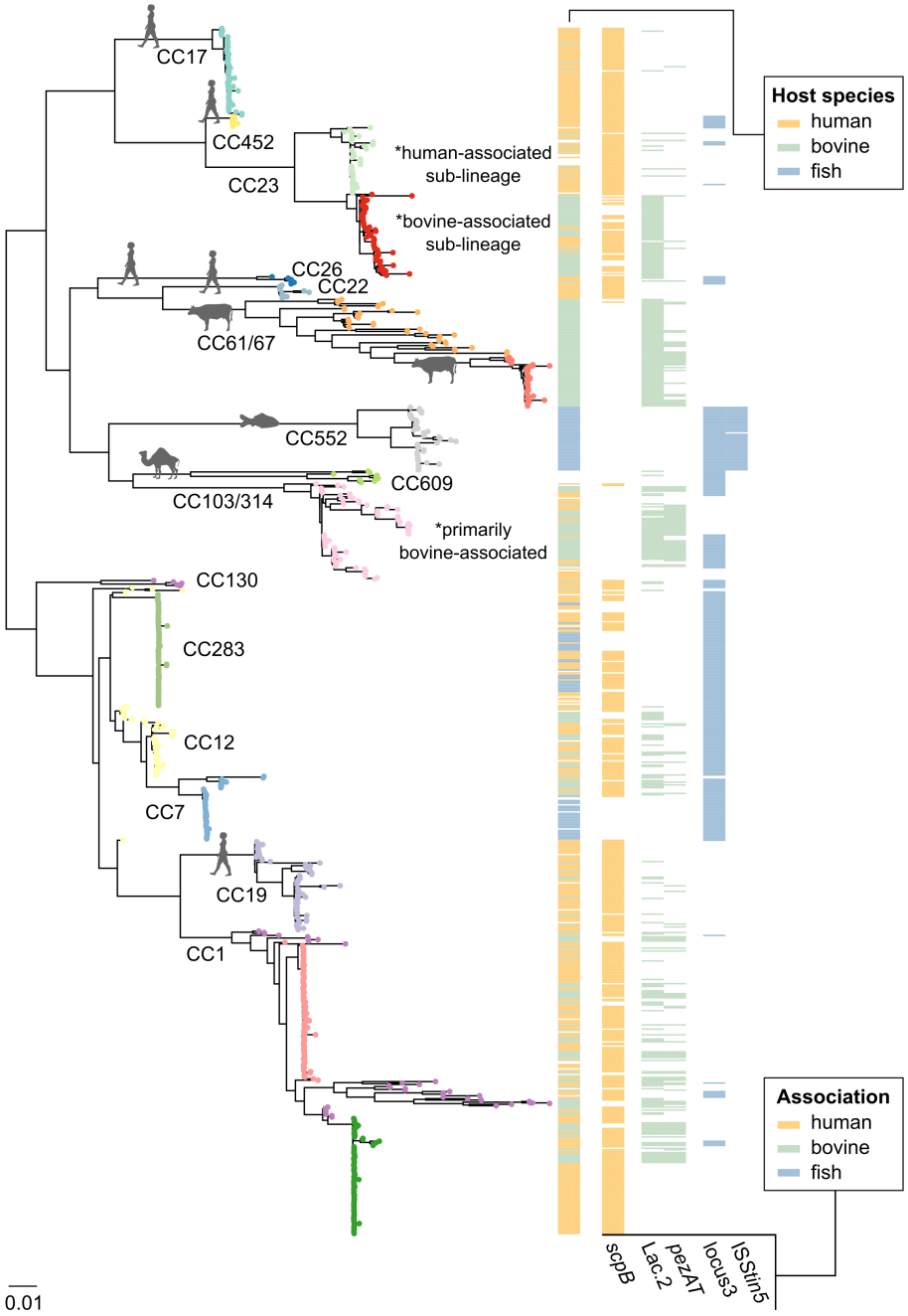


Figure 5.3: Maximum-likelihood phylogenetic tree of 850 group B *Streptococcus* (GBS) genomes reconstructed with IQtree v1.6.10 (as described in chapter 4). Leaves colours show the 18 BAPS clusters identified in chapter 4. Outer strips show: species of isolation for the three major host groups (human, bovine and fish, marked in orange, green and blue, respectively), and the distribution across the population of mobile genetic elements that are associated with each group (indicated with the same colour as their associated host), as per genome-wide association studies (GWAS) results. Host-specialist lineages have been indicated with host icons on their branches. Tree was rooted at midpoint.

frames) compared to the reference sequence (Fig. 5.4). Previously, Rosinski-Chupin et al., 2013, reported having identified 11-20 copies of IS*Sag1* in CC552 isolates. However, I was not able to confirm this: no reference sequence for IS*Sag1* is available on ISfinder or NCBI for blast identification. When I searched for the primer sequences reported by Rosinski-Chupin et al., 2013, on SnapGene v5.2.5 (`https://www.snapgene.com`), they did not show any binding sites in my C552 genomes; this was true also for the first complete ST261 genome Rosinski-Chupin et al., 2013, generated (2-22, FO393392), and for which they reported 11 IS*Sag1* copies. IS play a role in genome rearrangements, with their ability to carry other genes and transposing flanking DNA sequences, and in gene expression, being able to silence or activate genes, based on whether they insert within a gene or upstream a gene, respectively (Siguier et al., 2014; Curcio & Derbyshire, 2003). The insertion of IS within genes causes gene disruption and can lead to pseudogenisation and long-term genome reduction. This mechanism has been described in several bacterial species, such as *Mycobacterium ulcerans*, in which it was responsible for its evolution from the generalist environmental *Mycobacterium marinum* into a niche-adapted specialist (Rondini et al., 2007), and *Yersinia pestis*, a bacterial species that evolved through IS-mediated genome reduction from the less virulent *Yersinia pseudotuberculosis* (Chain et al., 2004). In *Escherichia coli* it has been shown how IS-mediated mutations can both promote and limit evolvability, however, a higher IS activity appears detrimental to adaptation over evolutionary time (Consuegra et al., 2021). The high prevalence of IS*Stin5* in fish CC552 genomes suggests that this element played a major role in the genome reduction observed in this lineage (Rosinski-Chupin et al., 2013) and in the inability of these isolates to escape the fish host and adapt to new niches. In addition, IS*Stin5* is an insertion sequence first described in *S. iniae*, another leading pathogen of the aquaculture industry, particularly in warmwater fish, with a host-spectrum that is very similar to that of GBS (Agnew & Barnes, 2007). It is likely that, similar to what is described in chapter 3 for plasmid exchange among GBS, GAS and *S. dysgalactiae* subsp. *equisimilis* which share the human oropharynx niche (Davies et al., 2005), genetic exchange of a fish-associated accessory gene pool occurs among bacteria that share the aquatic niche with GBS, such as *S. iniae*. This was confirmed by blasting amino acid sequences[14] of genes be-

[14]Thresholds for query coverage and of percentage of identity (ID) were set at 90% and 40%, respectively; the ID% of 40% is a standard threshold used to identify homologous sequences.

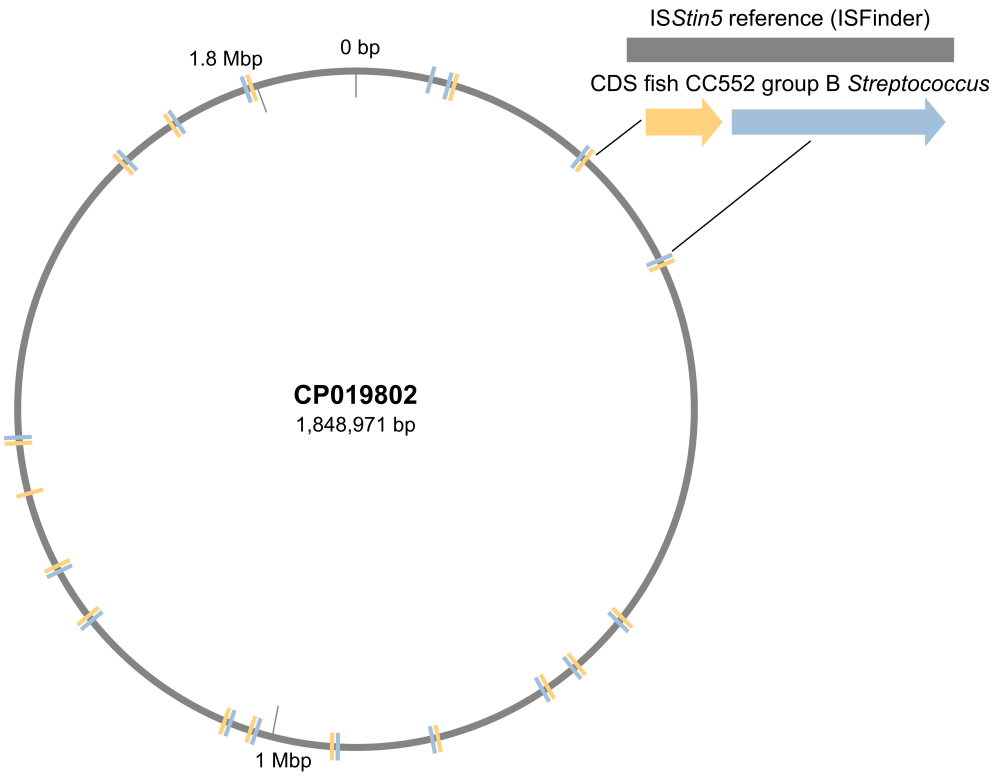


Figure 5.4: Example map of IS*Stin5* elements in a complete group B *Streptococcus* (GBS) genome from fish belonging to the CC552 lineage. Coding sequences (CDS) have been indicated with two colours, based on their identity to the reference IS*Stin5* sequence (first half orange, second half blue) obtained from ISfinder (Siguier et al., 2006), and mostly occur in pairs.

longing to loci 1-8 against all *S. iniae* genomes available to date (n=94; March 2021). Two loci gave positive hits: locus 3 and 4. While locus 4 genes were found in 2-4 genomes, eight genes from locus 3 were found in 100% of *S. iniae* genomes, in which they were divided in two islands: genes 2-4 (three genes for the Leloir pathway) and genes 5-9 (*galA*, sugar transporters and AraC transcriptional regulator). This highlights the importance of locus 3 genes in streptococcal adaptation to fish.

5.4.2 GWAS methodologies

Results from pyseer and scoary were in agreement with respect to the positive associations of the *scpB-lmb* transposon with the human phenotype and of Lac.2 (in particular its *lacEG* genes) with the bovine phenotype, albeit with differences in numerical estimates of signif-

icance, especially for genes with multiple alleles. In addition, when considering the blast prevalences of these genes/MGE (Fig. 5.2), it is clear how the difference in prevalence across species impacts on epidemiological characteristics such as SE, SP, PPV and NPV. As an example, genes from locus 3 have a perfect SE and NPV (100%), as they are found in all genomes from fish, but have poorer SP and PPV (Tab. 5.2), as they are also found in a good proportion of non-fish genomes (Fig. 5.2). This means that the detection of genes from locus 3 is not a very good predictor of a genome belonging to the fish phenotype (e.g. AraC PPV 34.6%), whilst the absence of locus 3 perfectly predicts negativity for the fish phenotype (100% NPV). Conversely, genes found uniquely in fish genomes and in none of the non-fish (e.g. locus 5 and IS*Stin5*, Tab. 5.2) have SP and PPV of 100%, therefore they are perfect predictors of a positivity to the fish phenotype. While they also conserve a good NPV (>90%), they have low SE (38.6-44.6%), as they are found in less than half of the fish genomes.

Pyseer reported an ICE as significant in both the human and bovine phenotypes; association was negative with the former and positive with the latter, in which the ICE carried a cattle-associated toxin/antitoxin system (*pezAT*). However, scoary did not confirm the high significance of ICE-encoded genes with either of these phenotypes (e.g. ICE-encoded LPxTG gene *p*-values for bovine: pyseer-unitigs 2.08×10^{-140}, scoary between 5.84×10^{-18} and 0.034, depending on the variant; for human: pyseer-unitigs 1.89×10^{-25}, scoary between 3.62×10^{-9} and 0.011, depending on the variant). The absence of association in the scoary output could be a result of multiple alleles of genes carried by the ICE (i.e. failure to detect association), whilst the positive association with the ICE with the bovine phenotype in pyseer could be a results of this MGE being in linkage with *pezAT* (possible false positive association). Unlike pyseer, scoary identified known fish-associated and fish-specific genes, as well as fish-specific genes that had not been identified before.

The main advantage of pyseer is that this program is not affected by the presence of gene variants. This is because *k*-mers/unitigs are mapped to the same gene, which gets an overall higher score than the single variants identified by scoary would. An example of this is *scpB*, of which four different variants are known. Pyseer found *scpB* to be more strongly associated with the human phenotype than *lmb*; this likely has no biological significance, as

blast showed that *scpB*+ genomes are also *lmb*+ (hence they have equal distributions), rather, the higher *scpB* score was due to the larger size of this gene compared to *lmb* (larger genes generate more *k*-mer/unitigs hits than smaller genes in pyseer, as it is the case also for *lacEG* in Lac.2). In contrast, scoary gave a lower score to *scpB* compared to *lmb* because it was influenced by the existence of these different alleles. Another advantage of pyseer is that it can identify associations to the SNP level, which can be useful when the phenotype being tested for is determined by SNP, such as mutations responsible for antimicrobial resistance, as shown for *Mycobacterium tuberculosis* (Desjardins et al., 2016). SNP can potentially influence host range, as observed for *S. aureus* in rabbits (Viana et al., 2015). However, given that the focus of my thesis is the mobilome, I decided to focus more on gene presence/absence, rather than on specific SNP within those genes. Finally, unitig-based analyses, which are to be preferred as described in the 'best practices' section of the pyseer documentation (`https://pyseer.readthedocs.io`), offer an advantage over *k*-mers, as they are computationally more efficient and require less running time.

A major limitation I experienced with pyseer was that the program seems to be strongly affected by population structure and possibly by the sample size for single phenotypes. In fact, even when correcting for population structure, my fish analysis did not yield significant results, as it failed to detect known fish-associated and fish-specific loci, and it deemed the *scpB* transposon to be significantly positively fish-associated. This could lead to unsatisfactory or misleading results also in the case of the investigation of genes associated with a single lineage (e.g. camel-specific lineage CC609, as detailed in chapter 6), which might be desired when e.g. a certain lineage within a bacterial species adapts to a new host species, or shows higher pathogenicity compared to other lineages. Another important limitation is that the final output of pyseer does not intuitively state whether the association of a certain gene with the phenotype of interest is a positive or negative one. To determine this, intermediate files need to be scanned for the presence of a positive or negative sign of the beta-coefficient associated with each *k*-mer/unitig mapped to that gene, as the final pyseer output only takes in absolute values. Lack of intuitiveness of the final output can also be caused by the reference genomes used for the final annotation step in which the significant *k*-mers or unitigs are mapped to gff annotation files, to determine to which gene they belong: if a certain gene

is annotated with different IDs in the different gff reference files, which is often the case, the same gene will be reported multiple times as significant in the final output, under its different IDs. Also, the annotation can be affected by the set of reference genomes chosen: if reference genomes from other phenotypes are included when annotating, genes that are actually not associated with the phenotype of interest might appear in the output. As an example, when I annotated the human significant unitigs with a set of reference genomes that also included bovine genomes (all reference genomes in Tab. D.1), Lac.2 appeared in the final output and plot (Fig. D.10). This is because Lac.2 has a significant *p*-value in human GBS, but it is negatively associated with it. However, as described above, it is not immediately possible to distinguish negatively and positively associated genes, and this makes the pyseer output difficult to interpret. The same issue appeared when I annotated the significant *k*-mers/unitigs of each phenotype with a gff reference file generated from the pangenome (an example for the human unitigs annotation can be found in Appendix D, Fig. D.11). In addition, pyseer is affected by linkage, the genetic phenomenon that causes genes that are close to each other to be consistently inherited together, as often is the case for MGE: if only one or a few genes within a MGE play a functional role in adaptation to a certain host, pyseer is likely to attribute high scores not only to these genes, but also to other genes belonging to that MGE, regardless of their functional importance. An example of this is Lac.2, which is responsible for lactose fermentation and which is therefore considered to help with niche-adaptation to the bovine mammary gland. Pyseer reported all genes in Lac.2 as highly significant in bovine (with a higher significance for *lacEG*), but also genes in the fructose operon, which is in linkage with Lac.2 but is not considered to play a major role in adaptation to the bovine host (Richards et al., 2013, 2011).

Scoary demonstrated several advantages over pyseer, the first being an extremely quick running time (for all three phenotypes: 12 min 35 sec, vs pyseer *k*-mers: human 22 hr 43 min 54 sec, bovine 22 hr 49 min 59 sec, fish 38 hr 33 min 56 sec; pyseer unitigs: human 20 min 58 sec, bovine 15 min 8 sec, fish 14 min 39 sec). It was also less affected by population structure, as it was able to recognise positive controls (genes or loci that were known to be host associated) not only in the human and bovine phenotypes, but in the fish phenotype as well, which pyseer was unable to do even when correcting for population structure (with the

minimum MAF cut-off at 4%). In addition, the scoary output is much more intuitive for several reasons. First, as scoary works on a presence/absence file, there is more consistency on gene annotation and their relative scores. Also, scoary helps in distinguishing whether the association of a certain gene with a phenotype is positive or negative very quickly, as it reports the number of genomes belonging to the phenotype of interest in which the gene is present and absent (Number_pos_present_in, Number_pos_not_present_in, respectively) and the number of genomes from the opposing phenotype in which the gene is present and absent (Number_neg_present_in, Number_neg_not_present_in, respectively).

The major limitation of scoary is that this program can be negatively affected by gene variants, which was the case for the lower score observed for *scpB* compared to *lmb*, as described above. This was also observed for the different variants of *lacG*, of which the most common variant was present in 122/277 bovine genomes, compared to *lacE*, for which the most common variant was present in 192/277 genomes: the *lacE* variant scored better than the *lacG*, which had a lower score even compared to genes in the fructose operon. In addition, genes that are uniquely present in a phenotype (or more prevalent in a phenotype compared to others) but not with proportions as high as e.g. that of Lac.2 in bovine, may be overlooked in the scoary output, as could have been the case for *pezAT* and IS*Stin5* if I had not filtered the output for variants that were uniquely present in bovine and fish, respectively.

Overall, I found the scoary output easier to interpret and more reliable, especially in light of previous knowledge on host-associated MGE in GBS. On the other hand, comparing the output from scoary with that from pyseer was useful to identify issues with gene variants in scoary (e.g. *scpB* in human, and *lacG* in bovine).

5.4.3 Final remarks

I showed that there is a limited number of genes/MGE that are significantly associated with three major host groups in GBS. These MGE carry genes that are mostly related to the metabolism and transport of carbohydrates, which confer an adaptive advantage in the presence of available substrates (i.e. lactose in dairy cattle and galactose in fishes), or related to inactivation of the immune response and increased adherence/invasiveness (i.e. *scpB* in

humans). I also found a few genes that are associated with certain phenotypes which did not show levels of significance as high as the aforementioned MGE, but which could nonetheless have played a role in host-adaptation (e.g. *pezAT* in cattle and IS*Stin5* in fish). In addition, I showed the existence of a fish-associated accessory gene pool related to the aquatic niche, which is shared between streptococcal fish pathogens (GBS and *S. iniae*). Overall, I preferred the performance of scoary over pyseer, although I greatly appreciated the advantages deriving from the comparison of multiple GWAS programs.

Chapter 6

Characterisation and identification of niche-associated genes of group B *Streptococcus* from camels

6.1 Introduction

Group B *Streptococcus* (GBS) has a wide range of host species, the major ones being humans, cattle and fish. A fourth host group, which has gained more attention within the scientific community in recent years (Seligsohn et al., 2020; Fischer et al., 2013; Zubair et al., 2013), is that of camels (*Camelus dromedarius*). GBS has been isolated from healthy (carriage in the nasopharynx, vagina and rectum) and diseased camels in the Horn of Africa (Younan & Bornstein, 2007; Bekele & Molla, 2001; Obied et al., 1996). Clinical syndromes can vary (e.g. chronic cough, gingivitis, wound infections, periarthricular abscesses), with the most important one being subclinical mastitis (Seligsohn et al., 2020; Fischer et al., 2013; Zubair et al., 2013). Camel milk is vital for pastoralist communities in East Africa both as a source of income and as part of their diet (Elhadi et al., 2015). There is a growing reliance on camels as source of meat and milk in this geographical area, which is in part attributable to climate change and camels' ability to tolerate droughts and more extreme environments compared to other species. Mastitis reduces camel milk production in terms of quality and quantity (Saleh et al., 2013), similar to dairy cattle, with negative impacts on food security and, possibly, food safety (Seligsohn et al., 2020). GBS isolates from dairy cattle mastitis

carry accessory genetic content that promotes their success in the bovine mammary gland (Lac.2, as described in chapter 3), and many lineages affecting dairy cattle are shared with the human host (chapter 4). Not only is there evidence for the potential for bovine-to-human transmission of GBS, as testified by several studies (Cobo-Ángel et al., 2019; Sørensen et al., 2019), but human-to-bovine transmission of host-generalist lineages, such as clonal complex (CC) 1, appears possible (as described in chapter 3). Colonisation/infection of a new host species provides opportunities for acquisition of host-associated accessory genome content in GBS isolates that can result in the amplification of such strains in the new host (chapter 4 and chapter 5) (Richards et al., 2019), and potentially in more virulent genotypes that pose a higher threat to human health, as shown for sequence type (ST) 283 (Barkham et al., 2019). This highlights the importance of expanding knowledge on GBS ecology in camels, particularly for mastitis-causing strains.

A limited number of genomic studies published to date focus on GBS from camels and only nine genome sequences had been published prior to 2021 (Fischer et al., 2013; Zubair et al., 2013), which is why camel GBS genomes are poorly represented in chapters 4 and 5. One study suggested the presence of a camel-specific linage (comprising ST616 and ST617 among others) and of a shared human-camel population (comprising ST609 and ST614 from camels, and ST26 from humans) (Fischer et al., 2013). However, this was solely based on multi-locus sequence typing (MLST) data, not on whole genome sequencing; therefore, inferences on relationships between isolates of camel and human origin should be made carefully due to the limitations of this typing system, as explained in chapter 4. In addition to core genes, some accessory genes were analysed, but they were limited to antimicrobial resistance (AMR) genes in isolates phenotypically resistant to tetracycline; in these isolates, the *tet*(M) gene was detected and it was carried by a Tn*916* integrative conjugative element (ICE) (Fischer et al., 2013). Another study that analysed a large GBS collection, which included one camel genome, reported this isolate as distinct from all other isolates in its gene content and as biochemically enriched, particularly for carbohydrate metabolism (Richards et al., 2019). This suggests that unique genetic features might be associated with the camel GBS population, although it is not sufficient to draw conclusions that apply to all GBS strains affecting camels. Specific camel-associated accessory genome content has not been exten-

sively investigated in GBS to date, and a knowledge gap exists on the structure of the GBS population in camels, and on the possible existence of shared human-camel ST and lineages.

As part of a collaboration with Dr Dinah Seligsohn and Dr Erika Chenais (National Veterinary Institute Sweden - SVA), I gained access to a large collection of GBS genomes from Kenyan camels (n=122). Two genetic studies were produced as part of this collaboration: the first was aimed at investigating the diversity of the GBS population both within and between herds in pastoralist communities (Seligsohn et al., 2021a), whilst the second explored the prevalence of extramammary GBS carriage in ranch settings in Kenya (Seligsohn et al., 2021b).

The aim of this work was to gain insight into the genetic mechanisms for host-adaptation of GBS in camels. To this end, the population structure of the largest collection of GBS genomes from camels available to date was analysed, both in terms of core and accessory genome content. Methods from previous chapters were combined: i) core genome analysis based on maximum-likelihood phylogeny and clustering algorithms (chapter 4); ii) detection of mobile genetic elements (MGE) such as prophages, phage-inducible chromosomal islands (PICI) (chapter 2) and ICE (Tn*916*) (chapter 3); iii) detection of camel-associated and camel-specific accessory genome content with genome-wide association studies (GWAS) (chapter 5). Because GBS from mastitis is of particular interest, genes associated with milk isolates compared to those from other sample types were investigated with the same pan-GWAS approach.

6.2 Materials and methods

All supplementary material for this chapter, including tables and figures, can be found in Appendix E (these are indicated with the letter E in front of the sequential number).

6.2.1 Dataset selection

A dataset of 122 GBS genomes from Kenyan camels was used for this work (a complete list can be found in Tab. E.1). Sample collection was performed by Dr Seligsohn and took

place during 2017 (February-April) and 2019 (November) in three Kenyan counties: Isiolo, Meru and Laikipia (Fig. 6.1), all classified as arid or semi-arid regions. Within the Isiolo and Meru counties, herds owned by pastoralist communities were sampled, while in Laikipia the majority of the herds belonged to ranches or smallholders. Isolates derived from twenty-five herds, different age groups (adults, *n*=98; calves, *n*=24) and sample types (milk[1], *n*=75; mouth, *n*=7; nose, *n*=37; rectum, *n*=3). Details of sampling and isolation procedures can be found in published work (Seligsohn et al., 2021a, 2021b). DNA extraction from pure colonies of phenotypically confirmed GBS isolates was performed at SVA by Dr Seligsohn using the IndiMag Pathogen kit (Indical Bioscience GmbH, Leipzig, Germany).

6.2.2 Sequencing and assembly

Sequence data were generated at Clinical Genomics, Science for Life Laboratory (Clinical Genomics, Solna, Sweden) with Illumina NovaSeq 6000, resulting in paired-end reads with an average length of 150 bp and average depth of 450x. Raw read processing and downstream analyses were performed in Glasgow by me.

Genomes were assembled *de novo*, using the pipeline shovill v1.0.9 (`https://github.com/tseemann/shovill`), which uses SPAdes v3.13.1 (Bankevich et al., 2012) to assemble the reads. As built-in pre-processing steps, shovill uses seqtk (`https://github.com/lh3/seqtk`) to reduce read depth to 100x (this step optimises the speed of assembly) and Trimmomatic (Bolger et al., 2014) to trim low quality reads. Assembly quality was checked with QUAST v5.0.2 (Gurevich et al., 2013) (Fig. E.1), and genome quality statistics were processed with the same quality control pipeline described in chapter 4. Species identity was confirmed with KmerFinder v3.2 (Clausen et al., 2018; Larsen et al., 2014; Hasman et al., 2014). Multi locus sequence typing (MLST) was carried out with SRST2 v0.2.0 (Inouye et al., 2014) and capsular serotype was detected *in silico* using a standard method also used in chapter 3 and 4 (Metcalf et al., 2017).

[1] All milk isolates included in this work derived from cases of mastitis, which had been diagnosed by Dr Seligsohn based on the presence of clinical signs (e.g. swelling, increased temperature, pain and redness) and/or positivity to the California mastitis test (CMT); CMT gives an indication of the degree of inflammation based on a semi-quantitative measurement of milk leukocytes. Dr Seligsohn considered a CMT ≥ 3 as positive for mastitis.

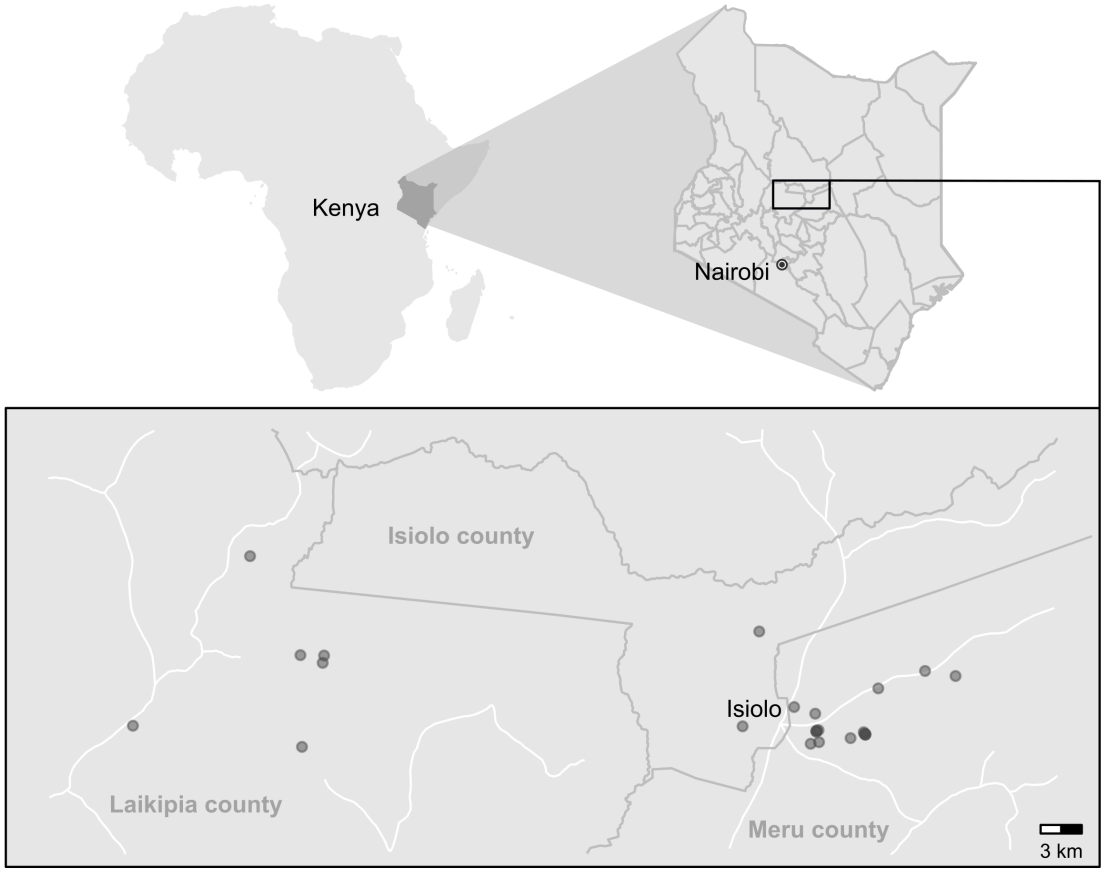


Figure 6.1: Geographical map showing sampling sites (*n*=25 herds) for 122 group B *Streptococcus* (GBS) isolates from Kenyan camels (lactating females and their calves), collected by Dr Dinah Seligsohn (Seligsohn et al., 2021a, 2021b). Sampling took place in three Kenyan counties: Isiolo, Meru and Laikipia. Regional borders are shown as dark grey lines, main roads are shown in white, sampling sites are shown as semi-transparent circles. Map was created with ggplot2 v 3.3.5 in RStudio v1.3.1093 (Allaire, 2012), R v4.0.3 (R Core Team, 2013), with shapefiles from ArcGIS (`www.arcgis.com`).

For genome-wide association studies (GWAS), this dataset of camel genomes was combined with the one described in chapter 4, section 4.2 (*n*=850 genomes), for a total of 972 GBS sequences (see below).

6.2.3 Core genome analysis

A core genome alignment was obtained with snippy v4.6.0 (`https://github.com/tseemann/snippy`) using ILRI112 (HF952106.1) as reference sequence, a serotype VI ST617 from a periarthricular abscess of a Kenyan camel (Zubair et al., 2013). A phyloge-

netic tree was reconstructed with RAxML-NG v0.9.0 (Kozlov et al., 2019) with a general time-reversible (GTR)+G model and the tree was visualised in iTol (Letunic & Bork, 2006). Fastbaps v1.0.4 (fast hierarchical Bayesian analysis of population structure) (Tonkin-Hill et al., 2019) was run on the core genome alignment generated with snippy within RStudio v1.3.1093 (Allaire, 2012), R v4.0.3 (R Core Team, 2013) (for all the commands used, see Appendix E, subsection E.2).

6.2.4 Analysis of accessory genome content

Host-associated genes previously detected with GWAS (chapter 5) were searched for with blast v2.9.0 (Camacho et al., 2009). The presence of the lactose operon (Lac.2) (Richards et al., 2011) was assessed with blastn based on a database of nucleotide sequences of four known Lac.2 genotypic variants (Crestani et al., 2021; Sørensen et al., 2019). Lac.2-negative isolates were further scanned for annotations related to Lac.2 in files obtained using Prokka v1.14.5 (Seemann, 2014) to confirm presence/absence (which was also validated by Dr Seligsohn with a PCR targeting a ∼2.5-kbp region straddling *lacEG* (Seligsohn et al., 2021a, 2021b)). Amino acid sequences of single genes forming the *scpB* transposon (n=6 *scpB* alleles, n=2 *lmb1* alleles) and locus 3 (n=17 genes in total, five of which had two alleles) were searched for with tblastn. Minimum thresholds for detection were set at 90% sequence identity (ID) and 90% query coverage (QC).

SRST2 was used to detect antimicrobial resistance (AMR) genes from raw sequence reads with the ARG-ANNOT v3 database (Gupta et al., 2014). Similar to Lac.2, the presence of Tn*916* was investigated with blastn (ID >90%, QC >80%). Annotation files of assemblies that were *tet*(M)-positive but blastn-negative for Tn*916* (n=4) were further explored.

Presence of prophages and PICI was assessed with PHASTER (Arndt et al., 2016), and complete prophages and PICI were classified based on integrase type, as described in Crestani et al., 2020.

6.2.5 Genome-wide association studies (GWAS)

Gene presence/absence matrices were generated with roary v3.13.0 (A. J. Page et al., 2015) from annotation files created with Prokka. Two pan-GWAS were performed with scoary v1.6.16 (Brynildsrud et al., 2016): a within-host comparison between isolates from camel milk samples and from other sample types, and a between-host comparison of camel against non-camel GBS genomes. I selected scoary over pyseer as this program showed better performance when there is a strong lineage effect (chapter 5, section 5.4.2), which is the case for camel GBS isolates, as they mainly belong to one camel-specific lineage (CC609, as described in chapter 4). For the within-host sample type comparison, the roary gene presence/absence matrix used as input file for scoary only comprised *n*=122 camel GBS genome assemblies generated from the SVA isolates (*n*=75 milk vs *n*=47 non-milk). For the between-host comparison, the matrix was built on GBS genomes included in this chapter (*n*=122) and genomes described in chapter 4, section 4.2 (*n*=850), for a total of *n*=972 (*n*=131 camel vs *n*=841 non-camel). Statistics such as sensitivity (SE), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV) were calculated as described in chapter 5.

Figures were created with iTol (core genome phylogeny) (Letunic & Bork, 2006), Easyfig v 2.2.2 (maps of genetic elements for comparisons) (Sullivan et al., 2011), RStudio v1.3.1093 (geographical map of sampling sites) (Allaire, 2012), and/or modified with Inkscape (`www.inkscape.org`).

6.3 Results

6.3.1 Isolate genotyping and core genome analysis

Isolates belonged to eight ST, of which the most well-represented was ST616 (*n*=59) (Tab. E.1). Other known ST were ST1, ST612, ST615, ST617 whilst three ST were new, and they were submitted to PubMLST (`https://pubmlst.org`) for ST assignment (ST1652, ST1653, ST1654). Five serotypes were identified, the most common being serotype III (*n*=61) and serotype VI (*n*=36), followed by serotype IV (*n*=14), II (*n*=10) and V (*n*=1). ST616 serotype III isolates represented the majority of milk isolates (77%). Each ST was as-

sociated with only one serotype (ST612 serotype IV; ST615 serotype II, ST616 and its single locus variants ST1653 and ST1654 serotype III; ST1652 serotype VI), with the exception of ST617 (*n*=11/17 isolates serotype VI, *n*=6/17 isolates serotype IV). One isolate belonged to the host-generalist lineage ST1 (serotype V) (Fig. 6.2).

Fastbaps identified six subpopulations[2] (which are all part of population 15, or CC609, as described in chapter 4). These largely corresponded to ST (population 1, ST615; population 2, ST612; population 3, ST616 and its single locus variants, SLV, ST1653 and ST1654; population 6, ST1). ST617 isolates belonged to two distinct subpopulations, population 4, which only comprised ST617 serotype IV, and population 5, which comprised ST617 and its SLV ST1652, both belonging to serotype VI (Fig. 6.2).

6.3.2 Analysis of accessory genome content

Only one genome (ST1 serotype V) carried human-associated genetic markers *scpB-lmb*; this assembly was also the only one that did not code for locus 3 genes.

Lac.2 was detected uniquely among milk isolates (*n*=54, 72% of all milk isolates, Fig. 6.2), except for one isolate that originated from the nose of a calf (ST617). This prevalence was higher when considering only milk isolates from population 3 (ST616) (82%), and much lower in milk isolates from population 5 (ST617 and ST1652 serotype VI) (33%). Two known Lac.2 variants were identified (Lac.2b, *n*=9; Lac.2d, *n*=31). A new Lac.2 variant was also detected and named Lac.2e, following the progressive nomenclature (Crestani et al., 2021; Sørensen et al., 2019). Lac.2e (length=9,535 bp) showed the same gene arrangement as Lac.2a, with the exception of an additional gene in Lac.2e, a glucokinase (*glk*, length=951 bp), upstream *lacA*.

Tetracycline-resistance (TcR) determinant *tet*(M) was detected in 96 genomes; four of these encoded a variant of Tn*916* that also carried a *tet*(L) gene, which caused the blast

[2]The identification of multiple subpopulations within the camel-specific lineage CC609 (which belongs to a single population in chapter 4), depends on the scale of the dataset being examined (i.e. major lineages will be identified if the whole GBS population is analysed, while subpopulations of these lineages will be defined when each lineage is analysed separately).

search to give a negative result (two fragments, which separately had a QC <80%) (Fig. E.2). *Tet*(M) and *tet*(L) were always carried by *Tn916* and *Tn916*-like elements (Fig. E.2), and this was confirmed by blastn searches. TcR was particularly prevalent among milk

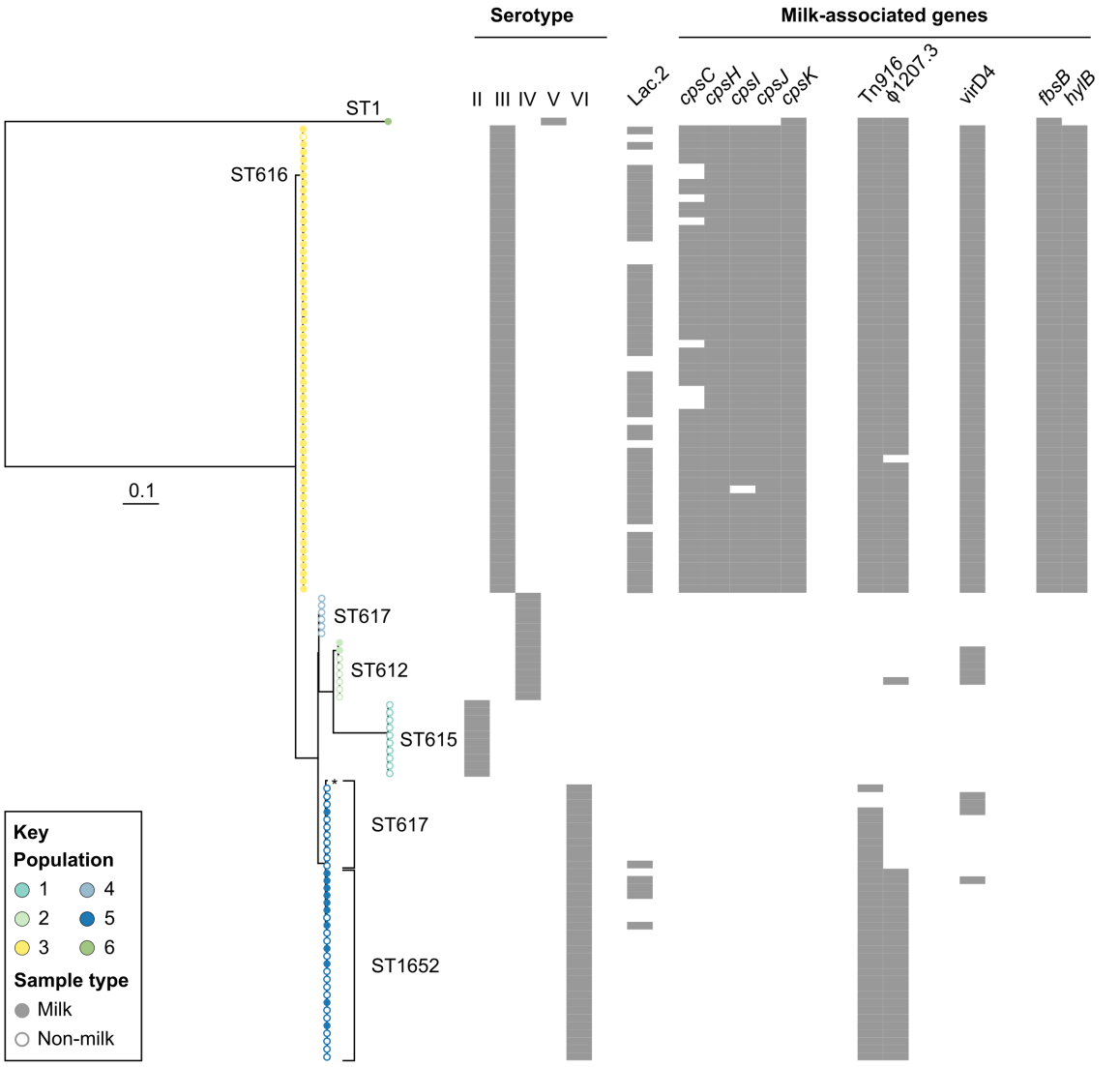


Figure 6.2: Maximum-likelihood phylogenetic tree of 122 group B *Streptococcus* (GBS) genomes from Kenyan camels. Leaves are coloured based on BAPS populations (n=6), with different shading for milk isolates (n=75; full circles) and non-milk isolates (n=47; open circles). Major sequence types (ST) are indicated close to branches or leaves (the two isolates belonging to ST1653 and ST1654 are comprised within the ST616 clade). Grey blocks on the right indicate the serotype each isolate belongs to, whether it encoded for the Lac.2 operon, and the presence of major genes associated with the milk phenotype, as per genome-wide association study (GWAS) results. Reference genome HF952106 is indicated with an asterisk. Tree was rooted at midpoint.

isolates (97%) compared to other sample types (49%), and prevalence was higher among isolates from lactating females (86%) compared to calves (50%), which is inherent to the sample-type association as only lactating females yield milk samples. All ST616 genomes were *tet*(M)/Tn*916*-positive (Fig. 6.2).

Fifty-six assemblies carried complete prophages, which all belonged to integrase types that had been previously classified (Crestani et al., 2020). The most common were prophages GBS11.1 (*n*=39), followed by GBS10 (*n*=18) and GBS11.2 (*n*=4). The majority of genomes carried only one complete prophage, whilst five encoded two (*n*=3 GBS10-GBS11.1; *n*=2 GBS10-GBS11.2). Ninety-eight genomes had between 1 to 6 incomplete/remnant prophages (ϕ1207.3 excluded, see below) (Tab. E.1). PICI were detected with PHASTER in 33 assemblies; of these, four carried PICI2, which is described in chapter 2 and had been identified in a GBS genome from a Kenyan camel (ILRI005) (Crestani et al., 2020). Two new PICI were detected in this dataset (Fig. 6.3): PICI3 (length=13,723 bp), which shares the same insertion site as PICI1 and PICI2 (*rpsD* - 30S ribosomal protein S4), and PICI4 (length=12,919

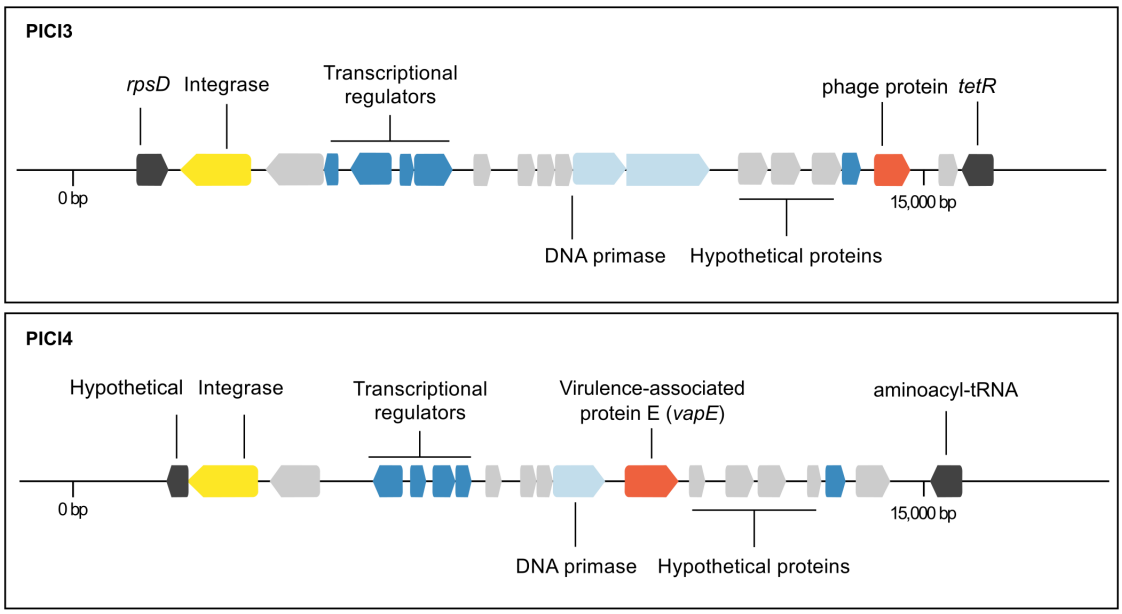


Figure 6.3: Annotated maps of genes in phage-inducible chromosomal island (PICI) 3 (isolate 84Ob) and PICI4 (isolate E7). The integration site of PICI3 is the same as that of PICI1 and PICI2 (described in chapter 2), the *rpsD* gene, whereas for PICI4 it is next to a hypothetical gene. Genes are colour-coded based on function (black: chromosomal genes; yellow: site-specific integrase; dark blue: lysogeny genes; light blue: replication genes; light grey: hypothetical; red: other genes).

bp), which is found between a hypothetical gene and an aminoacyl-tRNA synthetase gene. PICI2 (n=4) and PICI3 (n=2) were detected uniquely among ST617 serotype IV genomes (population 4), whilst PICI4 (n=27) was found in both ST617 and ST1652 (serotype VI, population 5). PICI4 carries a virulence gene, the virulence associated protein E (*vapE*).

6.3.3 Camel-associated genes

Through comparison with GBS genomes from other host species, scoary identified several genes that were positively associated with the camel host, many of which (n=23 among the first fifty best-scoring genes) were unique to this phenotype (camel-specific). Most high-scoring genes primarily mapped to four genomic islands (GEI) (Fig. 6.4).

Significant genes in island 1 included genes for carbohydrate metabolism (sugar-phosphatase),

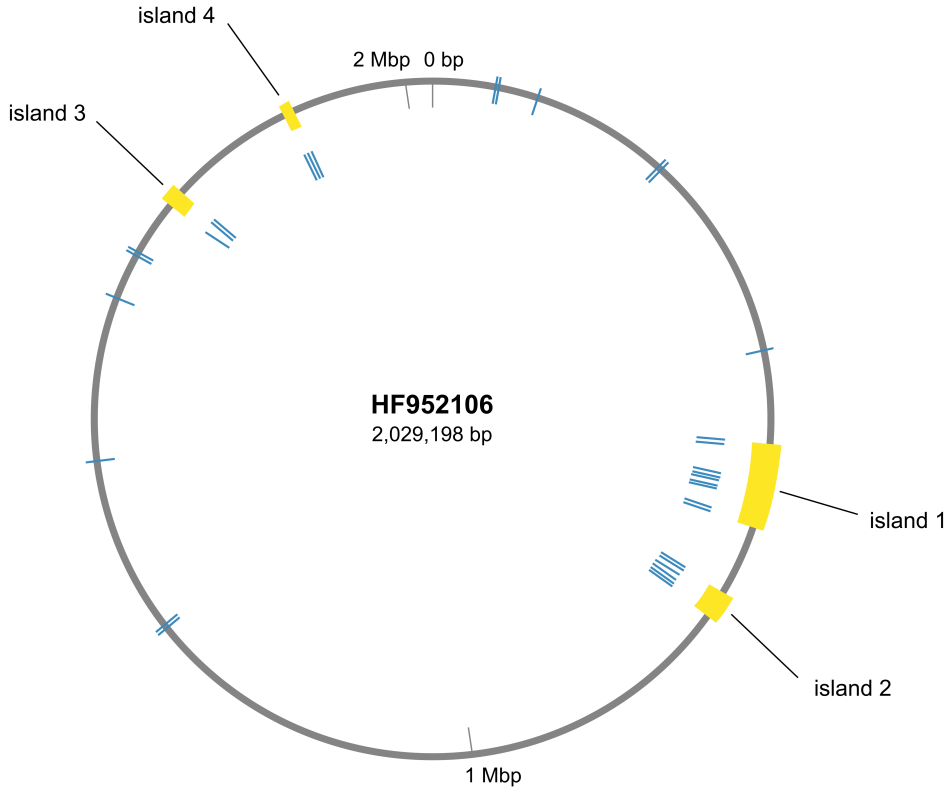


Figure 6.4: Camel-associated genes in group B *Streptococcus* (GBS), as detected by scoary, mapped to reference genome HF952106. Single genes are indicated with blue lines, while genomic islands (GEI) comprising areas of higher density of camel-associated genes are indicated as yellow blocks (GEI 1-4).

for DNA modification enzymes (DNA topoisomerase and DNA cytosine methyltransferase) and protein channels (voltage-gated chloride channel family *eriC*). In island 2, high-scoring genes included those coding for proteins involved in interaction with metals (sensor histidine kinase *cusS*, IMMA/IrrE family metallo-endopeptidase, CPBP family intramembrane metalloproteases) as well as a transcriptional regulator (Rgg/GadR/MutR), an oxidoreductase (Gfo/Idh/MocA), a histidine phosphatase, an N-acetyltransferase (GNAT family), an ATP-dependent DNA helicase and a pyrimidine nucleotidase (YjjG). In island 3, significant genes included a carbohydrate kinase (FGGY family) and a PTS sugar transporter subunit IIC, whilst island 4 genes were all annotated as hypothetical. Four insertion sequences (IS) appeared among the fifty best-scoring genes: IS*1562*, IS*Lre2*, IS*Sag9* and IS*Sag8*. The former two were unique to the camel phenotype (sensitivity: SE 97.7%, specificity: SP 100%, positive predictive value: PPV 100%, negative predictive value: NPV 99.6%), whilst the latter two were also carried by fifty non-camel genomes. In addition, a type I M restriction modification system (RMS) was found as being associated with camel GBS (present in *n*=121/131 camel genomes, SE 96.9%, SP 92.9%, NPV 99.5%), although its PPV was low (67.9%) as it was also found in sixty non-camel genomes.

Genes in island 2 were among the best-scoring entries in the scoary output, in particular the two CPBP family intramembrane metalloproteases (SE 98.5%, SP 100%, PPV 100%, NPV 99.8%), followed by non-GEI genes such as a hypothetical protein, IS*1562* and IS*Lre2*. Among the first fifty highest scoring genes, only eight were negatively associated with camels.

6.3.4 Camel milk-associated genes

In addition to camel-associated genes, genes specifically associated with camel milk isolates were detected (camel milk vs camel non-milk). Within the first fifty best-scoring genes, the majority (*n*=40) were positively associated with milk. Nine of these genes did not belong to GEI, among which *fbsB* (fibrinogen-binding protein B) and *hylB* (Fig. 6.2), and five genes belonged to the capsular serotype locus *cps* (*cpsC*, *cpsH*, *cpsI*, *cpsJ* and *cpsK*). Genes from the *cps*, *fbsB* and *hylB* had the best PPV for camel milk (98.1-98.4%), although NPV tended to be lower (66.7-76.7%) (Tab. 6.1). This is because these genes were highly specific

Table 6.1: Scoary results are reported for some of the most significant camel milk-associated genes. The Doc (death-on-curing) family toxin and *virD4* were chosen as representative genes for the two milk-associated genomic islands (GEI) Tn*916*-ϕ1207.3 and *virD4* GEI, respectively.

Gene	SE	SP	PPV	NPV	*p*-value
fbsB	81.3	97.9	98.4	76.7	7.63×10^{-20}
hylB	80.0	97.9	98.4	75.4	5.62×10^{-19}
cpsC	69.3	97.9	98.1	66.7	6.32×10^{-15}
cpsH	80.0	97.9	98.4	75.4	5.62×10^{-19}
cpsI	78.7	97.9	98.3	74.2	1.14×10^{-18}
cpsJ	80.0	97.9	98.4	75.4	5.62×10^{-19}
cpsK	81.3	97.9	98.4	76.7	7.63×10^{-20}
Doc toxin	96.0	68.1	82.8	91.4	1.14×10^{-14}
virD4	84.0	85.1	90.0	76.9	2.86×10^{-14}

(SP 97.9%) for camel milk isolates, as only one non-milk isolate was carrying these genes (ST616 from a calf mouth sample), but had low sensitivity (SE 69.3-81.3%), as a number of milk isolates did not code for them. Notably, these genes were associated with ST616, but were lacking from most of the other ST (with the exception of *cpsK* and *fbsB* in the ST1 isolate) (Fig. 6.2).

Genes positively-associated with milk, specifically most of those ranked between positions 26 and 142 by scoary, belonged to two GEI (Fig. 6.5): one is the *tet*(M)-carrying Tn*916*, and the other is a 14-gene cluster ($\sim$15,000 bp), which shows signatures of mobility (e.g. transposase, relaxase and mobilisation genes) and which carries the *virD4* gene. Interestingly, Tn*916* in the majority of camel milk isolates is linked to a long gene cluster ($\sim$43,000 bp, Fig. 6.5) that was also detected as significantly associated with milk isolates. This region was identified as an incomplete prophage by PHASTER, as it was lacking its integrase gene, indicating that it is likely hijacking Tn*916* to be mobilised. When the region surrounding the Tn*916*-prophage was blasted against the general ICEberg database with multigene blast (M. Liu et al., 2018), it showed high sequence similarity with a segment of Tn*1207.3* from *Streptococcus pyogenes* strain 2812A (Santagati et al., 2003) and strain MGAS10394 (Fig.

6.5). The same region also showed sequence similarity with two composite MGE from *Streptococcus suis* (CMGETZ080501 and CMGEYY060816) (Fig. 6.5). This prophage, named ϕ1207.3 (Iannelli et al., 2014), carries a type II toxin/antitoxin system with a Phd/YefM fam-

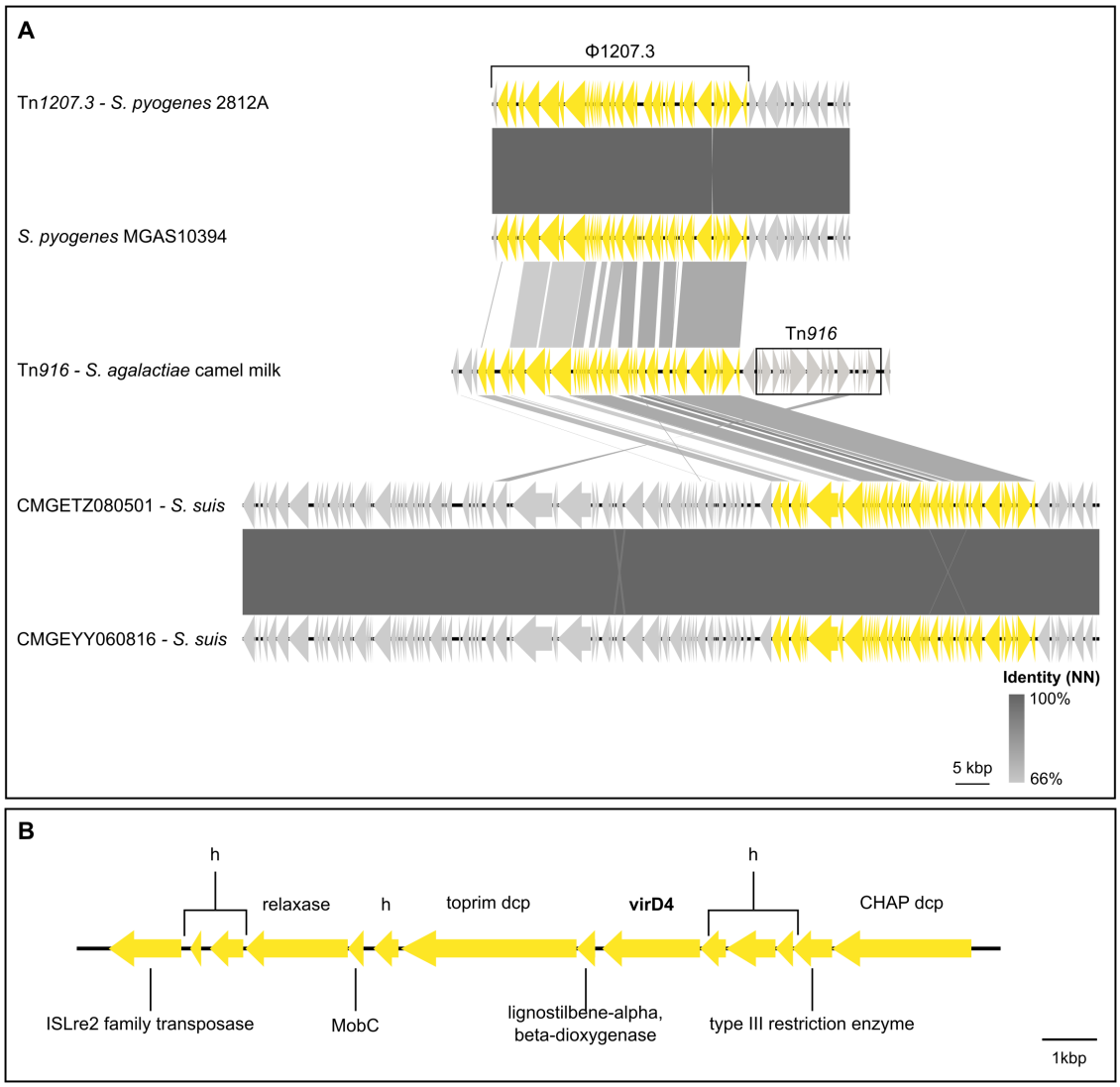


Figure 6.5: Diagrams of milk-associated genomic islands (GEI): Tn*916*-ϕ1207.3 (A) and *virD4* GEI (B). A) Visualisation of a blastn comparison between the Tn*916*-ϕ1207.3 associated with camel milk isolates (centre, from isolate 1M) and mobile elements from *Streptcoccus pyogenes* (top, Tn*1207.3 S. pyogenes* strain 2812A and MGAS10394) and from *Streptococcus suis* (bottom, CMGETZ080501 and CMGEYY060816). Figure was obtained with Easyfig v2.2.2 (Sullivan et al., 2011) and modified with Inkscape (`www.inkscape.org`). B) Visualisation of gene orientation and annotation of the *virD4* GEI from isolate 1M. The contig on which this element was located ended at the right hand side of the CHAP domain containing protein (dcp). Hypothetical proteins are indicated with h.

ily antitoxin followed by a Doc (death-on-curing) family toxin. Among the genomes that carried Tn*916* (*n*=96), ten did not carry ϕ1207.3 (*n*=9 ST617 serotype VI and *n*=1 ST616) (Fig. 6.2). Only one genome (ST612) was ϕ1207.3-positive but *tet*(M)/Tn*916*-negative. In contrast to *cps*, *fbsB* and *hylB*, ϕ1207.3, specifically its Doc toxin gene, showed a high NPV (91.4%), and a lower PPV (82.8%) (Tab. 6.1); this is linked to few milk isolates lacking ϕ1207.3 (SE 96.0%) but a good proportion of non-milk isolates carrying it (SP 68.0%), particularly among ST1652 (Fig. 6.2). In contrast, the *virD4* GEI has a better PPV (90.0%) and a lower NPV (76.9%). Similar to *cps*, *fbsB* and *hylB*, *virD4* is associated with ST616, but in contrast to these genes, *virD4* GEI was also detected among some non-milk isolates from ST612 and ST617.

Of note, genes belonging to the lactose operon (Lac.2) scored much lower than the MGE mentioned above (positions 198-246 in the scoary ranking). They showed very low sensitivity for the milk phenotype (SE 12-60%, NPV 41.6-60.5%, depending on the gene variant), as Lac.2 was not present in 28% of milk isolates, but higher specificity (SP 97.9-100%, PPV 96.8-100%), as only one non-milk isolate was carrying Lac.2 (an ST617 isolate from a calf's nose).

6.4 Discussion and conclusions

Prior to 2021, limited genomic data were available on GBS from camels (Fischer et al., 2013; Zubair et al., 2013), and consequently our understanding of how camel GBS isolates fit into the context of the wider GBS population was limited too.

In this study, the camel GBS population structure based on the core genome of a collection of GBS assemblies from lactating camels and their calves originating from Kenya was analysed. I detected eight ST, which were grouped into six populations, and that mostly belonged to ST that have previously been reported from camels (Fischer et al., 2013). With the exception of one ST1 isolate, all genomes clustered within the camel-specific lineage CC609 that I describe in chapter 4. Analysis of the camel GBS population structure carried out by Fischer et al., 2013, identified two groups of camel GBS isolates based on a MLST maximum-likelihood (ML) phylogeny: one clustered between ST17 and ST23 (ST609 and

ST614), while the other clustered at the opposite end of the unrooted phylogeny, with the closest clade being that of CC12 (ST610-ST613, ST615-ST618). The authors analysed these data in the context of human GBS isolates, identifying the first cluster as sharing ancestry and the occurrence of possible genetic exchange with ST26, which is considered a human-associated lineage. During phylogenetic analyses of a large dataset of GBS isolates I conducted in chapter 4, I showed how, when considering the entirety of the core genome, ST609 and ST614 cluster together with the other ST identified in camels, forming a monophyletic camel-specific lineage (CC609). This finding highlights the limitations of phylogenetic inferences based on MLST data, as explained in detail chapter 4 and in the paper from Sørensen et al., 2010. In addition, I showed that CC609 and CC26 are not very close relatives, rather, the closest relatives of CC609 are CC103/314 (chapter 4, Fig. 4.1) and CC552, similar to what was reported by Richards et al., 2019. Moreover, there is little shared recombination between CC609 and CC26 (chapter 4, Fig. 4.3), which is indicative of limited genetic exchange between the two lineages, as opposed to what was reported by Fischer et al., 2013, with their MLST-based results. Although the majority of isolates from camels belong to a GBS population that has never been isolated from other host species (host-specialist), the detection of a ST1 serotype V isolate in a milk sample indicates the possibility of interspecies transmission, as reported in Seligsohn et al., 2021a. CC1 is a host-generalist, and ST1 serotype V is often isolated from humans and dairy cattle (Sørensen et al., 2019; Lyhs et al., 2016), with variable prevalence of host-associated markers, *scpB-lmb* and Lac.2 (chapter 5). The fact that this specific ST1 isolate showed genetic markers of human host-adaptation (*scpB-lmb* transposon) and that it did not code for any of the camel-associated genes and GEI (data not shown), likely points to a human-to-camel transmission event. Reverse zoonotic transmission of GBS between human and dairy cattle has also been described (Crestani et al., 2021), with the introduction of GBS generalist lineage CC1 in dairy cattle from humans, as detailed in chapter 3. Close contact between humans and camels, coupled with low hygiene standards in particular during milking (e.g. no hand washing prior to milking or between animals), in part due to the limited access to water sources, may allow for human-to-camel transmission of GBS (Seligsohn et al., 2020). The possibility of strains from the camel-specific lineage (CC609) colonising or infecting people, although unlikely based on genomic characteristics of this lineage, cannot be ruled out. To date, no genomic data are available from GBS iso-

lates from camel keepers, but high-risk practices such as the routine consumption of raw milk within pastoralist communities pose a potential threat to human health (Seligsohn et al., 2020). Only further work involving parallel sampling of camels and their herds-persons, similar to what has been done for dairy cattle and farm workers in Colombia and Denmark (Cobo-Ángel et al., 2019; Sørensen et al., 2019), can help clarify the full extent to which interspecies transmission might be happening, and highlight which lineages can pose a risk for zoonotic transmission. In addition, camel GBS could act as a reservoir of capsular serotypes that are rare or emerging in humans (IV and VI) (Lyhs et al., 2016; Teatero, Athey, et al., 2015) and that would not be covered by vaccines currently under development (Ia, Ib, II, III, V) (Kobayashi et al., 2019). Similar to what has been observed in *Streptococcus pneumoniae* (Brueggemann et al., 2007), the introduction of a vaccine could potentially generate escape mutants in case of capsular switching, a well-documented phenomenon in GBS (Neemuchwala et al., 2016; Bellais et al., 2012; Martins et al., 2010), between camel and human isolates.

In addition to the population structure of camel GBS based on the core genome, the accessory genome content in terms of known host- and niche-associated genes, AMR genes and MGE was assessed. Previous work on accessory genes from camel GBS had uniquely evaluated the presence of GEI and AMR genes. GEI (n=6-7) were detected in two camel GBS isolates by Zubair et al., 2013, but no further analyses are reported, whereas Fischer et al., 2013, explored the accessory genome only in terms of AMR, reporting the presence of TcR, with resistant isolates carrying the *tet*(M) gene on ICE Tn*916*. In my work, I found an overall higher prevalence of TcR compared to Fischer et al. (79% vs 34%), which was particularly associated with milk isolates (97.3%) compared to other samples types (48.9%) (Tab. E.2). This is in agreement with previous reports of low levels of TcR among GBS from nasal isolates (Mutua et al., 2017). Geographical origin must also be considered: when excluding milk isolates from Somalia (all TcR negative) from Fischer et al. and only considering those originating in Kenya, the prevalence of TcR is higher (72.4%) but still less than what detected in my dataset (Tab. E.2). Likewise, TcR in non-milk samples from this work is higher than that reported in GBS from camels for the same country by Fischer et al., (20.4%). In Kenya, there is a lack of control over the sales and usage of antimicro-

bials (Heffernan & Misturelli, 2000) and they are commonly employed in the treatment of bacterial infections in camels (Lamuka et al., 2017; Younan, 2002). High levels of TcR are reported in people and animals in several bacterial species from other African countries, such as Tanzania, where antimicrobials are available over-the-counter (Subbiah et al., 2020; Caudell et al., 2017). In this study, in addition to the classic Tn*916* variant that is reported in GBS from humans, cattle and fish (Crestani et al., 2021; Barkham et al., 2019; Da Cunha et al., 2014), I detected for the first time a variant of Tn*916* that carries both *tet*(M) and *tet*(L) genes. Several other ICE carrying AMR determinants have been reported in GBS, mostly among human isolates and carrying genes for erythromycin resistance (ErmR), such as Tn*3872* and ICE*sp2905* (Oppegaard et al., 2020), or multidrug resistance, ICE*Sa2603* (ErmR, some TcR and streptothricin) (Oppegaard et al., 2020), which has also been reported in bovine GBS (Huang et al., 2016). The detection of a novel Tn*916* with multiple TcR determinants highlights the plasticity of ICE in the acquisition of new resistance genes and confirms the importance of ICE as vehicles of AMR determinants in GBS, and more generally in streptococci (Oppegaard et al., 2020).

GWAS analysis of milk vs non-milk isolates confirmed that Tn*916* is associated with milk isolates, and also that this ICE is linked to an incomplete prophage (ϕ1207.3) in most milk isolates. This same prophage had been described as part of other ICE/MGE in *S. pyogenes* and *S. suis*, which are associated with human and porcine hosts, respectively (Iannelli et al., 2014; Santagati et al., 2003). Unlike the element described in *S. pyogenes*, ϕ1207.3 in GBS did not code for macrolide resistance genes, but it encoded the same toxin/antitoxin (TA) system (Phd/Doc). Phd/Doc TA systems have been described in various gram-positive (Behrooz et al., 2018; Chan et al., 2014) and gram-negative bacterial species (Guérout et al., 2013; Lehnherr et al., 1993). Under stressful environmental conditions for the bacteria, the antitoxins are usually rapidly degraded, and the toxins are free to interfere with essential bacterial cellular processes, such as replication, translation and cell-wall synthesis (Q. E. Yang & Walsh, 2017; Chan et al., 2014). The Doc toxin of the Phd/Doc systems blocks protein synthesis, however, high levels of toxin have been shown to de-repress rather than repress transcription (Q. E. Yang & Walsh, 2017; R. Page & Peti, 2016; De Gieter et al., 2014). The biological impact of TA systems in virulence and pathogenicity remains unclear (Chan et al.,

2014).

In addition to incomplete prophage ϕ1207.3, remnant prophages were detected in the vast majority of genomes (91.8%). Also, complete prophages of type GBS10 and GBS11 (Crestani et al., 2020) were highly prevalent in this dataset (46%). Only one previous study had investigated the presence of prophages in camel GBS, but this was limited to one genome in which only GBS1 was identified (Richards et al., 2019). GBS10 and GBS11 prophages have been described uniquely among human isolates, and one seal isolate in case of GBS11.1 (Crestani et al., 2020). Prophages can carry important toxin genes that have an impact on pathogenicity, such as for the Shiga toxin in *Escherichia coli* (Meltz Steinberg & Levin, 2007), as well as virulence and AMR genes, as shown in *Staphylococcus aureus* (Dini et al., 2019). Further bioinformatic analyses on prophages from camel GBS are needed to determine whether they carry such genes, which could be followed by wet-lab experiments in case these are identified. Other MGE that can severely impact on pathogenicity of isolates, and which are bound to prophages for the completion of their life cycle, are PICI. As an example, in *S. aureus* PICI carry the toxic shock syndrome toxin-1 (TSST-1) (Lindsay et al., 1998), which is responsible for the clinical manifestations of the toxic shock syndrome (Todd et al., 1978). Previous work on PICI in a large GBS dataset identified a low diversity of PICI (Crestani et al., 2020) compared to other bacterial species (Penadés & Christie, 2015), with only two types being detected (chapter 2). PICI1 was widespread across isolates from multiple host species (humans, fish, cattle, a dog and a dolphin), whilst PICI2 was found uniquely in a camel GBS genome from Kenya (Crestani et al., 2020). In contrast with the low diversity of PICI in other host species, I detected several PICI types (three in total, of which two were new types) in this dataset of camel GBS. The first new PICI, PICI3, had the same integration site as PICI1 and PICI2, the *rpsD* gene, confirming this site is an important hotspot for recombination of PICI in GBS. Interestingly, the second novel PICI, PICI4, carried a virulence-associated protein (*vapE*). This virulence gene has been detected in *Rhodococcus equi* (Takai et al., 2000) and *S. suis*, in which it has been associated with pathogenicity of serotype 2 strains in pigs (Ji et al., 2016). It has also been recently described as significantly associated with PICI in *S. pneumoniae* (Shaw, 2021). Of note, PICI4 was associated with a sublineage (population 5) whose isolates comprise both

colonisers of extramammary body sites of healthy animals and mastitis-causing isolates (Fig. 6.2). Further work is needed to clarify the possible role of this virulence gene in camel GBS.

One particular sublineage was significantly associated with mastitis, ST616, which agrees with previous findings (Fischer et al., 2013). Other sublineages, such as ST612, ST615 and ST617-serotype IV were mostly associated with extramammary body sites of healthy animals, whilst population 5 (ST617-serotype VI and ST1652) showed a mixed pattern (Fig. 6.2). This suggests different niche predilection within the wider camel-specialist lineage CC609, with ST616 being an udder-specialist, ST612-ST615-ST617 (serotype IV) being extramammary and particularly nose colonisers, whilst ST1652-ST617 (serotype VI) is a tissue-generalist sublineage that has the ability to colonise and/or cause infection in different body sites (nose coloniser/mastitis-causing). In the study from Fischer et al., 2013, ST617 serotype VI was isolated primarily from abscesses but also from milk, suggesting that skin, soft tissue and udder infections may be opportunistic. The acquisition of advantageous accessory genes could play a role in adaptation of these isolates to the camel mammary gland. Genes that have been associated with the successful colonisation, survival and infection of the mammary gland in another major GBS host species, dairy cattle, are those belonging to the Lac.2 operon (Richards et al., 2011) (chapter 5). These genes are responsible for the utilisation of lactose (Lyhs et al., 2016; Richards et al., 2011), of which the bovine udder is rich (niche-adaptation), providing a survival advantage thanks to substrate utilisation (S. K. Sheppard et al., 2018). Lac.2 has also been identified as a genetic determinant of mastitis-associated *Klebsiella pneumoniae* (Holt et al., 2015). Although the concentration of lactose in camel milk is similar to that of cows (Yoganandi et al., 2014), the overall Lac.2 prevalence was lower in camel milk isolates (72%) compared to bovine mastitis isolates (98%, chapter 5). This was higher when considering uniquely the mastitis-associated sublineage ST616 (82%), whereas it was much lower among the niche-generalist ST617-ST1652 (33%) and close to absent in non-milk isolates. These data suggest that Lac.2 is not necessary for GBS to successfully colonise and establish infections in the udder of camels, but that it offers an evolutionary advantage in adaptation to the mammary gland. This is further supported by the GWAS results on camel milk vs non-milk isolates, which showed a high specificity of Lac.2 for milk/mastitis (useful/associated trait), but a low sensitivity (not nec-

essary). In addition, I detected a new variant of the Lac.2 which encodes a glucokinase (*glk*), a gene for the metabolism of glucose (phosphorylation). In *Streptococcus mutans*, *glk* and *lacR* regulate each other's expression (Zeng & Burne, 2021a, 2021b), optimising cell growth in the presence of different substrates, i.e. growth in a lactose-rich environment lowers *glk* activity in *S. mutans* compared to growth on glucose (Zeng & Burne, 2021b). The presence of *glk* within the lactose operon of certain camel GBS isolates could constitute an advantage in the presence of both high concentrations of lactose and glucose (which is also a product of lactose fermentation). Additional wet-lab experiments are needed to clarify the role of this gene for the successful survival of GBS in camel milk.

The majority of milk-associated genes reported by GWAS were also strongly associated with the sublineage ST616 (Seligsohn et al., 2021a). Although scoary has shown to successfully correct for population structure compared to other GWAS methods such as pyseer (chapter 5), if a single lineage is predominantly associated with a phenotype/trait of interest, as is the case of ST616 and camel mastitis, lineage-associated features will be reported as significant in GWAS regardless of their functional role. This is the case as an example for capsular genes of serotype III, which was only detected among ST616 isolates in the camel dataset, and that consequently showed association with the milk phenotype. Serotype III is not specifically associated with mastitis in dairy cattle (chapter 5) (Lyhs et al., 2016; Dogan et al., 2005) and capsule pseudogenisation in bovine GBS suggests that it has no role in causing bovine mastitis (Almeida et al., 2016). In contrast to Lac.2, for which functional association with mastitis can be attributed to the utilisation of a highly available substrate, there is no explanation for a biological association of serotype III and camel mastitis. Interestingly, other milk and ST616-associated genes found using GWAS were all genes associated with invasion. In particular in human GBS, the fibrinogen-binding protein *fbsB* promotes invasion of epithelial cells (Gutekunst et al., 2004) and the hyaluronate lyase *hylB* is able to degrade certain components of the extracellular matrix and is believed to contribute significantly to invasion (Herbert et al., 2004a; Rolland et al., 1999). However, human invasive isolates, in particular those associated with high risk of neonatal meningitis (Domelier et al., 2006), can show an inactive *hylB* due to the insertion of IS*1458* within the gene. Although no difference has been found in intracellular invasion of GBS between human isolates carrying

the disrupted gene and bovine isolates functionally expressing the gene (Sukhnanand et al., 2005), a potential functional role of *hylB* in the establishment of camel mammary infections cannot be ruled out. The *virD4* gene, a previously identified type IV secretion system (T4SS) (Schulein et al., 2005), was recently described in GBS as significantly associated with CC19 (Gori et al., 2020). This gene is found in various bacterial species, including *S. suis* serotype 2 (X. Jiang et al., 2016), and it is associated with conjugation, translocation of virulence factors (Wallden et al., 2010; Alvarez-Martinez & Christie, 2009), anti-phagocytic activity and a pro-inflammatory effect (X. Jiang et al., 2016). Similar to Lac.2, *fbsB*, *hylB* and *virD4* do not seem to be necessary to cause mastitis in camels, as non-ST616 isolates from milk lack them, but they could provide an advantage in the colonisation and invasion of the mammary gland. This is true also for genes encoded by Tn*916*-ϕ1207.3. More work will need to be carried out in order to understand their possible role in camel mastitis.

Finally, through GWAS, I detected for the first time camel-associated genes, which mainly clustered in four GEI. These GEI, in particular the larger islands 1 and 2, do not show classical signatures of mobility, such the presence of an integrase (e.g. close to Lac.2), or relaxases and T4SS, which are typical of ICE. Interestingly, three GEI carried genes involved in various metabolic processes, in particular carbohydrate and metal utilisation. In this aspect, camel-associated GEI are similar to Lac.2 and locus 3 (Delannoy et al., 2016), which are genomic islands that carry genes for metabolism of carbohydrates (lactose and galactose, respectively). The bovine mammary gland and the central nervous system of fishes are rich in those carbohydrates, and utilisation of substrates that are present in certain niches or host species is a well-described mechanism of bacterial adaptation (S. K. Sheppard et al., 2018). In addition, the ability to uptake and utilise essential metals can counteract the host immune defences, which usually sequester these molecules to protect the host from infection (Mortensen & Skaar, 2013). Essential nutrient metals vary based on bacterial species and the niche they inhabit (Mortensen & Skaar, 2013), and this could also be the case for camel GBS. At this stage, it is not clear which carbohydrate or metal molecules in particular are utilised by these GEI, and therefore further research into the function of these genes and the role they play in adaptation to the camel host is needed. One limitation of this work is that all GBS genomes from camels included in this work originated from only two countries, Kenya

and Somalia. Additionally, due to the nature of the studies from which sequence data were obtained, in particular Seligsohn et al., 2021a, which aimed at investigating the diversity of GBS from camel milk between farms, multiple isolates per herd per year were included, in contrast with the approach from chapters 4 and 5. Sequencing of GBS from other camel populations, such as those of the Middle East, India and of feral and farmed camels in Australia (Saalfeld & Edwards, 2010), could provide a more complete picture of the camel GBS population in terms of lineages and of genes associated with this species.

In conclusion, I showed the existence of one camel-specific GBS lineage (CC609) which comprises the vast majority of isolates from carriage and disease in this host species. I detected toxin genes and genes for substrate utilisation - carried on several MGE types - that could play a role not only in adaptation to camels, but also to the different niches within this host (nose, mammary gland). Further bioinformatic work, such as comparative genome analysis including sequenced data from camel GBS originating from countries outside the Horn of Africa, and laboratory experiments will be important to better understand the role of these genes in the ecology of GBS in camels.

Chapter 7

General discussion

7.1 The mobilome: dynamic molecular parasites shaping bacterial populations

During the past two decades, the advances of next generation sequencing (NGS) technologies enabled scientists to carry out the first studies aimed at comparing bacterial genomes, initially on a small scale and then on a progressively larger scale (Tettelin et al., 2008, 2005; Medini et al., 2005). These studies showed that bacterial genomes consist of a core - a set of genes that is present among all isolates of the same species - and of an accessory component, that is variably present in some but not all isolates (Frost et al., 2005), which together form the pangenome (i.e. 'whole' genome) (Tettelin, 2009). Some bacterial species, such as *Staphylococcus lugdunensis* (Argemi et al., 2018), *Staphylococcus aureus*, *Streptococcus pyogenes* (group A *Streptococcus*, GAS) and *Bacillus anthracis* (Tettelin et al., 2008, 2005; Medini et al., 2005), have so called 'closed' pangenomes (Vernikos et al., 2015), whilst others, such as group B *Streptococcus* (GBS) and *Streptococcus pneumoniae* (Tettelin et al., 2008), have 'open' pangenomes, whose size continues to increase when adding new sequenced isolates. Pangenomes can be considered good representations of the bacterial species' gene pools, which is broader in these latter species compared to the former, assuming there is no sampling bias (Vernikos et al., 2015); in fact, for multi-host pathogens, the inclusion of isolates limited to one host species when studying the pangenome is a common cause of bias in genomic studies. As an example, in GBS, the inclusion of the fish-specific

lineage (clonal complex (CC) 552), which has a reduced genome size (1.8-1.9 Mbp as opposed to an average of 2.1 Mbp for the other lineages, Fig. C.2), would reduce the GBS core genome, compared to a pangenome analysis uniquely comprising human isolates. On the other hand, excluding bovine GBS isolates could reduce the accessory genome, as multiple mobile genetic elements (MGE) (e.g. plasmids and insertion sequences, IS) are found in the bovine-specific lineage CC61/67 (chapter 3).

The accessory genome, of which a substantial proportion consists of MGE, is thought to play an important role in ecology and evolution, particularly of multi-host bacterial pathogens (S. K. Sheppard et al., 2018; Richardson et al., 2018). MGE can be regarded as molecular parasites that move between cells through horizontal gene transfer (HGT) (Koonin & Wolf, 2008), and are collectively referred to as the 'mobilome' (Frost et al., 2005). The mobilome was first mentioned in relation to GBS in the work of Lopez-Sanchez et al., 2012, who proposed that a particular type of clustered regularly interspaced short palindromic repeats (CRISPR) system is responsible for the modulation of MGE in its cells. CRISPR are part of the bacterial immune system together with restriction modification systems (RMS) (Rodic et al., 2017). RMS can also regulate the presence and prevalence of MGE within bacterial genomes and they can shape bacterial populations and lineages (Budroni et al., 2011; Lindsay, 2010). Unlike CRISPR, RMS are often carried by MGE, which use RMS as a way of promoting their own survival within a new cell (Sánchez-Busó et al., 2019; Koonin & Wolf, 2008). Like for RMS, other MGE that constitute the mobilome, such as bacteriophages, plasmids and transposable elements (chapter 1), carry accessory genes that can be crucial for the adaptation and survival of a cell in a new environment or host species (e.g. new metabolic pathways) (S. K. Sheppard et al., 2018), and they can be vectors of virulence genes, toxins or antimicrobial resistance (AMR) genes (Koonin & Wolf, 2008). In addition, the mobilome not only determines the acquisition or loss of genes that are part of the accessory genome, but it often has an impact on the core genes as well (Koonin & Wolf, 2008). As an example, bacteriophages can package large segments of the chromosome (Chen et al., 2018) that are transferred to a new cell and that can substitute the original DNA sequence thanks to homologous recombination (Frost et al., 2005). Moreover, some MGE can also play a role in gene expression, silencing or upregulating genes, based on whether the integration site is located

within or upstream the gene, respectively (Siguier et al., 2014; Curcio & Derbyshire, 2003).

The role played by the mobilome in GBS has been studied for some of its epidemiological aspects, such as: i) the impact on pathogenicity of bacteriophages (van der Mee-Marquet et al., 2018; Salloum et al., 2010; Domelier et al., 2009; van der Mee-Marquet et al., 2006), IS (IS*1458*), and group II introns (GBSi1) (Domelier et al., 2009); ii) AMR linked with plasmids (Sendi et al., 2016; Compain et al., 2014; DiPersio et al., 2011; Horodniceanu et al., 1976) and integrative conjugative elements (ICE, Tn*916*), the latter also exerting an effect on the GBS population structure due to the selection[1] of resistant clones (Da Cunha et al., 2014). However, the role of MGE in terms of GBS adaptability to new hosts and niches had never been investigated extensively prior to this work. In my PhD project, I have run large-scale analyses on MGE in GBS, detecting both previously described and novel MGE (chapters 2, 3 and 6), I have carried out genome-wide association studies to detect genes associated with a particular host/ecological niche that could play a role in adaptation (chapter 5), and I have analysed the GBS population structure in terms of both core and accessory genome content, using the most representative and complete genomic dataset available to date, with representation of all major host groups (people, dairy cattle, fish) as well as a relatively poorly studied but increasingly important fourth host group, i.e. dromedary camels (chapters 4 and 6).

7.2 The detection of novel mobile genetic elements shows a high diversity of molecular parasites in GBS

MGE have been shown to play a role in various aspects of the ecology of GBS, in particular in the pathogenicity of certain lineages, as described above. Similar to other bacterial host species, such as *S. aureus* (Richardson et al., 2018), *Campylobacter jejuni* (S. K. Sheppard et al., 2013) and *Salmonella enterica* (Foley et al., 2013), MGE that carry advantageous

[1] Selection, in biology, is the preferential survival and reproduction or preferential elimination of individuals with certain genotypes, by means of natural or artificial controlling factors (Encyclopaedia Britannica, 2021).

accessory genes could also help GBS in adapting to different niches, in particular to different hosts, and could condition its tissue tropism. This is the case for the lactose operon (Lac.2) (Richards et al., 2013, 2011), an MGE that is responsible for the fermentation of lactose, which is found in almost all GBS bovine isolates (host-association, chapter 5) (Lyhs et al., 2016) and many camel mastitis isolates (Seligsohn et al., 2021a, 2021b). More broadly, Lac.2 represents an important marker of adaptation to the lactose-rich mammary gland niche (substrate utilisation), as confirmed by the presence of this element in other pathogens of the udder, such as *Streptococcus uberis*, *Streptococcus dysgalactiae* subsp. *dysgalactiae* (Richards et al., 2011) and *Klebsiella pneumoniae* (Holt et al., 2015), as discussed in more detail in the following section. Throughout my PhD project I detected a high number and variety of MGE types in GBS, some of which had already been described, such as certain prophages (van der Mee-Marquet et al., 2018) and some AMR plasmids (Sendi et al., 2016; Compain et al., 2014; DiPersio et al., 2011; Herbert et al., 2005), as well as several new MGE. The detection of a high number of novel MGE in GBS, including prophages, phage-inducible chromosomal islands (PICI) and plasmids, further supports the high diversity of MGE types in GBS and suggests their importance in its ecology.

In the case of bacteriophages, I broadened the knowledge on prophage diversity in GBS by expanding on the number of human genomes analysed - from a wider range of lineages - and by analysing for the first time genomes of animal origin for the presence of these MGE (chapter 2). One of the most interesting findings of this work was the detection of micro-evolution of prophage sub-lineages that acquired different integrases, which are the genes that determine where in the chromosome the prophages will insert. Through integrase shifts, prophages can acquire the ability to integrate in new chromosomal insertion sites, which has important implications in genome plasticity. As mentioned above, bacteriophages have been shown to be able to package considerable segments of the chromosome that are adjacent to their integration site in *S. aureus* (Chen et al., 2018). Acquiring the ability to insert in novel insertion sites means prophages may be able to package new portions of the host genome and this would have an impact on which sections of the chromosome will be affected by homologous recombination in the receiver cell. Not only could prophages package large sections of the GBS core genome downstream the twelve insertion sites I identified,

but they could also package other MGE that are inserted close to them (e.g. in the case of the prophage GBS5, for which the prophage was surrounded by other genes with signatures of an ICE, Fig. A.4), with phage particles acting as Trojan horses for both chromosomal genes, MGE and prophages. In addition, the identification of a high number of chromosomal insertion sites for prophages, even more than in *S. aureus* in which lateral transduction has been shown to influence genomic architecture (Chen et al., 2018), fits with GBS being a highly recombinogenic pathogen with a 'mosaic' genome (Tettelin et al., 2002). This mosaic structure was also shown by the presence of several Lac.2 integration sites (chapter 3), and highlights even more the importance of integration and recombination events in GBS evolution.

My work on GBS prophages does not include the characterisation of the presence of virulence, toxin and AMR genes that could be carried by these MGE and, if any, whether they are correlated with particular lineages or highly pathogenic isolates. This is something that could be further explored bioinformatically. In addition, both the prevalence of prophages and of their virulence/toxin/AMR genes by lineage and host species should be further investigated using a representative population sample of GBS from different host species, continents and type of isolates (carriage vs disease). The dataset on which I carried out analyses of prophages in chapter 2 was not the result of balanced sampling strategy, rather it comprised all the sequenced GBS genomes publicly available at the time, which made it impossible to draw conclusions about the prevalence of prophages in different lineages and host species. Nevertheless, the new prophage classification system I propose has so far shown itself to be comprehensive, as no new prophage types were detected in a new database of camel genomes, a host species for which only a few sequenced isolates were available prior to 2021 (chapter 6). Also, I did not investigate the direct impact of prophages on recombination events and if/how often these occur between different GBS lineages. Functional experiments similar to the ones carried out by Chen et al., 2018, in which genetic markers are inserted at progressively increasing distance from the integrated lysogenic prophages in the chromosome, could be designed to investigate the magnitude of lateral transduction in GBS. Furthermore, prophages' mobility between different lineages could be assessed with wet-laboratory experiments, as well as their ability to acquire new integrase genes.

In addition to new prophage types, I discovered other MGE in GBS: PICI and plasmids. I describe for the first time PICI as such in GBS and I observed an overall low diversity of PICI types, which is in contrast with other gram-positive (e.g. SaPI1 to SaPI5 and SaPIbov1 to SaPIbov5 in *S. aureus*) and gram-negative (e.g. fourteen putative PICI in *Escherichia coli*) bacterial species (Fillol-Salom et al., 2018; Chen et al., 2018; Martínez-Rubio et al., 2017). However, one type of PICI was widespread and highly prevalent across lineages, which could point to a possible fitness advantage of genes encoded by this element, for which however the functional role is not known (most genes are annotated as hypothetical). As for prophages, further bioinformatic work on PICI prevalence and association with certain isolate types (e.g. carriage vs invasive) in a more balanced dataset, together with an extensive analysis of PICI genetic content, is necessary to test the hypothesis of PICI1 conferring a fitness advantage. Although the diversity of PICI types was limited in GBS from the three major host groups (chapter 2), it was high within the camel-specific lineage (chapter 6), suggesting these elements might play a more important role in the evolution of this lineage compared to the broader GBS population. As PICI can be major drivers of bacterial pathogenicity, such as for *S. aureus* in which they carry the toxic shock syndrome toxin-1 (TSST-1) (Lindsay et al., 1998; Todd et al., 1978), it would be important to follow up on these bioinformatic findings with laboratory work aimed at characterising the functional role of genes encoded by these PICI, particularly for the camel-specific subpopulation.

I also identified for the first time plasmids in GBS of animal origin, in particular from dairy cattle. Plasmids are rarely reported in human GBS (Richards et al., 2019) and they are not usually considered important drivers of GBS evolution. AMR plasmids described in human GBS to date (Sendi et al., 2016; Compain et al., 2014; DiPersio et al., 2011; Herbert et al., 2005) do not encode genes for resistance to β-lactams[2], which remain the antimicrobials of first choice for the treatment of GBS infections in people (unless hypersensitive to penicillin) and dairy cattle. Different from human isolates, I showed that plasmids are present in one third of bovine isolates, which possibly indicates a more important role in GBS evolution than previously thought. GBS plasmids do not carry AMR genes in this host species.

[2]Resistance to β-lactams in GBS is mainly associated with structural changes in the *pbp* gene (Hayes et al., 2020; C. Li et al., 2020; van der Linden et al., 2020), rather than with β-lactamase genes carried on plasmids.

Interestingly, two novel plasmids I describe showed high sequence similarity with plasmids from human-pathogenic streptococci: GAS and *Streptococcus dysgalactiae* subsp. *equisimilis*, which co-exist with GBS in the human oropharynx (Davies et al., 2005). This illustrates HGT of MGE between streptococcal species that share the same ecological niche, as discussed in more detail in the next section, which involves complex dynamics of competition between isolates, as indicated for example by the presence of a plasmid-encoded anti-GAS bacteriocin in GBS (chapter 3). The main limitation of my work is the fact that these are bioinformatic results which do not provide information about the functional importance of these plasmids in GBS, particularly for the one that is widespread in the bovine-specific lineage CC61/67. As an example, it would be interesting to perform laboratory experiments to test the performance of GBS isolates carrying the bacteriocin plasmid (detected in a sequence type (ST) 314 bovine isolate) and isolates lacking it when in competition with GAS. Also, the possibility and the frequency of exchange of these plasmids between different GBS lineages (from the same and different host species), as well as between GBS and other streptococci should be evaluated. In addition, sequencing more isolates from both human and animal sources with long-read sequencing technologies could lead to the discovery of new plasmids and could expand our knowledge and understanding of the role of these MGE in GBS. In fact, as for one of the novel plasmids I detected in bovine GBS that was present in a high proportion of my CC61/67 genomes and that showed hits in published Illumina data in which it had not been reported (as described in chapter 3), other widespread plasmids may have gone undetected so far.

Likewise, the prevalence and the possible functional relevance of prophage GBS7 in ST283, which is integrated at the 5' end of the hyaluronate lyase (*hylB*) gene, may have been overlooked so far due to limitations of short-read sequencing and assembly. This prophage was detected in the majority of closed genomes from Singaporean isolates (n=6/9, chapter 2) and it is usually not detected in draft assemblies generated from Illumina data (n=0 in ST283 genomes from chapter 4, data not shown). This is likely due to the fact that GBS1, which in itself is highly associated with ST283 (n=64/77 in ST283 genomes from chapter 4, data not shown), has an identical integrase gene (GBS*Int*1) to that of GBS7, and since GBS7 is uniquely found in genomes in which GBS1 is present, assembly issues with mapping

of the integrase reads to the same site as GBS1 may result in false negative results from short-read sequencing. Therefore, the application of long-read sequencing to ST283 isolates might help clarify the extent of the presence of this prophage in this lineage, which might be exerting a functional role in the expression on the *hylB* gene. *HylB* is a virulence gene that is able to degrade certain components of the extracellular matrix and is believed to contribute significantly to invasion (Herbert et al., 2004a; Tettelin et al., 2002; Glaser et al., 2002), and the insertion of GBS7 at its 5' end might be upregulating its expression. Another possibility is that GBS7 might be downregulating the expression of *hylB* if inserted in its promoter region, also potentially contributing to invasion; in fact, the disruption of *hylB* due to the integration within the gene of IS*1458* has been linked with high risk of invasive disease in neonates (Domelier et al., 2006). The expression of *hylB* in ST283 genomes from Singapore and other countries could be assessed with RT (reverse transcription)-PCR and RNA-sequencing, in isolates with prophage GBS7 and in knockout mutants (ΔGBS7). In addition, the pathogenicity of GBS7+ and of ΔGBS7 isolates could be assessed in infection models such as the one developed by Six et al., 2019, for GBS in *Galleria mellonella* larvae.

7.3 GBS ecotypes share genetic content with other streptococcal species based on the niche they inhabit

As described in the previous section for novel plasmids, GBS shares accessory genome content with other gram-positive and gram-negative bacterial species, and particularly with other streptococci. This phenomenon has been linked to shared ecological niches, such as the human oropharynx (Franken et al., 2001) and the bovine mammary gland (Richards et al., 2011), that represent interfaces for HGT. Therefore, the concepts of genetic and ecological species (Cohan, 2002), as described for *Thermotoga* spp. (Nesbø et al., 2006), also apply to streptococci. Genetic species (a.k.a. biological species) are characterised by intra-species sequence similarity due to long-term within-species homologous recombination, but genetic barriers largely prevent them from exchanging chromosomal material with other species (Hanage et al., 2005; Dykhuizen & Green, 1991). Ecological species (a.k.a. ecotypes)

are groups of bacteria that occupy the same ecological niche and use similar ecological resources, thanks to common genetic traits, to out-compete isolates that are not adapted to that ecological niche (Cohan, 2002). Cohan proposed that each named genetic species contains numerous ecotypes, 'each with the fundamental properties of a [genetic] species'. In this optic, the niche environment that is populated by adapted ecotypes represents a source of useful accessory genes, which are often carried and transferred between bacterial cells by MGE, that can help isolates in adapting to a new environment and succeed in the niche itself.

One of the best examples in GBS is that of the bovine mammary gland environment with the Lac.2 being commonly carried by mastitis-causing pathogens (Holt et al., 2015; Richards et al., 2011), which together form the bovine mammary gland ecotype. Lac.2+ isolates of these bacterial species have an adaptive advantage over competing Lac.2- isolates, as the former are able to ferment lactose, one of the most abundant substrates available in the udder. The acquisition of metabolic pathways in response to nutrient availability is one of the major drivers of adaptation to specific niches (S. K. Sheppard et al., 2018), and, therefore, to the creation of ecotypes.

For bacteria that form the human ecotype, the most important example of shared genetic content is the *scpB-lmb* transposon, an MGE that is known to be highly prevalent among human GBS isolates (Morach et al., 2018; Franken et al., 2001). The *scpB-lmb* transposon is found among other human pathogenic streptococci, such as GAS, *S. dysgalactiae* subsp. *equisimilis* and *Streptococcus canis* (Fig. 7.1) and it has been associated with their ability to colonise or infect the human host (Franken et al., 2001). Other examples of HGT between GBS, GAS and *S. dysgalactiae* subsp. *equisimilis* within the human niche are bacteriophages and transposons, as documented by Davies et al., 2005. The human oropharynx is regarded as the most likely site of genetic exchange between these species (Franken et al., 2001).

In addition to the bovine and the human ecological niches, I found evidence of shared genetic content between GBS lineages associated with fish (locus 3 and IS*Stin5*) and another important warm-water fish pathogen, *Streptococcus iniae* (Fig. 7.1). To my knowledge, no previous studies had investigated the existence of shared accessory genome content between GBS and other bacterial species that populate the aquatic niche, with the exception of

Delannoy et al., 2016, who had found homologs for only two genes of locus 3 in *S. iniae*. The results of my analyses show that GBS fish-associated and fish-specific accessory genome content, including locus 3 and IS*Stin5*, is shared with *S. iniae*. These two species together are part of the aquatic ecotype.

A fourth ecological niche is that of the camel host. In particular, within camels differ-

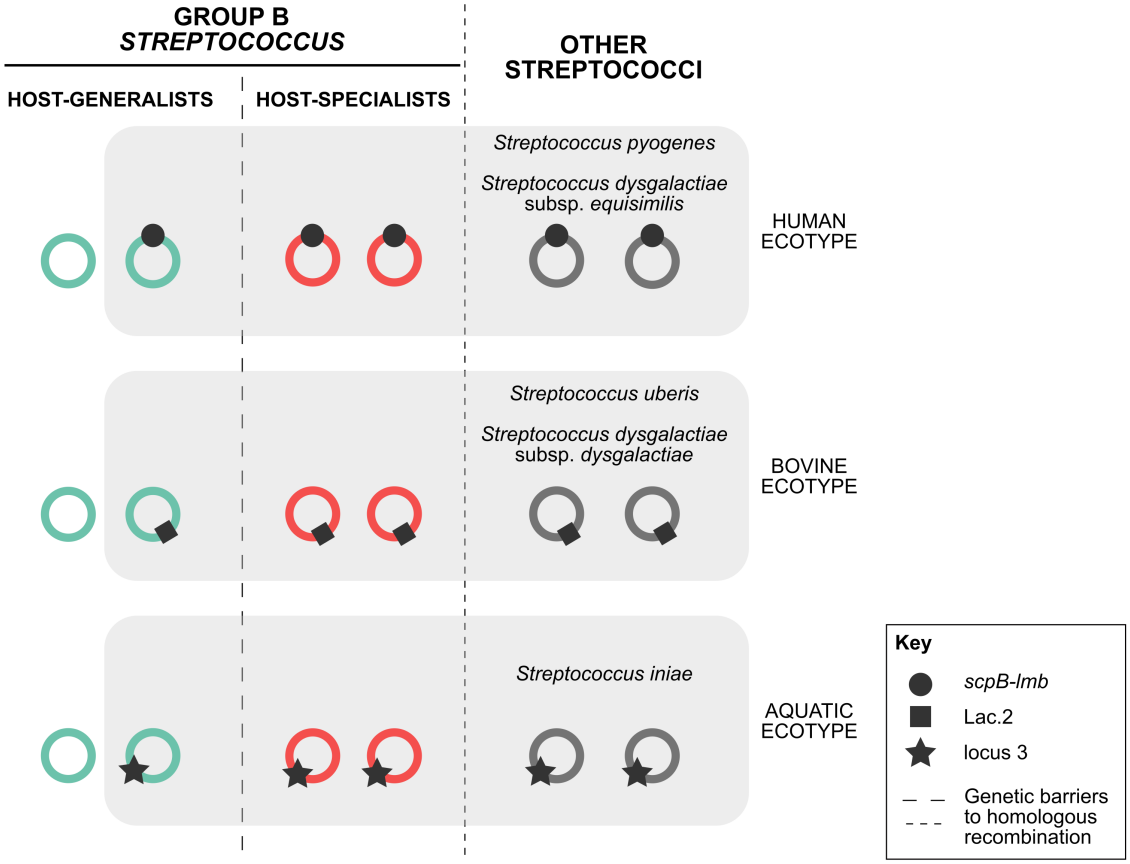


Figure 7.1: A simplified diagram illustrating genetic and ecological species concepts applied to streptococci. Grey boxes indicate ecotypes and comprise isolates from their respective hosts, whereas isolates outside the grey boxes originate from other hosts. Group B *Streptococcus* (GBS) is a genetically distinct species from other streptococci, but it shares accessory genome content that is important for adaptation to three major host groups with other bacteria in this genus. The *scpB-lmb* transposon is associated with the human ecotype, and it is shared among GBS, *Streptococcus pyogenes* (group A *Streptococcus*, GAS), and *Streptococcus dysgalactiae* subsp. *equisimilis*. The Lac.2 (lactose operon) is found in mastitis-causing streptococci (bovine ecotype), such as GBS, *Streptococcus dysgalactiae* subsp. *dysgalactiae* and *Streptococcus uberis*. In the aquatic niche, GBS shares locus 3 with *Streptococcis iniae* (aquatic ecotype).

ent GBS lineages show predilection for two niches: the nasopharynx (carriage strains e.g. ST612, ST615) and the mammary gland (mastitis-causing strains e.g. ST616). I detected Lac.2 in the majority of GBS isolates from mastitis cases in camels, although this was not as strongly associated with mastitis as in dairy cattle (chapters 3, 5, 6). This suggests that this gene cluster is beneficial for survival in the camel udder, albeit not necessary. Bacterial competition assays comparing the success in a lactose-rich environment (or specifically in camel milk) of Lac.2+ isolates and their knockout mutants (ΔLac.2) could be carried out to better understand the importance of Lac.2 in camel strains associated with milk (ST616). Moreover, camel milk isolates share accessory genome content with GAS and *Streptococcus suis* (chapter 6); however, it is hard to connect these genetic exchanges to a physical interface in which camel GBS from milk, GAS and *S. suis* may come into contact. As GBS from camels forms a monophyletic clade with deep branching (chapter 4, Fig. 4.1), it is highly likely that it evolved a long time ago, and the presence of shared accessory genome content with other streptococci may signal ancestral genetic exchanges, rather than contemporary ones. The analysis of the acquisition and loss events of these accessory genes on a time-scaled phylogeny, as described in the next section, could help our understanding of the evolutionary history of camel GBS.

7.4 A limited number of accessory genes is associated with specific ecotypes in GBS

Large-scale GWAS were carried out to detect host/niche-associated accessory genome content. To my knowledge, this is the first time this approach was applied to GBS with such an aim. One previous GWAS had analysed a broad dataset of GBS genome sequences to identify genes associated with clonal complexes (CC) of particular interest, such as the hypervirulent CC17 (Gori et al., 2020). However, the dataset used by the authors had a high number of human genomes compared to animal genomes (n=1901 and n=89, respectively, vs n=420 and n=404 in my dataset, excluding food market fish isolates), and these only derived from five countries. In addition, for one of these countries, genomes originated from a study that focused specifically on ST1 (Flores et al., 2015), introducing a sampling bias. Hence, this dataset cannot be considered a good representation of the global GBS population, and

this likely influenced the authors' findings. In my work, a particular focus was dedicated to dataset curation in order to select a representative sample of the sequenced isolates available to date and to reduce sampling bias, as described in chapter 4.

The total number of highly significant host-associated genes for the three major host groups (human, bovine, fish) was limited. This differs from what has been described in *S. aureus*, for which the molecular basis of its adaptation has been linked to numerous MGE as well as to chromosomal gene clusters (Richardson et al., 2018): human-associated MGE comprised several prophages, PICI, plasmids and transposons (Richardson et al., 2018); ruminant-associated MGE included PICI and bacteriophages (Richardson et al., 2018; Guinane et al., 2010; Viana et al., 2010); bird-associated MGE comprised plasmids and bacteriophages (Richardson et al., 2018; Lowder et al., 2009), as was the case for pigs (Richardson et al., 2018). The most significant genes I detected in GBS had all been described before as host-associated in GBS (see previous section) (Delannoy et al., 2016; Richards et al., 2011; Franken et al., 2001), which is remarkable considering that the genome-wide comparison was run on a large dataset that aimed at being as representative as possible of the diversity of the GBS population. In addition to these host-associated elements, I investigated for the first time genes associated with the camel host, filling a knowledge gap in this area. I detected four major genomic islands (GEI), which mostly encoded genes for surface proteins and sugar metabolism. The latter, i.e. the ability of isolates to utilise particular sugar molecules, appears to be a major driver of host-adaptation/tissue tropism in animal GBS (e.g. lactose fermentation in dairy cattle, galactose fermentation in fish, as described in chapter 5). Within the camel host, I also detected GEI and single genes for adhesion and invasion that are associated with the camel mammary gland niche (chapter 6). Further work, particularly competition essays between wild type and knockout mutants, is needed to investigate the possible functional role of these GEI in adaptation to the camel host and to its mammary gland.

Interestingly, when considering the accessory genome in its entirety, this content does not form separate clusters based on host species independent of lineages. Rather, the whole repertoire of accessory genes of an isolate seems to depend first and foremost on its lineage, and only secondarily on host species, which is observable in the network analysis of acces-

sory genes in chapter 4. This is an important observation, especially when coupled with the GWAS results: these findings together suggest that a limited number of accessory genes could potentially exert a great impact on host adaptation of GBS isolates. This hypothesis is supported by preliminary results of functional experiments carried out by Dr John Bell at Moredun Research Institute, under the supervision of Prof Ruth Zadoks (Ruth Zadoks, personal communication). Dr Bell created various knockout mutants of locus 3 and its segments on GBS isolates from fish. Subsequent experiments on live fish showed that complete locus 3 knockout mutants were non-pathogenic (ST7) or attenuated (ST283). Therefore, locus 3 appears to be necessary or important for GBS to cause invasive infections in fish, which is also supported by bioinformatic findings: locus 3 is found in all fish isolates and its absence from an isolate perfectly predicts negativity for the fish phenotype (100% negative predictive value, NPV). This differs from e.g. Lac.2 in camel mastitis isolates, in which this MGE might provide an adaptive advantage to the udder niche, but it does not seem to be necessary to establish infection in the camel mammary gland, as mentioned in the previous section. These observations highlight how the wider concept of host-adaptation includes different nuances. On the one hand certain genetic assets might be essential for a bacterium to survive or to cause disease in a specific host; on the other hand some accessory genes might not be necessary, but could offer an adaptive advantage over competing strains. In the case of locus 3, it would be interesting to test whether isolates of human or bovine origin can successfully adapt and cause disease in fish when complemented with locus 3, as it is thought to have happened for ST283 isolates (i.e. human isolate acquired locus 3, then jumped into fish, where they were amplified, and transmitted back to humans) (Barkham et al., 2019).

The degree to which a specific gene cluster is necessary for the success of an isolate in a certain host species also affects the possibility for it to be used as a predictor of the host of origin. Continuing with locus 3 as an example, this gene cluster is not very specific for fish, as it is found in a good proportion of genomes from other hosts (both in isolates from humans that were infected from raw fish contaminated with the fish-associated lineage CC283 and in isolates from lineages unrelated to fish such as the camel lineage CC609 and the bovine lineage CC103/314, chapter 5). Therefore locus 3 is not a good positive predictor for fish origin (in contrast to fish-specific loci like locus 5 and IS*Stin5*). However, if a genome lacks

locus 3, a fish origin of the isolate can be excluded. Differently, Lac.2 was found to be highly specific for the bovine host, although not essential, as a few bovine isolates are Lac.2- (false negative genomes due to limitations of short-read sequencing and assembly cannot be ruled out). When camel GBS from milk, in particular ST616, is not considered, Lac.2 is found in few non-bovine isolates and it appears to be a good predictor of bovine origin. However, the high specificity of Lac.2 genes for the bovine host (97.2-99.7% for *lacE* and *lacG* variants), as described in chapter 5, would be lower if genomes from camel milk were included in the GWAS analyses (future work). The *scpB-lmb* transposon falls in the middle of the spectrum, with certain *scpB* alleles being very good positive predictors of a human origin compared to other ones (chapter 5), and *lmb1* being a moderately good negative predictor (i.e. if a genome does not code for *lmb1*, it is highly likely that it does not derive from a human host).

In chapter 3, I use the tetracycline resistance (TcR) gene *tet*(M) as a predictor of human origin of newly emerged lineages in dairy cattle because extensive usage on tetracycline in the 1960s it is thought to have selected a few TcR clones that spread globally in the human population (Da Cunha et al., 2014). The ICE carrying *tet*(M) tends to be retained even in the absence of selective pressure (e.g. TcR in dairy cattle in Sweden, as described in chapter 3, a production system in which there is a strict regulation of antibiotic usage, and in particular tetracycline is not commonly used to treat the main diseases in cattle), with the exception of recent CC283 isolates from Thailand (Barkham et al., 2019), and this makes *tet*(M) a historical (long-term) predictor of human origin. Of the four *tet*(M) alleles detected by roary, three have a high specificity for humans (92-99%, data not shown), but they are still moderate predictors of this phenotype (65-86% positive predictive value, PPV). This is explained by the fact that nowadays a significant proportion of animal GBS isolates carry the ICE responsible for TcR, Tn*916* (Crestani et al., 2021; Barkham et al., 2019). However, the time-frame of emergence of TcR in GBS in the various host species must also be considered. At the same time in which human TcR clones underwent the selective pressure of tetracycline (1960s), bovine isolates remained *tet*(M)-, as illustrated by my results of historical GBS isolates from dairy cattle in Sweden. Therefore, this background knowledge on the timing of acquisition and spread of TcR in the different hosts, together with knowledge of which lineages are shared between humans and cattle, supports the hypothesis of a human origin of newly

emerged lineages in dairy cattle in Northern Europe. On the other hand, it would be important to study the evolutionary history of GBS in terms of acquisition of these genetic markers (*scpB-lmb*, Lac.2, locus 3, *tet*(M)) and how these map to host-jump events on time-scaled phylogenies. Acquisition of these MGE may have determined the subsequent emergence of host-adapted lineages, as described in *S. aureus* for human-to-livestock host-jumps corresponding to the period of animal domestication (Richardson et al., 2018; Weinert et al., 2012). Similar to the time-scaled phylogenetic analyses carried out by Barkham et al., 2019, who dated the period of emergence of CC283 around 1985 (95% HPD 1980–1990), analyses of a representative sample of the global GBS population could help in tracing the emergence of the various GBS lineages, as well as their host of origin (as was done for *S. aureus* for the human host (Richardson et al., 2018)), subsequent host-to-host jumps and their correlation with the acquisition and loss of relevant genetic material. At present, I have run preliminary data analysis with TEMPEST v1.5.3 (Rambaut et al., 2016) to detect a temporal signal. This was followed by BEAST v1.10.4 (Suchard et al., 2018) analyses of a subset of the dataset described in chapter 4, based on a representative sampling of BAPS population, serotype, host of origin and year of isolation. On this subset, I ran combinations of nucleotide substitution models (GTR+G4, HKY), clock models (uncorrelated relaxed clock, strict clock) and population size (constant, exponential growth and Bayesian skyline) to select the best model. The selected model was GTR+G4 with a strict clock and Bayesian skyline population. Unfortunately, due to time constraints and modulated access to computational resources, analysis of the full dataset is still ongoing and results could not be included in my thesis.

The timing of acquisition of MGE which are considered important for host adaptation can not only be studied in terms of the long-term evolutionary history of GBS (e.g. impact of animal domestication after the Neolithic period as described in *S. aureus* (Richardson et al., 2018)), for which acquisition of these elements likely determined host-jumps and host-adapted lineages, but it could be assessed in the short-term as well. As an example, the speed at which Lac.2 is acquired by isolates that are Lac.2- (both within GBS and between GBS and other mastitis-causing streptococci) could be assessed with laboratory experiments similar to those described by Prof Jan Roelof van der Meer at the Microbiology Society Annual Conference 2021 for ICE*clc* in *Pseudomonas* spp. (Carraro et al., 2020). Briefly,

fluorescent tags such as mCherry and GFP (Green Fluorescent Protein) can be used to mark Lac.2+ and Lac.2- cells, and these can be observed under a fluorescent microscope to monitor the timing of transfer of Lac.2 in a lactose-rich environment, such as that Lac.2- cells will change fluorescence after acquiring Lac.2. This could help better understand whether such genetic exchanges are common within the bovine mammary gland niche, and the speed at which they occur.

7.5 Host-specificity is associated with different levels of genome plasticity among GBS lineages

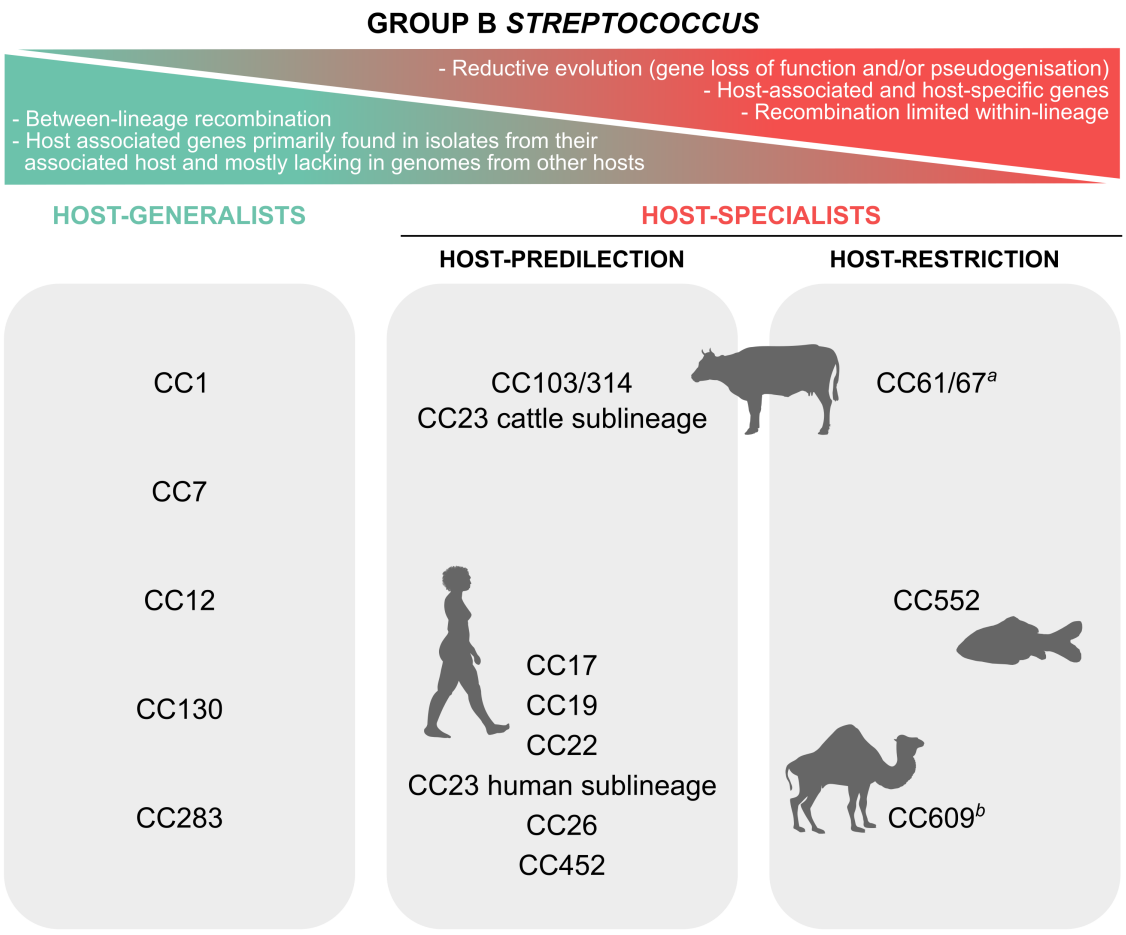
The host range of a GBS isolate does not only depend on the carriage of the three host-associated elements described in the previous section. Other genetic phenomena can impact on the host range, particularly at the lineage level, such as single nucleotide polymorphisms (SNP), as described in rabbits for *S. aureus* (Viana et al., 2015), and reductive evolution/pseudogenisation, as in *S. enterica* subsp. *enterica* serovar Gallinarum and Pullorum in poultry (Langridge et al., 2015), *S. enterica* subsp. *enterica* serovar Typhi in humans (Parkhill et al., 2001), *Mycobacterium leprae* in humans (Cole et al., 2001) and *S. aureus* lineage CC133 in ruminants (Guinane et al., 2010). Certain GBS lineages are known to be highly specific for a particular host, such as CC61/67 for dairy cattle, whilst others are more generalist and commonly detected in multiple hosts, such as CC1 in humans and cattle. In addition, the degree of lineage host-association can vary (Fig. 7.2): from host-predilection, as an example for CC103/314, which is predominantly described in cattle (Europe, South America, Asia) (Sørensen et al., 2019; Cobo-Ángel et al., 2019; Y. Yang et al., 2013) but which has been isolated in rare cases from humans, guinea pigs and cats (Tab. C.2) (Boonyayatra et al., 2020; Sørensen et al., 2019; Hsu et al., 2019; Wu et al., 2019), to host-restriction, as is the case for CC552 in cold-blooded species (Richards et al., 2019; Rosinski-Chupin et al., 2013). This latter lineage is limited to some poikilotherm species (primarily fish, with a few records also from frogs (Rosinski-Chupin et al., 2013)): there are no records of CC552 isolates from other host species, including warm-blooded (e.g. humans, cattle, camels, sea mammals) and cold-blooded species (e.g. crocodiles), and this has been attributed to extended gene loss of function and pseudogenisation with genome reduction (Richards et al., 2019; Rosinski-

Chupin et al., 2013). These modifications likely 'trapped' CC552 in the aquatic niche. Similar to CC552 in fish GBS, a higher level of pseudogenisation has been observed in *S. aureus* among isolates from ruminants compared to other host species, suggesting that this niche has stronger selection for gene loss of function compared to other niches (Richardson et al., 2018; Guinane et al., 2010). In bovine GBS, a similar mechanism of host restriction has been demonstrated for the bovine-specific lineage CC61/67, but in this case the pseudogenisation has been described uniquely for the genes of the capsular operon (Almeida et al., 2016). The capsule is an important virulence factor in humans, and its pseudogenisation would therefore prevent these strains from successfully colonising and causing disease in humans. However, this mechanism does not seem to be as restrictive as the genome-wide reduction of CC552, since three recent cases of CC61/67 have been reported in humans (L. Li et al., 2018). For other host-specialist[3] lineages, genetic explanations for host-predilection other than the presence/absence of host-associated accessory genes, as shown by my work, are still unclear.

Prior to this PhD thesis, no previous studies had investigated the GBS population structure identifying genomic differences between host-specialist and host-generalist lineages. In contrast with previous studies that had analysed the phylogenetic structure of the GBS population (Barkham et al., 2019; Richards et al., 2019), the dataset I used was purposefully curated to reduce the bias toward human invasive isolates as much as possible given the available data, as described above. This was done to achieve a more realistic representation of the global GBS population, minimising the skew of the data towards particular ST and serotypes that had been selected for sequencing in certain studies (Teatero, McGeer, et al., 2015; Teatero, Athey, et al., 2015; Flores et al., 2015).

The GBS population structure was analysed based on core genome content (maximum-likelihood phylogeny and fastbaps clustering) and on its set of accessory genes (distance network). Both analyses showed a separation between host-specialist and host-generalist lineages, and the accessory gene content was particularly good at discriminating between the

[3]In this thesis, the terminology 'host-specialist' is used to indicate both host-associated lineages that are restricted to one host (i.e. uniquely found in one host group) and host-associated lineages that show predilection for one host (i.e. predominantly isolated from one host group, but that can occasionally occur in other hosts). Host-generalist lineages comprise strains that commonly affect multiple host groups.



[a]Pseudogenisation in CC61/67 has been described uniquely for the capsular operon
[b]Pseudogenisation and gene decay has not been investigated in CC609, but this lineage has only been isolated from camels so far

Figure 7.2: Diagram illustrating host-specificity levels in group B *Streptococcus* (GBS) in three lineage categories (host-generalists, host-specialists with predilection or restriction). Host generalist lineages show extended between-lineage homologous recombination (chapter 4) and the host associated accessory genes (*scpB-lmb*, Lac.2, locus 3, chapter 5) are primarily found in isolates from their associated host, while they are lacking from isolates from other hosts (e.g. Lac.2 tends to be present in most isolates from cattle and absent from most isolates from humans). Host-specialist lineages that show host-restriction are associated with reductive evolution (e.g. gene loss of function and genome reduction as in CC552, pseudogenisation of the capsular operon in CC61/67), they carry host-associated genes (e.g. CC61/67 all carry Lac.2, CC552 all carry locus 3) and some carry host-specific genes (e.g. locus 1-2 and 4-8 in CC552), and they either show absence of recombination (CC552) or recombination limited within the lineage (CC61/67). Host-specialists that show host-predilection are primarily associated with one host and they show all characteristics of the host-restricted lineages except for genome reduction and pseudogenisation.

two groups. In addition, host-generalists shared a more similar repertoire of accessory genes among themselves, whilst host-specialists showed very diverging accessory gene sets within the group (Fig. 7.1). This finding is also reflected in the homologous recombination observed in the core genes: host-generalists show a higher level of recombination within the group across lineages, compared to host-specialists in which recombination is low and mostly limited within lineages (chapter 4, Fig. 4.3). These are indications that host-specialist lineages tend to evolve independently of each other, whilst genetic exchanges in host-generalists are common across lineages. As described previously, a high level of homologous recombination in the core genes can be a reflection of a higher number and variety of MGE that transfer segments of the chromosome, such as prophages and ICE (Chen et al., 2018; Everitt et al., 2014). Therefore, the interpretation of results from homologous recombination analyses and accessory genome content/MGE acquisition should not be done independently but they should be considered as part of the same complex scenario of genome dynamics.

The absence of recombination between host-specialist and generalist lineages is an indication of the existence of barriers to recombination between these GBS subgroups (Fig 7.1). In *C. jejuni*, an inhabitant of the gastrointestinal tract of multiple hosts whose population also comprises host-generalist and specialist lineages, the absence of recombination between a cattle and a chicken specialist lineage has been shown (S. K. Sheppard et al., 2014), similar to what I observed in GBS host-specialists. However, the authors found that these lineages could recombine with two host-generalist lineages, and that these latter two did not share recombination, as opposed to what I describe in GBS. In *C. jejuni*, the absence of recombination between host-generalists within the same niche (i.e. cattle or chicken) has been attributed to cryptic niche structure that limits opportunities for genetic exchange within the host in nature, rather than to genetic barriers (S. K. Sheppard et al., 2014). In GBS, the fact that extended recombination is shared between host-generalists regardless of their host of origin, and that host-specialists either do not show recombination or show recombination limited within-lineage, is indicative of genetic barriers to recombination; however, the exact mechanisms underpinning this phenomenon are still unknown. In addition, it is currently unclear what causes host-generalists to be more recombinogenic than host-specialists. Although GBS is not known to be a naturally transformable bacterium, it carries genes for

competence[4] (e.g. *comX*, chapter 2), and different GBS lineages might be expressing higher or lower levels of competence, which would influence their recombination rates. Some factors that have been shown to influence genetic competence in multiple bacterial species are carriage of certain bacteriophages (Brooks et al., 2020; Rabinovich et al., 2012; Loessner et al., 2000) and particular environmental/niche conditions (Solomon & Grossman, 1996) (chapter 4). Moreover, as explained in the first section of this chapter, RMS could be playing a role in the ability of each lineage to retain alien DNA. In particular, I detected a difference in type I RMS between host-generalists, which only code for I M and I S genes but which lack the restriction enzyme (R) that is responsible for cleaving non-self DNA, and host specialists, which either lacked type I RMS or coded for a functional type I RMS, including the I R enzyme (as explained in chapter 4). Laboratory experiments are currently being performed at the Roslin Institute, Edinburgh, UK, by Dr Nicola Linskey and Dr Connor Bowen to test the electroporation efficiency[5] of plasmid pDL278 on isolates with different type I RMS profiles. In addition, I have run a preliminary GWAS with scoary on the generalists vs specialists phenotypes. The highest scoring genes for the specialists are mostly negatively associated with this phenotype, and they largely comprise MGE-associated genes (e.g. phage integrases, Cro/Cl transcriptional regulators, type I M gene) and genes with metabolic functions (e.g. various sugar metabolic pathways), whereas these were significantly positively associated with the generalist phenotype.

All of these findings taken together suggest that host-generalist lineages have more plastic genomes and might have a superior ability to uptake and retain foreign DNA compared to host-specialists, from which they differ considerably in terms of recombinogenic potential, and that these two groups largely evolve independently of each other. This has implications for disease management and public health, as specialist lineages generally represent a low threat outside of their preferred host, and elimination efforts can lead to successful outcomes, as shown for CC61/67 in Sweden (chapter 3). In contrast, generalist lineages represent a higher threat for two main reasons: i) they can undermine control programmes due to the sympatric presence of other known hosts that can represent a source of introduction,

[4]Genetic competence is the ability of a bacterium to uptake external DNA.

[5]Electroporation is the usage of high-voltage electric shocks to introduce DNA into bacterial cells; this technique is commonly used as part of cloning protocols to introduce plasmids into recipient bacterial isolates.

such as hypothesised for human CC1 in the bovine population in Sweden (chapter 3); and ii) they have a high potential to jump and adapt to new host species, as it is thought to have occurred for ST283 after its emergence in the mid-1980s (likely human-to-fish jump with the acquisition of locus 3) (Barkham et al., 2019). Control of GBS therefore requires ongoing monitoring of pathogen diversity across host species (One Health approach) and adaptive management in response to changing selective pressures and emergence of new strains.

7.6 Final thoughts

Results from large-scale comparative genomic studies, in particular those which focus on understanding between-host genetic differences and genome dynamics of host-adaptation, are major assets to our understanding of multi-host pathogens such as GBS. In this PhD thesis, I provide significant contributions to the field, in particular:

- An extensive characterisation of MGE in GBS across host groups (human, bovine, fish and camel), in particular prophages and PICI types (with the creation of a typing and detection method), ICE, and plasmids (with the first description of plasmids in animal GBS) (chapter 2, 3, 6). I also described other GEI and their possible correlation with pathogenicity (chapter 2, 6) and host-adaptation (chapter 6);
- A large-scale analysis of the global GBS population structure comprising human and animal genomes and an evaluation of genomic properties of its various lineages, with a particular focus on genome plasticity (chapter 3) of host-specialist vs host-generalist lineages (chapter 4);
- Comparative genomic analyses of the ensemble of accessory genes (chapter 4) and large-scale identification of significantly host-associated genes and MGE using two GWAS methods (chapter 5, 6).

Moreover, in this chapter I propose how additional insights could be gained through further work in the form of several possible wet-lab experiments, which are based on relevant bioinformatic findings and which could help expand our understanding of the functional role of the mobilome in GBS host-adaptation. Overall, my findings represent valuable contributions to our understanding of GBS ecology.

GBS is a complex multi-host pathogen with numerous facets: it is a leading pathogen of human neonates across the globe (Seale et al., 2017), an emerging pathogen in human adults with and without underlying clinical conditions (Chaiwarith et al., 2011; Lambertsen et al., 2010; Skoff et al., 2009; High et al., 2005), as well as a highly virulent emerging foodborne pathogen in Asia (Barkham et al., 2019; Ong et al., 2018; Kalimuddin et al., 2017). GBS is also a well-established threat to food security and possibly to food safety for animal production in dairy cattle (milk), particularly in countries with less-developed dairy industries (e.g. South America). Moreover, it represents an emerging threat in aquaculture (fish) in warm-water countries of the Southern hemisphere (e.g. South America, Asia) and in camels (milk and meat) especially in pastoralist communities in the Horn of Africa. The absence of approved vaccines for both humans and animals (Kobayashi et al., 2019; Heath, 2016) forces medical doctors and veterinarians to uniquely rely on other preventive measures (e.g. screening of pregnant women in humans, biosecurity and increased hygiene protocols in farmed animals) and on the utilisation of antimicrobials for treatment of diseased individuals. In the event of a vaccine being approved and introduced, GBS from animals could pose a threat to human health due to possible capsular switching with capsular serotypes that are rare in humans and not covered by the vaccine (e.g. IV and VI in camels), as observed in *S. pneumoniae*. Finally, the possibility of amplification of human-pathogenic lineages (particularly host-generalists) in other hosts, which would act as reservoirs of GBS, represents a double threat: i) a threat to human health if human-to-animal host jumps, are followed by human reinfection, as shown for ST283 (Barkham et al., 2019); ii) a threat to animal health and food security due to reverse zoonotic transmission, with possible implications for the success of GBS control programs on farms. Continuing efforts in genomic research, through the funding of biobanking systems and of interdisciplinary sequencing projects targeting GBS from different hosts, are crucial to our understating of dynamics of host-adaptation and evolution of GBS.

Appendix A

Supporting information Chapter 2

A.1 Tables and figures

Table A.1: List of 69 group B *Streptococcus* genomes downloaded from NCBI included in dataset 1. Accession numbers, isolate names, host species, country of origin, sequence type (ST) and serotype are shown.

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
CP007482.1	138P	fish	USA	ST261	Ib
CP007565.1	138spar	fish	USA	ST261	Ib
CP011328.1	GX026	fish	China	ST261	Ib
CP015976.1	S25	fish	Brazil	ST552	Ib
CP018623.1	S13	fish	Brazil	ST552	Ib
CP019800.1	SA30	fish	Brazil	ST552	Ib
CP019801.1	SA33	fish	Brazil	ST552	Ib
CP019802.1	SA53	fish	Brazil	ST260	Ib
CP019803.1	SA73	fish	Brazil	ST260	Ib
CP019804.1	SA1	fish	Brazil	ST552	Ib
CP019805.1	SA5	fish	Brazil	ST552	Ib
CP019806.1	SA9	fish	Brazil	ST552	Ib
CP019807.1	SA16	fish	Brazil	ST552	Ib
CP019808.1	SA75	fish	Brazil	ST260	Ib
CP019809.1	SA79	fish	Brazil	ST552	Ib
CP019810.1	SA81	fish	Brazil	ST552	Ib
CP019811.1	SA85	fish	Brazil	ST927	Ib
CP019812.1	SA95	fish	Brazil	ST927	Ib
CP019813.1	SA97	fish	Brazil	ST927	Ib
CP019814.1	SA102	fish	Brazil	ST927	Ib
CP019815.1	SA132	fish	Brazil	ST260	Ib
CP019816.1	SA136	fish	Brazil	ST260	Ib
CP019817.1	SA159	fish	Brazil	Unknown ST	Ib

Table A.1 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
CP019818.1	SA184	fish	Brazil	ST552	Ib
CP019819.1	SA191	fish	Brazil	ST260	Ib
CP019820.1	SA195	fish	Brazil	ST552	Ib
CP019821.1	SA201	fish	Brazil	ST552	Ib
CP019822.1	SA209	fish	Brazil	ST552	Ib
CP019823.1	SA212	fish	Brazil	ST552	Ib
CP019824.1	SA218	fish	Brazil	ST927	Ib
CP019825.1	SA220	fish	Brazil	ST552	Ib
CP019826.1	SA245	fish	Brazil	ST260	Ib
CP019827.1	SA256	fish	Brazil	ST260	Ib
CP019828.1	SA289	fish	Brazil	ST260	Ib
CP019829.1	SA330	fish	Brazil	ST552	Ib
CP019830.1	SA333	fish	Brazil	ST552	Ib
CP019831.1	SA341	fish	Brazil	ST552	Ib
CP019832.1	SA343	fish	Brazil	ST552	Ib
CP019833.1	SA346	fish	Brazil	ST552	Ib
CP019834.1	SA374	fish	Brazil	ST552	Ib
CP019835.1	SA375	fish	Brazil	ST552	Ib
CP019836.1	SA623	fish	Brazil	ST552	Ib
CP019837.1	SA627	fish	Brazil	ST552	Ib
CP025026.1	SGEHI2015-113	fish	Singapore	ST283	III
CP025027.1	SGEHI2015-107	fish	Singapore	ST283	III
CP025028.1	SGEHI2015-95	fish	Singapore	ST283	III
CP025029.1	SGEHI2015-25	fish	Singapore	ST283	III
FO393392.1	2-22	fish	Israel	ST261	Ib
HF952106.1	ILRI112	camel	Kenya	ST617	VI
NZ_CP008813.1	C001	bovine	China	ST103	III
NZ_CP012503.1	NGBS357	human	Canada	ST297	V
NZ_CP013908.1	GBS-M002	human	Taiwan	ST1	VI
NZ_CP016391.1	FWL1402	frog	China	ST739	III
NZ_CP016501.1	WC1535	fish	China	ST7	Ia
NZ_CP019978.1	Sag37	human	China	ST12	Ib
NZ_CP019979.1	Sag158	human	China	ST19	III
NZ_CP020449.1	FDAARGOS_254	-	-	ST22	II
NZ_CP021862.1	CUGBS591	human	Hong Kong	ST12	Ib
NZ_CP021863.1	SG-M163	human	Singapore	ST283	III
NZ_CP021864.1	SG-M158	human	Singapore	ST283	III
NZ_CP021865.1	SG-M50	human	Singapore	ST283	III
NZ_CP021866.1	SG-M29	human	Singapore	ST283	III
NZ_CP021867.1	SG-M25	human	Singapore	ST19	III
NZ_CP021868.1	SG-M8	human	Singapore	ST1	VI
NZ_CP021869.1	SG-M6	human	Singapore	ST17	III
NZ_CP021870.1	SG-M4	human	Singapore	ST23	III
NZ_CP022537.1	874391	human	Japan	ST17	III

Table A.1 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
NZ_LT545678.1	SA111	bovine	Portugal	ST61	II
NZ_LT714196.1	BM110	human	USA	ST17	III

Table A.2: List of 503 group B *Streptococcus* genomes included in dataset 2 (Richards et al., 2019). Accession numbers, isolate names, host species, country of origin, sequence type (ST) with single locus variants (SLV) and serotype are shown.

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
AE009948	2603V/R	human	Italy	ST110	V
AEXT0100	FSL_S3-026	bovine	USA	ST67	III
AL732656	NEM316	-	-	ST23	III
ALQP0100	CCUG_37738	human	Sweden	ST19	III
BCNJ0100	JP17	fish	Thailand	ST283	III
CP000114	A909	human	USA	ST7	Ia
CP003810	GD201008-001	fish	China	ST7	Ia
CP003919	SA20	fish	Brazil	SLV257	Ib
CP006910	CNCTC 10/84	human	USA	ST26	V
CP007570	GBS1-NY	human	USA	ST22	II
CP007571	GBS2-NM	human	USA	ST22	II
CP007572	GBS6	human	USA	ST22	II
CP007631	NGBS061	human	Canada	ST459	IV
CP007632	NGBS572	human	Canada	ST452	IV
CP010319	GBS85147	human	Brazil	ST103	Ia
CP010867	SS1	human	USA	ST1	V
CP010874	CU_GBS_08	human	Hong Kong	ST283	III
CP010875	CU_GBS_98	human	Hong Kong	ST283	III
CP011325	HN016	fish	China	ST7	Ia
CP011326	YM001	fish	China	ST7	Ia
CP011327	GX064	fish	China	ST7	Ia
CP011329	H002	human	China	SLV736	III
CP012419	SG-M1	human	Singapore	ST283	III
CP012480	NGBS128	human	Canada	ST17	III
CP013202	GBS ST1	dog	USA	ST1	V
ERR048526	BE-NI-001	human	Belgium	ST23	Ia
ERR048527	BE-NI-005	human	Belgium	ST8	Ib
ERR048528	BE-NI-007	human	Belgium	ST315	III
ERR048529	BE-NI-008	human	Belgium	ST10	V
ERR048530	DK-NI-001	human	Denmark	ST17	III
ERR048531	DK-NI-002	human	Denmark	ST23	V
ERR048532	DK-NI-003	human	Denmark	ST523	Ib
ERR048534	DK-NI-005	human	Denmark	ST23	Ia
ERR048535	DK-NI-007	human	Denmark	ST17	III
ERR048536	DK-NI-008	human	Denmark	ST9	Ib
ERR048537	DK-NI-009	human	Denmark	ST28	II

Table A.2 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
ERR048538	DK-NI-010	human	Denmark	ST19	III
ERR048539	DK-NI-011	human	Denmark	ST19	III
ERR048540	DK-NI-012	human	Denmark	SLV1	V
ERR048541	DK-NI-013	human	Denmark	ST1	V
ERR048542	DK-NI-014	human	Denmark	ST88	Ia
ERR048543	DK-NI-015	human	Denmark	ST10	V
ERR048546	DK-NI-021	human	Denmark	ST17	III
ERR048547	DK-NI-022	human	Denmark	SLV1	V
ERR048548	BG-NI-001	human	Bulgaria	ST144	Ia
ERR048549	BG-NI-002	human	Bulgaria	ST17	III
ERR048550	BG-NI-003	human	Bulgaria	ST8	Ib
ERR048551	BG-NI-004	human	Bulgaria	ST23	Ia
ERR048552	BG-NI-005	human	Bulgaria	ST23	Ia
ERR048553	BG-NI-006	human	Bulgaria	ST12	Ib
ERR048554	BG-NI-007	human	Bulgaria	ST12	Ib
ERR048555	BG-NI-009	human	Bulgaria	ST28	II
ERR048556	BG-NI-010	human	Bulgaria	ST12	II
ERR048557	BG-NI-011	human	Bulgaria	SLV1	V
ERR048561	DE-NI-001	human	Germany	SLV1	V
ERR048562	DE-NI-003	human	Germany	ST10	V
ERR048563	DE-NI-004	human	Germany	ST10	V
ERR048564	DE-NI-006	human	Germany	ST144	Ia
ERR048567	DE-NI-012	human	Germany	ST88	Ia
ERR048568	DE-NI-0013	human	Germany	ST17	III
ERR048569	DE-NI-014	human	Germany	ST23	Ia
ERR048570	DE-NI-0017	human	Germany	SLV17	III
ERR048571	DE-NI-0019	human	Germany	ST17	III
ERR048572	DE-NI-022	human	Germany	ST387	V
ERR048573	DE-NI-032	human	Germany	ST17	III
ERR048574	DE-NI-033	human	Germany	ST23	Ia
ERR048575	DE-NI-036	human	Germany	SLV17	III
ERR048576	DE-NI-037	human	Germany	ST17	III
ERR048577	DE-NI-040	human	Germany	ST23	Ia
ERR048579	DE-NI-042	human	Germany	ST19	III
ERR048581	IT-NI-007	human	Italy	ST17	III
ERR048582	IT-NI-008	human	Italy	ST17	III
ERR048583	IT-NI-009	human	Italy	ST17	III
ERR048584	IT-NI-016	human	Italy	ST130	IX
ERR048586	IT-NI-019	human	Italy	ST467	III
ERR048587	IT-NI-020	human	Italy	SLV17	III
ERR048588	IT-NI-028	human	Italy	ST1	V
ERR048589	IT-NI-0031	human	Italy	SLV1	V
ERR048591	IT-NI-033	human	Italy	ST17	III
ERR048592	IT-NI-034	human	Italy	ST17	III

Table A.2 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
ERR048594	IT-NI-037	human	Italy	ST26	V
ERR048595	CZ-NI-001	human	Czech Republic	ST23	Ia
ERR048596	CZ-NI-002	human	Czech Republic	ST19	III
ERR048597	CZ-NI-003	human	Czech Republic	ST19	III
ERR048598	CZ-NI-004	human	Czech Republic	SLV1	V
ERR048599	CZ-NI-005	human	Czech Republic	ST23	Ia
ERR048600	CZ-NI-006	human	Czech Republic	ST1	V
ERR048601	CZ-NI-007	human	Czech Republic	ST255	Ib
ERR048602	CZ-NI-008	human	Czech Republic	ST1	V
ERR048603	CZ-NI-009	human	Czech Republic	ST1	V
ERR048605	CZ-NI-013	human	Czech Republic	ST1	V
ERR048606	CZ-NI-014	human	Czech Republic	ST479	II
ERR048607	CZ-NI-015	human	Czech Republic	ST1	V
ERR048608	CZ-NI-016	human	Czech Republic	ST459	IV
ERR048611	GB-NI-003	human	United Kingdom	ST17	III
ERR048612	GB-NI-004	human	United Kingdom	ST17	III
ERR048613	GB-NI-005	human	United Kingdom	SLV17	III
ERR048614	GB-NI-006	human	United Kingdom	ST23	Ia
ERR048615	GB-NI-007	human	United Kingdom	ST19	III
ERR048616	GB-NI-009	human	United Kingdom	ST1	V
ERR048617	GB-NI-010	human	United Kingdom	ST1	V
ERR048618	GB-NI-011	human	United Kingdom	ST19	III
ERR054970	B09PS	human	Australia	ST1	V
ERR054971	B15VD	human	Australia	ST17	III
ERR054972	B24VD	human	Australia	ST28	II
ERR054973	B37VS	human	Australia	ST335	III
ERR054974	B42VD	human	Australia	ST23	III
ERR054975	B50VD	human	Australia	ST23	Ia
ERR054976	B68VD	human	Australia	SLV19	III
ERR054982	RBH02	human	Australia	ST2	IV
ERR054983	RBH03	human	Australia	ST19	III
ERR054985	RBH05	human	Australia	ST1	V
ERR054987	RBH07	human	Australia	ST23	Ia
ERR054988	RBH08	human	Australia	ST23	Ia
ERR054990	RBH11	human	Australia	ST19	III
ERR054992	B96P	human	Australia	ST17	III
ERR054993	B41VS	human	Australia	ST652	II
ERR054994	B50VS	human	Australia	ST23	Ia
ERR054997	B96V	human	Australia	ST17	III
ERR829829	MRI Z1-116	bovine	Denmark	ST604	Ib
ERR829883	MRI Z2-084	human	Finland	ST28	II
HF952104	09mas018883	bovine	Sweden	ST1	V
HF952105	ILRI005	camel	Kenya	ST609	V
HG939456	COH1	human	USA	ST17	III

Table A.2 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
JPOV0100	NGS-ED-1000	human	United Kingdom	ST7	Ia
SRR1213207	NGBS024	human	Canada	ST459	IV
SRR1213208	NGBS046	human	Canada	ST459	IV
SRR1213210	NGBS058	human	Canada	ST459	IV
SRR1213213	NGBS070	human	Canada	ST459	IV
SRR1213214	NGBS100	human	Canada	ST452	IV
SRR1213215	NGBS122	human	Canada	ST452	IV
SRR1213216	NGBS146	human	Canada	ST459	IV
SRR1213217	NGBS151	human	Canada	ST3	IV
SRR1213218	NGBS187	human	Canada	ST452	IV
SRR1213219	NGBS191	human	Canada	ST459	IV
SRR1213220	NGBS197	human	Canada	ST452	IV
SRR1213221	NGBS199	human	Canada	ST459	IV
SRR1213223	NGBS290	human	Canada	ST459	IV
SRR1213224	NGBS314	human	Canada	ST452	IV
SRR1213226	NGBS379	human	Canada	ST3	IV
SRR1213227	NGBS400	human	Canada	SLV459	IV
SRR1213228	NGBS410	human	Canada	ST459	IV
SRR1213229	NGBS447	human	Canada	ST196	IV
SRR1213230	NGBS472	human	Canada	SLV196	IV
SRR1213231	NGBS493	human	Canada	ST459	IV
SRR1213232	NGBS507	human	Canada	ST459	IV
SRR1213233	NGBS521	human	Canada	ST459	IV
SRR1213234	NGBS525	human	Canada	ST459	IV
SRR1213235	NGBS528	human	Canada	ST459	IV
SRR1213236	NGBS556	human	Canada	ST452	IV
SRR1213238	NGBS572	human	Canada	ST452	IV
SRR1213239	NGBS588	human	Canada	ST682	IV
SRR1213240	NGBS597	human	Canada	ST452	IV
SRR1213241	NGBS598	human	Canada	ST452	IV
SRR1213242	NGBS612	human	Canada	ST452	IV
SRR1213243	NGBS615	human	Canada	ST459	IV
SRR1790740	NGBS010	human	Canada	ST1	V
SRR1790741	NGBS107	human	Canada	ST1	V
SRR1790742	NGBS110	human	Canada	ST1	V
SRR1790743	NGBS117	human	Canada	ST1	V
SRR1790749	NGBS180	human	Canada	ST1	V
SRR1790751	NGBS021	human	Canada	ST1	V
SRR1790752	NGBS210	human	Canada	ST1	V
SRR1790753	NGBS022	human	Canada	ST1	V
SRR1790758	NGBS246	human	Canada	ST1	V
SRR1790759	NGBS025	human	Canada	ST1	V
SRR1790760	NGBS267	human	Canada	ST1	V
SRR1790761	NGBS272	human	Canada	ST1	V

Table A.2 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
SRR1790765	NGBS028	human	Canada	ST1	V
SRR1790766	NGBS283	human	Canada	ST1	V
SRR1790767	NGBS287	human	Canada	ST1	V
SRR1790768	NGBS288	human	Canada	ST1	V
SRR1790769	NGBS298	human	Canada	ST1	V
SRR1790770	NGBS030	human	Canada	ST1	V
SRR1790771	NGBS303	human	Canada	ST1	V
SRR1790773	NGBS323	human	Canada	ST1	V
SRR1790775	NGBS330	human	Canada	ST1	V
SRR1790779	NGBS348	human	Canada	ST1	V
SRR1790780	NGBS035	human	Canada	ST1	V
SRR1790782	NGBS359	human	Canada	ST1	V
SRR1790783	NGBS360	human	Canada	ST1	V
SRR1790785	NGBS380	human	Canada	ST1	V
SRR1790786	NGBS381	human	Canada	ST1	V
SRR1790788	NGBS418	human	Canada	ST1	V
SRR1790789	NGBS425	human	Canada	ST1	V
SRR1790790	NGBS434	human	Canada	ST1	V
SRR1790792	NGBS444	human	Canada	ST1	V
SRR1790793	NGBS462	human	Canada	ST1	V
SRR1790794	NGBS492	human	Canada	ST1	V
SRR1790795	NGBS494	human	Canada	ST1	V
SRR1790796	NGBS497	human	Canada	ST1	V
SRR1790797	NGBS499	human	Canada	ST1	V
SRR1790799	NGBS519	human	Canada	ST1	V
SRR1790800	NGBS536	human	Canada	ST1	V
SRR1790801	NGBS054	human	Canada	ST1	V
SRR1790802	NGBS553	human	Canada	ST1	V
SRR1790803	NGBS558	human	Canada	ST1	V
SRR1790805	NGBS571	human	Canada	SLV1	V
SRR1790806	NGBS579	human	Canada	ST1	V
SRR1790807	NGBS580	human	Canada	ST1	V
SRR1790808	NGBS586	human	Canada	ST1	V
SRR1790809	NGBS604	human	Canada	ST1	V
SRR1790811	NGBS063	human	Canada	ST1	V
SRR1790812	NGBS630	human	Canada	ST1	V
SRR1790814	NGBS068	human	Canada	ST1	V
SRR1790815	NGBS008	human	Canada	ST1	V
SRR1790816	NGBS009	human	Canada	ST1	V
SRR1790817	NGBS092	human	Canada	ST1	V
SRR1790819	NGBS094	human	Canada	ST1	V
SRR1790820	NGBS099	human	Canada	ST1	V
SRR2062051	NGBS680	human	Canada	ST459	IV
SRR2062052	NGBS686	human	Canada	ST459	IV

Table A.2 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
SRR2062054	NGBS762	human	Canada	ST459	IV
SRR2062055	NGBS767	human	Canada	ST459	IV
SRR2062056	NGBS768	human	Canada	ST452	IV
SRR2062057	NGBS783	human	Canada	ST459	IV
SRR2062058	NGBS788	human	Canada	ST459	IV
SRR2062059	NGBS789	human	Canada	ST459	IV
SRR2062060	NGBS791	human	Canada	ST459	IV
SRR2062063	NGBS795	human	Canada	ST459	IV
SRR2062064	NGBS798	human	Canada	ST710	IV
SRR2062065	NGBS800	human	Canada	ST459	IV
SRR2062066	NGBS698	human	Canada	ST459	IV
SRR2062068	NGBS801	human	Canada	ST459	IV
SRR2062069	NGBS806	human	Canada	ST459	IV
SRR2062071	NGBS808	human	Canada	ST459	IV
SRR2062072	NGBS809	human	Canada	ST459	IV
SRR2062074	NGBS813	human	Canada	ST459	IV
SRR2062075	NGBS815	human	Canada	ST459	IV
SRR2062076	NGBS824	human	Canada	ST711	IV
SRR2062077	NGBS825	human	Canada	ST459	IV
SRR2062079	NGBS830	human	Canada	ST459	IV
SRR2062080	NGBS836	human	Canada	ST459	IV
SRR2062081	NGBS700	human	Canada	ST459	IV
SRR2062082	NGBS855	human	Canada	ST459	IV
SRR2062084	NGBS860	human	Canada	ST459	IV
SRR2062085	NGBS877	human	Canada	SLV459	IV
SRR2062086	NGBS899	human	Canada	SLV459	IV
SRR2062087	NGBS904	human	Canada	ST459	IV
SRR2062088	NGBS933	human	Canada	ST3	IV
SRR2062090	NGBS956	human	Canada	ST459	IV
SRR2062091	NGBS960	human	Canada	ST3	IV
SRR2062092	NGBS964	human	Canada	ST459	IV
SRR2062093	NGBS965	human	Canada	ST459	IV
SRR2062094	NGBS702	human	Canada	ST459	IV
SRR2062097	NGBS977	human	Canada	SLV459	IV
SRR2062099	NGBS979	human	Canada	ST459	IV
SRR2062100	NGBS984	human	Canada	ST459	IV
SRR2062101	NGBS991	human	Canada	ST196	IV
SRR2062103	NGBS996	human	Canada	ST459	IV
SRR2062104	NGBS1006	human	Canada	ST459	IV
SRR2062105	NGBS1009	human	Canada	ST459	IV
SRR2062107	NGBS1017	human	Canada	ST459	IV
SRR2062109	NGBS706	human	Canada	ST459	IV
SRR2062110	NGBS1021	human	Canada	ST459	IV
SRR2062112	NGBS1024	human	Canada	ST459	IV

Table A.2 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
SRR2062125	NGBS727	human	Canada	SLV452	IV
SRR2062139	NGBS736	human	Canada	ST459	IV
SRR2062154	NGBS737	human	Canada	ST459	IV
SRR2062160	NGBS741	human	Canada	ST459	IV
SRR2068019	NGBS1041	human	Canada	ST459	IV
SRR2068021	NGBS1043	human	Canada	ST459	IV
SRR2068022	NGBS1045	human	Canada	ST459	IV
SRR2068023	NGBS1046	human	Canada	ST459	IV
SRR2068024	NGBS1047	human	Canada	ST459	IV
SRR2068025	NGBS1048	human	Canada	ST459	IV
SRR2068026	NGBS1049	human	Canada	ST459	IV
SRR2068027	NGBS1050	human	Canada	ST452	IV
SRR2068028	NGBS1051	human	Canada	ST459	IV
SRR2068029	NGBS1052	human	Canada	ST459	IV
SRR2068031	NGBS1054	human	Canada	ST459	IV
SRR2068032	NGBS1056	human	Canada	ST459	IV
SRR2068033	NGBS1058	human	Canada	ST459	IV
SRR2068034	NGBS1059	human	Canada	ST459	IV
SRR2068035	NGBS1061	human	Canada	ST459	IV
SRR2068036	NGBS1062	human	Canada	ST459	IV
SRR2068037	NGBS1063	human	Canada	ST459	IV
SRR2068038	NGBS1064	human	Canada	ST459	IV
SRR2068039	NGBS1065	human	Canada	ST459	IV
SRR2068040	NGBS1066	human	Canada	ST459	IV
SRR2068041	NGBS1067	human	Canada	ST459	IV
SRR2068042	NGBS1068	human	Canada	ST459	IV
SRR2068043	NGBS1071	human	Canada	ST459	IV
SRR2068044	NGBS1072	human	Canada	ST459	IV
SRR2068045	NGBS1074	human	Canada	ST459	IV
SRR2068046	NGBS1075	human	Canada	ST459	IV
SRR2068047	NGBS1079	human	Canada	ST452	IV
SRR2068048	NGBS1080	human	Canada	SLV459	IV
SRR2068049	NGBS1082	human	Canada	ST459	IV
SRR2068050	NGBS1083	human	Canada	ST459	IV
SRR2451885	NGBS129	human	Canada	ST31	III
SRR2451888	NGBS147	human	Canada	ST17	III
SRR2451889	NGBS149	human	Canada	ST17	III
SRR2451892	NGBS169	human	Canada	ST17	III
SRR2451894	NGBS205	human	Canada	ST17	III
SRR2451896	NGBS222	human	Canada	ST17	III
SRR2451897	NGBS238	human	Canada	ST17	III
SRR2451898	NGBS239	human	Canada	ST17	III
SRR2451901	NGBS277	human	Canada	ST17	III
SRR2451902	NGBS282	human	Canada	ST17	III

Table A.2 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
SRR2451904	NGBS296	human	Canada	ST17	III
SRR2451905	NGBS297	human	Canada	ST17	III
SRR2451906	NGBS299	human	Canada	ST17	III
SRR2451907	NGBS306	human	Canada	ST17	III
SRR2451908	NGBS312	human	Canada	ST17	III
SRR2451909	NGBS317	human	Canada	ST17	III
SRR2451910	NGBS318	human	Canada	ST290	III
SRR2451914	NGBS356	human	Canada	ST17	III
SRR2451915	NGBS361	human	Canada	ST17	III
SRR2451916	NGBS362	human	Canada	ST17	III
SRR2451917	NGBS368	human	Canada	ST17	III
SRR2451919	NGBS374	human	Canada	ST17	III
SRR2451922	NGBS398	human	Canada	ST17	III
SRR2451923	NGBS403	human	Canada	ST17	III
SRR2451925	NGBS421	human	Canada	ST17	III
SRR2451926	NGBS422	human	Canada	ST17	III
SRR2451929	NGBS456	human	Canada	ST17	III
SRR2451930	NGBS464	human	Canada	ST17	III
SRR2451931	NGBS469	human	Canada	ST17	III
SRR2451932	NGBS470	human	Canada	ST17	III
SRR2451933	NGBS483	human	Canada	ST17	III
SRR2451934	NGBS485	human	Canada	ST17	III
SRR2451935	NGBS486	human	Canada	ST17	III
SRR2451936	NGBS500	human	Canada	ST17	III
SRR2451938	NGBS502	human	Canada	ST95	III
SRR2451939	NGBS515	human	Canada	ST17	III
SRR2451942	NGBS534	human	Canada	ST17	III
SRR2451943	NGBS551	human	Canada	ST17	III
SRR2451945	NGBS583	human	Canada	ST17	III
SRR2451946	NGBS593	human	Canada	ST17	III
SRR2451947	NGBS594	human	Canada	ST17	III
SRR2451948	NGBS596	human	Canada	ST17	III
SRR2451949	NGBS607	human	Canada	ST17	III
SRR2451950	NGBS608	human	Canada	ST17	III
SRR2451951	NGBS609	human	Canada	ST17	III
SRR2451952	NGBS613	human	Canada	ST17	III
SRR2451954	NGBS618	human	Canada	ST17	III
SRR2451955	NGBS622	human	Canada	ST148	III
SRR2451958	NGBS632	human	Canada	ST17	III
SRR2451960	NGBS641	human	Canada	ST17	III
SRR2451961	NGBS644	human	Canada	ST17	III
SRR2451962	NGBS650	human	Canada	ST17	III
SRR2981533	SGBS103	human	USA	ST1	V
SRR2981534	SGBS104	human	USA	ST1	V

Table A.2 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
SRR2981535	SGBS105	human	USA	SLV1	V
SRR2981536	SGBS106	human	USA	ST1	V
SRR2981541	SGBS111	human	USA	ST1	V
SRR2981542	SGBS114	human	USA	ST1	V
SRR2981543	SGBS115	human	USA	ST1	V
SRR2981545	SGBS118	human	USA	ST1	V
SRR2981546	SGBS119	human	USA	ST1	V
SRR2981547	SGBS120	human	USA	ST1	V
SRR2981548	SGBS122	human	USA	ST1	V
SRR2981550	SGBS126	human	USA	ST1	V
SRR2981554	SGBS133	human	USA	ST1	V
SRR2981555	SGBS135	human	USA	ST1	V
SRR2981558	SGBS140	human	USA	ST1	V
SRR2981559	SGBS141	human	USA	ST1	V
SRR2981560	SGBS143	human	USA	ST1	V
SRR2981561	SGBS144	human	USA	ST1	V
SRR2981562	SGBS145	human	USA	ST1	V
SRR2981563	SGBS146	human	USA	ST1	V
SRR2981564	SGBS147	human	USA	ST1	V
SRR2981565	SGBS148	human	USA	ST1	V
SRR2981566	SGBS150	human	USA	ST1	V
SRR2981568	SGBS152	human	USA	ST1	V
SRR2981569	SGBS031	human	USA	ST1	V
SRR2981570	SGBS032	human	USA	ST1	V
SRR2981571	SGBS033	human	USA	ST1	V
SRR2981572	SGBS034	human	USA	ST1	V
SRR2981573	SGBS035	human	USA	ST1	V
SRR2981574	SGBS036	human	USA	ST1	V
SRR2981575	SGBS037	human	USA	ST1	V
SRR2981576	SGBS038	human	USA	ST1	V
SRR2981577	SGBS039	human	USA	ST1	V
SRR2981578	SGBS040	human	USA	ST1	V
SRR2981579	SGBS041	human	USA	ST1	V
SRR2981580	SGBS042	human	USA	ST1	V
SRR2981582	SGBS044	human	USA	ST1	V
SRR2981583	SGBS045	human	USA	ST1	V
SRR2981584	SGBS046	human	USA	ST1	V
SRR2981585	SGBS047	human	USA	ST1	V
SRR2981586	SGBS048	human	USA	ST1	V
SRR2981587	SGBS049	human	USA	SLV1	V
SRR2981589	SGBS051	human	USA	ST1	V
SRR2981590	SGBS052	human	USA	ST1	V
SRR2981591	SGBS053	human	USA	ST1	V
SRR2981592	SGBS054	human	USA	ST1	V

Table A.2 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
SRR2981593	SGBS056	human	USA	ST1	V
SRR2981594	SGBS057	human	USA	ST1	V
SRR2981595	SGBS058	human	USA	ST1	V
SRR2981597	SGBS060	human	USA	ST1	V
SRR2981604	SGBS067	human	USA	ST1	V
SRR2981606	SGBS069	human	USA	ST1	V
SRR2981610	SGBS074	human	USA	ST1	V
SRR2981611	SGBS075	human	USA	ST1	V
SRR2981612	SGBS076	human	USA	ST1	V
SRR2981613	SGBS077	human	USA	ST1	V
SRR2981614	SGBS078	human	USA	ST1	V
SRR2981615	SGBS079	human	USA	ST153	V
SRR2981618	SGBS082	human	USA	ST1	V
SRR2981620	SGBS084	human	USA	ST1	V
SRR2981621	SGBS085	human	USA	ST1	V
SRR2981623	SGBS087	human	USA	ST1	V
SRR2981626	SGBS092	human	USA	ST1	V
SRR2981628	SGBS094	human	USA	ST1	V
SRR2981630	SGBS096	human	USA	ST1	V
SRR2981632	SGBS098	human	USA	ST1	V
SRR494266	CCUG_24810	human	Sweden	ST19	III
SRR494270	CCUG_37430	human	Sweden	ST19	II
SRR494271	CCUG_29376	human	Sweden	ST12	Ib
SRR494272	CCUG_30636	human	Sweden	ST1	V
SRR494276	FSL_S3-137	human	USA	ST8	Ib
SRR494279	FSL_S3-001	human	USA	ST1	V
SRR494280	FSL_S3-003	human	USA	ST19	III
SRR494281	CCUG_91	human	Sweden	ST28	II
SRR494284	FSL_S3-102	human	USA	ST31	III
SRR494285	FSL_F2-343	human	USA	ST88	Ia
SRR494286	FSL_S3-014	human	USA	ST8	Ib
SRR494288	FSL_S3-090	human	USA	ST23	Ia
SRR494289	FSL_S3-023	human	USA	ST1	V
SRR494292	LMG_15085	human	USA	ST17	III
SRR494295	CCUG_49086	human	Sweden	ST17	III
SRR494296	CCUG_49100	human	Sweden	ST1	V
SRR494297	CCUG_44140	human	Sweden	ST1	V
SRR494298	LMG_15081	human	USA	ST25	Ia
SRR494299	LMG_15083	human	USA	ST7	Ia
SRR494300	LMG_15084	human	USA	ST19	II
SRR494302	CCUG_49072	human	Sweden	ST524	V
SRR494303	CCUG_49087	human	Sweden	ST17	III
SRR494306	CCUG_47293	human	Sweden	SLV9	Ib
SRR494309	CCUG_44074	human	Sweden	ST23	Ia

Table A.2 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
SRR494311	CCUG_39096_A	human	Sweden	ST9	Ib
SRR494317	CCUG_37739	human	Sweden	ST23	Ia
SRR494322	BSU253	human	Germany	ST23	Ia
SRR494323	BSU247	human	Germany	ST26	V
SRR494325	BSU248	human	Germany	ST12	Ib
SRR494327	BSU252	human	Germany	ST1	V
SRR494328	BSU454	human	Germany	ST8	Ib
SRR494330	LMG_15094	human	Belgium	ST17	III
SRR494331	LMG_15095	human	Belgium	ST17	III
SRR494332	LMG_15090	human	Belgium	ST8	Ib
SRR494336	LMG_15091	human	Belgium	SLV786	IV
SRR494339	BSU451	human	Germany	ST103	Ia
SRR494340	BSU96	human	Germany	ST17	III
SRR494341	BSU165	human	Germany	ST28	II
SRR494342	BSU174	human	Germany	ST41	V
SRR494343	BSU92	human	Germany	ST196	IV
SRR494344	BSU133	human	Germany	ST6	Ib
SRR494346	BSU260	human	Germany	ST88	Ia
SRR494355	GB00202	human	Canada	ST10	Ib
SRR494358	GB00097	human	Canada	ST17	III
SRR494359	GB00111	human	Canada	ST32	III
SRR494360	GB00115	human	Canada	ST17	III
SRR494361	GB00190	human	Canada	ST23	Ia
SRR494364	GB00083	human	Canada	ST1	VI
SRR494365	GB00084	human	Canada	ST1	VIII
SRR494366	GB00003	human	Canada	ST12	Ib
SRR494367	GB00012	human	Canada	ST1	V
SRR494368	GB00018	human	Canada	ST444	Ia
SRR494369	GB00082	human	Canada	ST2	IV
SRR494370	GB00013	human	Canada	ST1	V
SRR494371	GB00020	human	Canada	ST1	V
SRR494372	GB00002	human	Canada	ST23	Ia
SRR494374	GB00864	human	USA	ST10	II
SRR494375	GB00663	human	Canada	ST19	III
SRR494376	GB00679	human	Canada	ST2	II
SRR494377	GB00654	human	Canada	ST17	III
SRR494378	GB00651	human	Canada	ST8	Ib
SRR494380	GB00640	human	Canada	ST26	V
SRR494383	GB00588	human	Canada	ST447	II
SRR494386	GB00555	human	Canada	ST12	Ib
SRR494388	GB00264	human	Canada	ST10	II
SRR494389	GB00279	human	Canada	ST2	II
SRR494390	GB00300	human	Canada	ST130	IX
SRR494392	GB00241	human	Canada	ST1	V

Table A.2 continued from previous page

ACCESSION/RUN	ISOLATE	HOST	COUNTRY	ST	SEROTYPE
SRR494393	GB00226	human	Canada	ST28	II
SRR494394	GB00245	human	Canada	ST23	Ia
SRR494395	GB00247	human	Canada	ST24	Ia
SRR494396	GB00219	human	Canada	ST8	Ib
SRR494568	GB00932	human	USA	ST23	Ia
SRR494611	GB00887	human	USA	ST23	Ia
SRR494612	GB00888	human	USA	ST41	V
SRR494631	GB00914	human	USA	ST8	Ib
SRR494632	GB00922	human	USA	ST88	Ia
SRR494635	GB00911	human	USA	ST452	IV
SRR494636	GB00867	human	USA	ST23	Ia
SRR494637	GB00884	human	USA	ST19	III
SRR494638	GB00955	human	USA	SLV1	V
SRR494645	GB00924	human	USA	ST1	V
SRR494656	GB00933	human	USA	ST452	IV
SRR494658	GB00901	human	USA	ST459	IV
SRR494659	GB00909	human	USA	ST12	Ib
SRR494660	GB00874	human	USA	ST1	II
SRR496544	MRI_Z1-211	bovine	Italy	ST1	V
SRR496556	MRI_Z1-025	bovine	Denmark	ST1	V
SRR496920	MRI_Z1-199	seal	United Kingdom	ST23	Ia
SRR497007	GB00893	human	USA	ST8	Ib
SRR497011	STIR-CD-14	fish	Vietnam	ST491	III
SRR497118	MRI_Z1-198	dolphin	United Kingdom	ST12	Ib
SRR525043	GB00548	human	Canada	ST88	Ia
SRR628712	GB00999	human	USA	ST1	V
SRR6996453	QMA0323	fish	Australia	ST261	Ib

Table A.3: Clonal complexes (CC) of the group B *Streptococcus* genomes included in dataset 2. Sequence types (ST) and single locus variants (SLV) identified and grouped among the different CC according to Richards et al. (2019) are shown.

CLONAL COMPLEX (CC)	SEQUENCE TYPES (ST)
1	1, 2, 153, 196, 387, 459, 524, 682, 710
7	6, 7, 41
12	3, 8, 9, 10, 12, 255, 523, 604, 652, 711
17	17, 31, 32, 95, 148, 290, 315, 467
19	19, 28, 110, 335, 447, 479
22	22
23	23, 24, 25, 88, 144, 452
26	26
67	67
103	103
130	130
260/261	261, 257 (SLV)
283	283, 491
609	609

Table A.4: Prophage integrase family type identity and similarity table. Progressive numbers are assigned to sixteen integrase types, at twelve different integration sites. Numbers indicate query coverage and percentage of identity (QC/ID) based on blastp comparison of the amino acid sequences. Comparisons were always run with the longest sequence as the database and the shortest sequence as the query.

PHAGE INTEGRASE	1	2.1	2.2	3	4	5	6.1	6.2	1at7	8	9.1	9.2	10	11.1	11.2	11.3	12
1	100/	84/	87/	92/	91/	51/	97/	98/	100/	99/	92/	93/	18/	38/	4/	92/	39/
	100	23.68	25.15	26.69	28.14	26.71	35.7	33.16	100	28.72	24.35	22.25	32.26	24.24	27.78	26.76	36.84
2.1	84/	100/	100/	49/	45/	67/	94/	96/	84/	80/	99/	99/	17/	25/	34/	78/	90/
	23.68	100	71.35	28.10	24.04	21.61	26.86	27.73	23.68	28.99	59.32	58.92	50	40.74	25.58	24.23	25.14
2.2	87/	100/	100/	42/	87/	45/	89/	93/	87/	76/	99/	99/	13/	15/	5/	77/	86/
	25.15	71.35	100	30.46	20.30	24.14	23.35	26.29	25.15	23.67	54.65	54.26	45.45	34.78	57.14	27.21	22.69
3	92/	49/	42/	100/	95/	58/	94/	89/	92/	97/	78/	81/	29/	21/	17/	98/	90/
	26.69	28.10	30.46	100	31.75	23.70	29.84	26.11	26.69	34.75	26.48	23.97	26.80	28.26	26.32	50.83	19.83
4	91/	45/	87/	95/	100/	91/	80/	93/	91/	97/	39/	85/	17/	31/	0/	97/	39/
	28.14	24.04	20.30	31.75	100	19.49	27.30	28.12	28.14	34.04	22.70	22.09	32.56	35.29	0	32.55	24.07
5	51/	67/	45/	58/	91/	100/	59/	72/	51/	77/	46/	74/	15/	53/	20/	72/	22/
	26.71	21.61	24.14	23.70	19.49	100	30.16	24.31	26.71	23.08	25.77	19.76	50.00	30.00	25.93	24.70	42.86
6.1	97/	94/	89/	94/	80/	59/	100/	98/	97/	81/	94/	93/	17/	45/	30/	90/	90/
	35.7	26.86	23.35	29.84	27.30	30.16	100	58.06	35.7	28.16	27.20	26.15	47.06	33.33	17.74	29.18	22.75
6.2	98/	96/	93/	89/	93/	72/	98/	100/	98/	82/	97/	96/	43/	11/	30/	95/	91/
	33.16	27.73	26.29	26.11	28.12	24.31	58.06	100	33.16	27.85	27.12	26.57	29.82	50.00	32.43	28.61	23.99
1at7	100/	84/	87/	92/	91/	51/	97/	98/	100/	99/	92/	93/	18/	38/	4/	92/	39/
	100	23.68	25.15	26.69	28.14	26.71	35.7	33.16	100	28.72	24.35	22.25	32.26	24.24	27.78	26.76	36.84

Table A.4 continued from previous page

PHAGE INTEGRASE	1	2.1	2.2	3	4	5	6.1	6.2	1at7	8	9.1	9.2	10	11.1	11.2	11.3	12
8	99/	80/	76/	97/	97/	77/	81/	82/	99/	100/	83/	93/	36/	34/	3/	96/	72/
	28.72	28.99	23.67	34.75	34.04	23.08	28.16	27.85	28.72	100	23.30	23.03	36.84	35.14	36.84	36.54	23.10
9.1	92/	99/	99/	78/	39/	46/	94/	97/	92/	83/	100/	100/	13/	20/	16/	78/	84/
	24.35	59.32	54.65	26.48	22.70	25.77	27.20	27.12	24.35	23.30	100	83.71	38.46	45.16	25.00	23.45	23.19
9.2	93/	99/	99/	81/	85/	74/	93/	96/	93/	93/	100/	100/	12/	19/	18/	90/	85/
	22.25	58.92	54.26	23.97	22.09	19.76	26.15	26.57	22.25	23.03	83.71	100	39.13	40.74	25.00	24.49	22.75
10	18/	17/	13/	29/	17/	15/	17/	43/	18/	36/	13/	12/	100/	99/	98/	37/	32/
	32.26	50	45.45	26.80	32.56	50.00	47.06	29.82	32.26	36.84	38.46	39.13	100	44.54	47.16	36.84	30.95
11.1	38/	25/	15/	21/	31/	53/	45/	11/	38/	34/	20/	19/	99/	100/	99/	16/	28/
	24.24	40.74	34.78	28.26	35.29	30.00	33.33	50.00	24.24	35.14	45.16	40.74	44.54	100	57.92	27.27	23.08
11.2	4/	34/	5/	17/	0/	20/	30/	30/	4/	3/	16/	18/	98/	99/	100/	7/	2/
	27.78	25.58	57.14	26.32	0	25.93	17.74	32.43	27.78	36.84	25.00	25.00	47.16	57.92	100	42.86	54.55
11.3	92/	78/	77/	98/	97/	72/	90/	95/	92/	96/	78/	90/	37/	16/	7/	100/	7/
	26.76	24.23	27.21	50.83	32.55	24.70	29.18	28.61	26.76	36.54	23.45	24.49	36.84	27.27	42.86	100	25.93
12	39/	90/	86/	90/	39/	22/	90/	91/	39/	72/	84/	85/	32/	28/	2/	7/	100/
	36.84	25.14	22.69	19.83	24.07	42.86	22.75	23.99	36.84	23.10	23.19	22.75	30.95	23.08	54.55	25.93	100

Table A.5: Insertion sites of sixteen site-specific integrases for prophages identified in group B *Streptococcus* isolated from across multiple host species and countries. Putative attachment sites (*att*) are shown when known. For *att* sites that differed slightly at the two ends of the lysogenic prophages, left and right (*att*L and *att*R) sequences are specified and differences are highlighted in bold.

INSERTION SITE	PHAGE INTEGRASE TYPE	PHAGE INTEGRASE LENGTH (AA)	GENE	PUTATIVE ATTACHMENT SITE
GBS1	GBS*Int*1	369	*comX* (sigma-70 family RNA polymerase sigma factor) - 3' end	*att*L TTTTTTGTTATAATATAA**GA** *att*R TTTTTTGTTATAATATAA**TA**
GBS2	GBS*Int*2.1 GBS*Int*2.2	360 363	tRNA methyltransferase - 3' end	ATCCCCTCCTCTCCTTTAAT
GBS3	GBS*Int*3	368	*rplS* - 3' end	*att*L GATTCC**G**GCAGGGGACAT *att*R GATTCC**A**GCAGGGGACAT
GBS4	GBS*Int*4	389	HU, histone-like DNA-binding protein	CTCTTAAAGACGCTGTTAAATA ATTCGTCTAGAAAAACCTTGTC ATATCAATGTTTATTGATAGCGAC AAGGTTC
GBS5	GBS*Int*5	331	*rpsI* - 3' end (within ICE)	-
GBS6	GBS*Int*6.1 GBS*Int*6.2	366 366	CatB-related O-acetyltransferase - 5' end	TGGAGCCGGTGGGAGT
GBS7	GBS*Int*1	369	*hylB* - 5' end	*att*L TTTTTTGTTATAATAT**AAGA** *att*R TTTTTTGTTATAATAT**GAGA**

Table A.5 continued from previous page

INSERTION SITE	PHAGE INTEGRASE TYPE	PHAGE INTEGRASE LENGTH (AA)	GENE	PUTATIVE ATTACHMENT SITE
GBS8	GBS*Int*8	382	YbaB/EbfC family nucleoid-associated protein	TTTTGCATATTCATCATA
GBS9	GBS*Int*9.1	360	*nhaK* (sodium/proton antiporter)	AAGGCGGTAGACGGATTTGAA
	GBS*Int*9.2	359	- 3' end	
GBS10	GBS*Int*10	476	DNA-binding protein WhiA - 3' end	-
GBS11	GBS*Int*11.1	489	*gspF* or *gspF*+competence proteins/	CTTTTAGAATGTTTGGTA
	GBS*Int*11.2	486	type II secretion system proteins	-
	GBS*Int*11.3	367	- 3' end	-
GBS12	GBS*Int*12	386	5-formyltetrahydrofolate cyclo-ligase - 5' end	-

Table A.6: Results of mobile genetic elements (MGE) screening (prophages, phage-inducible chromosomal islands or PICI, integrative conjugative elements or ICE, integrative and mobilizable elements or IME, and plasmids) of 503 group B *Streptococcus* genomes, divided by major host species.

MGE type	HOST SPECIES (tot number of genomes)		
	HUMAN (*n*=486)	BOVINE (*n*=5)	FISH (*n*=8)
Prophages (complete)	274	0	3
Prophages (incomplete)	400	7	4
PICI	321	3	2
ICE	234	6	6
IME	361	5	10
Plasmids	10	-	-
Tot	1600	21	25

Table A.7: Distribution of complete prophages among major clonal complexes (CC) in dataset 2.

CLONAL COMPLEX (total number of genomes)	NUMBER OF PROPHAGES
1 (260)	136
7 (7)	5
12 (38)	31
17 (90)	44
19 (29)	23
22 (3)	0
23 (56)	30
26 (4)	3
67 (1)	0
103 (2)	1
130 (2)	0
260/261 (2)	1
283 (5)	5
609 (1)	1

Table A.8: Distribution of complete prophages classified based on their integrase types among major sequence types (ST) in dataset 2.

PROPHAGE	SEQUENCE TYPE (total number of isolates)				
INTEGRASE	ST1 (147)	ST17 (77)	ST19 (16)	ST23 (28)	ST459 (85)
GBS*Int*1	-	6	-	-	-
GBS*Int*2.1	-	-	-	-	1
GBS*Int*2.2	-	-	1	-	52
GBS*Int*3	20	-	7	6	13
GBS*Int*4	-	9	2	-	1
GBS*Int*6.1	-	-	1	-	-
GBS*Int*6.2	1	-	1	1	-
GBS*Int*8	1	4	-	-	-
GBS*Int*9.1	-	1	-	-	-
GBS*Int*9.2	-	7	-	-	2
GBS*Int*10	-	2	-	-	-
GBS*Int*11.1	-	1	2	1	13
GBS*Int*11.2	8	4	1	3	10
GBS*Int*12	-	-	2	-	-

Table A.9: Distribution of complete prophages among continents in dataset 2.

CONTINENT (total number of genomes)	NUMBER OF PROPHAGES
Africa (1)	1
Asia (10)	5
Europe (117)	68
North America (351)	195
Oceania (18)	9
South America (2)	0
Unknown (4)	2

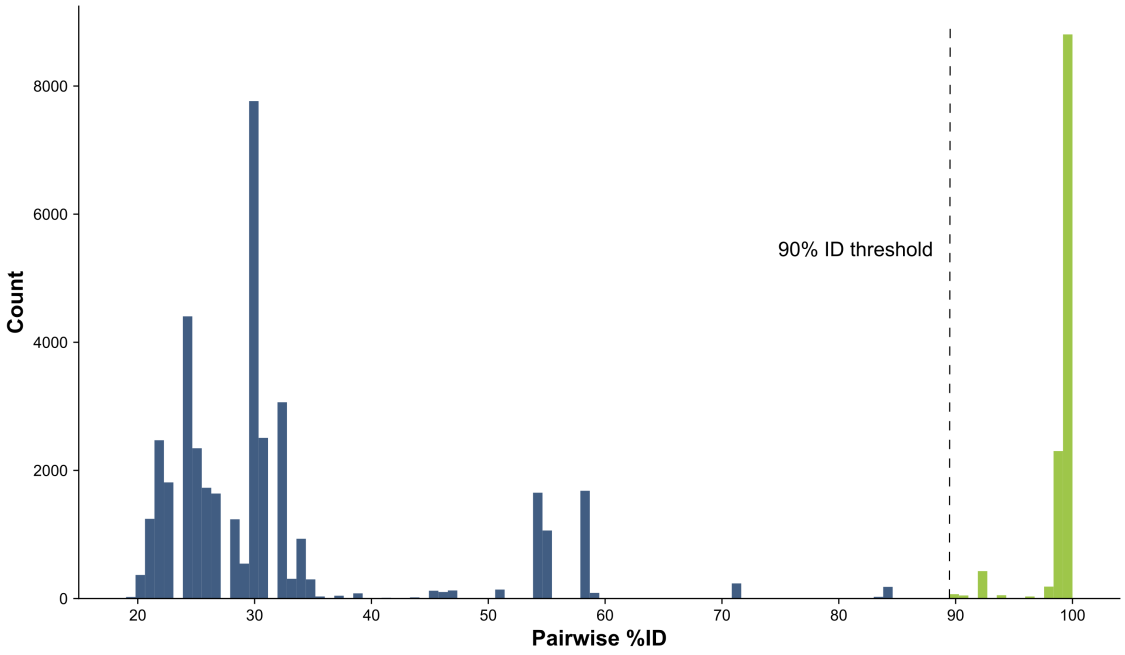


Figure A.1: Histogram plot showing the distribution of the blastp percentage of identity (%ID) scores between all pairs of group B *Streptococcus* prophage integrase amino acid sequences identified in this study. Green bars represent %ID of matching integrase type pairs, blue bars show %ID of unmatched pairs. A minimum threshold of 90%ID blastp score (dashed line) was adopted to consider two integrase protein sequences as the same.

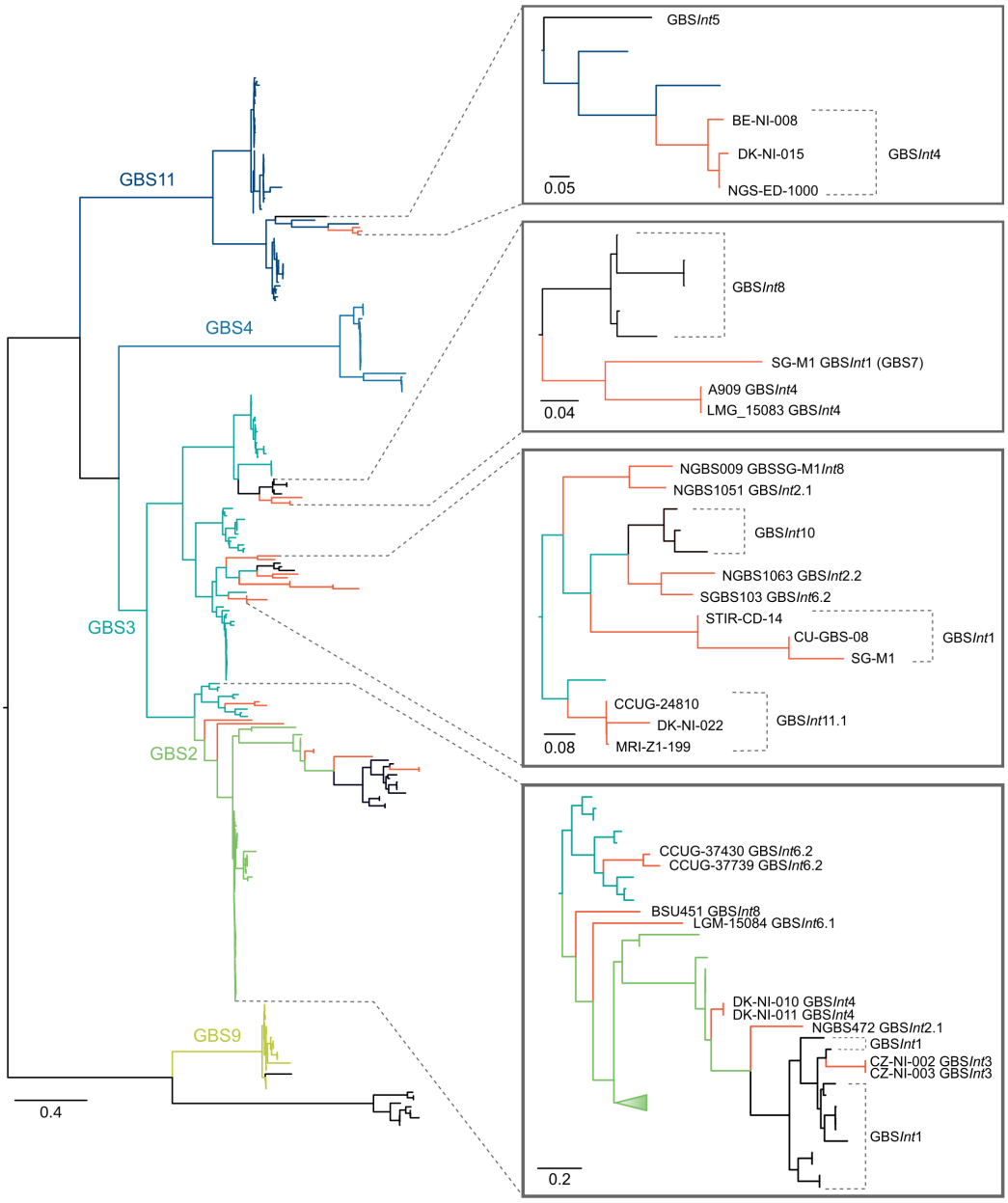


Figure A.2: Approximate maximum-likelihood phylogenetic tree of complete prophages (*n*=266) from group B *Streptococcus* (GBS) identified in dataset 2 and 22 prophages from van der Mee-Marquet et al. (2018). Phage clusters that are concordant with a particular insertion site and its integrase type/subtypes have been indicated (GBS2: green; GBS3: turquoise; GBS4: blue; GBS9: yellow; GBS11: navy). Red branches correspond to prophages that cluster within a group of phages with a different insertion site. Magnifications of such exceptions are shown and the integrase type has been indicated. Black branches correspond to minor clusters (GBS1, GBS5, GBS8, GBS10, GBS12) that are embedded within larger ones. Tree was rooted at midpoint.

Figure A.3: Approximate maximum-likelihood phylogenetic tree of 266 phage integrase protein sequences identified in group B *Streptococcus* (GBS) in dataset 2. Integrases of the same type largely clustered within their assigned group, with the exception of GBS*Int*11.3 which clustered separate from GBS*Int*11.1 and GBS*Int*11.2. Tree was rooted at midpoint.

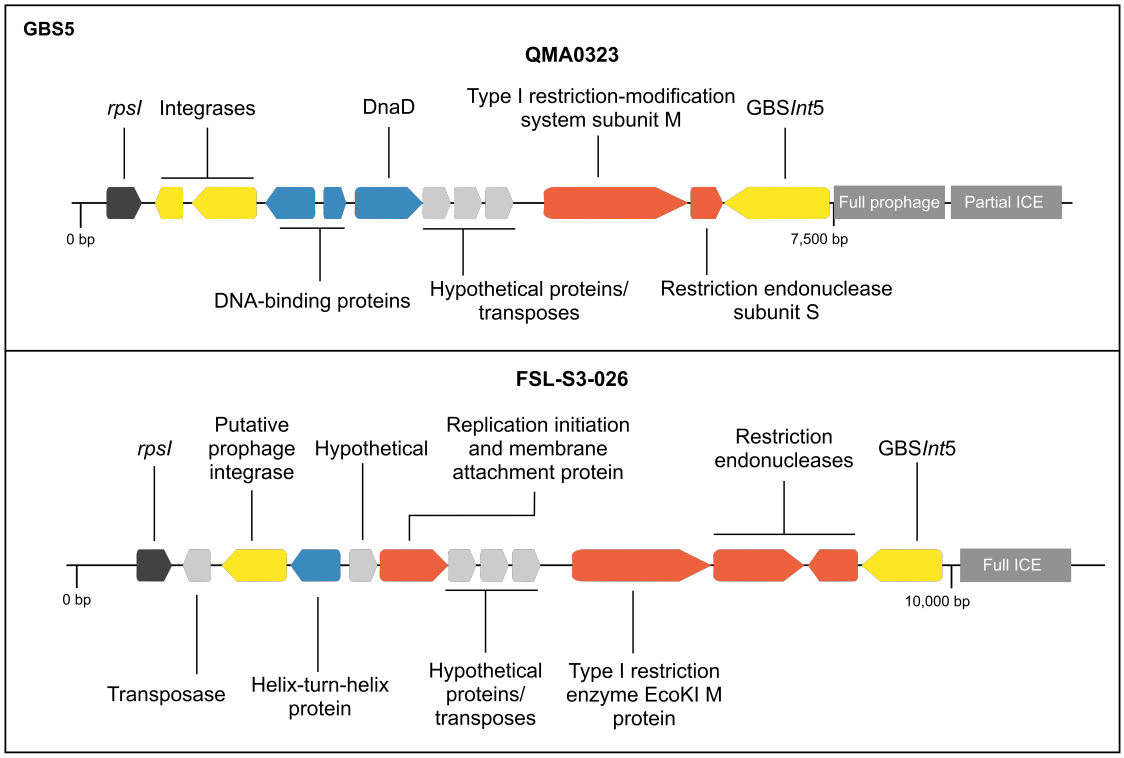


Figure A.4: GBS5 insertion site as observed in group B *Streptococcus* (GBS) genome QMA0323, where GBS*Int*5 is followed by a full prophage and by part of an ICE, and in genome FSL_S3-026, where the integrase was found as a singleton within a larger ICE. The gene *rpsI* is the closest upstream chromosomal gene to GBS*Int*5. The partial ICE after the prophage in QMA0323 showed similarity with part of the ICE in FSL_S3-026 (∼9,000bp, grey boxes at the right hand side). Genes are colour-coded based on function (black: chromosomal genes; yellow: site-specific integrase; dark blue: transcriptional regulators; light grey: hypothetical; red: other genes).

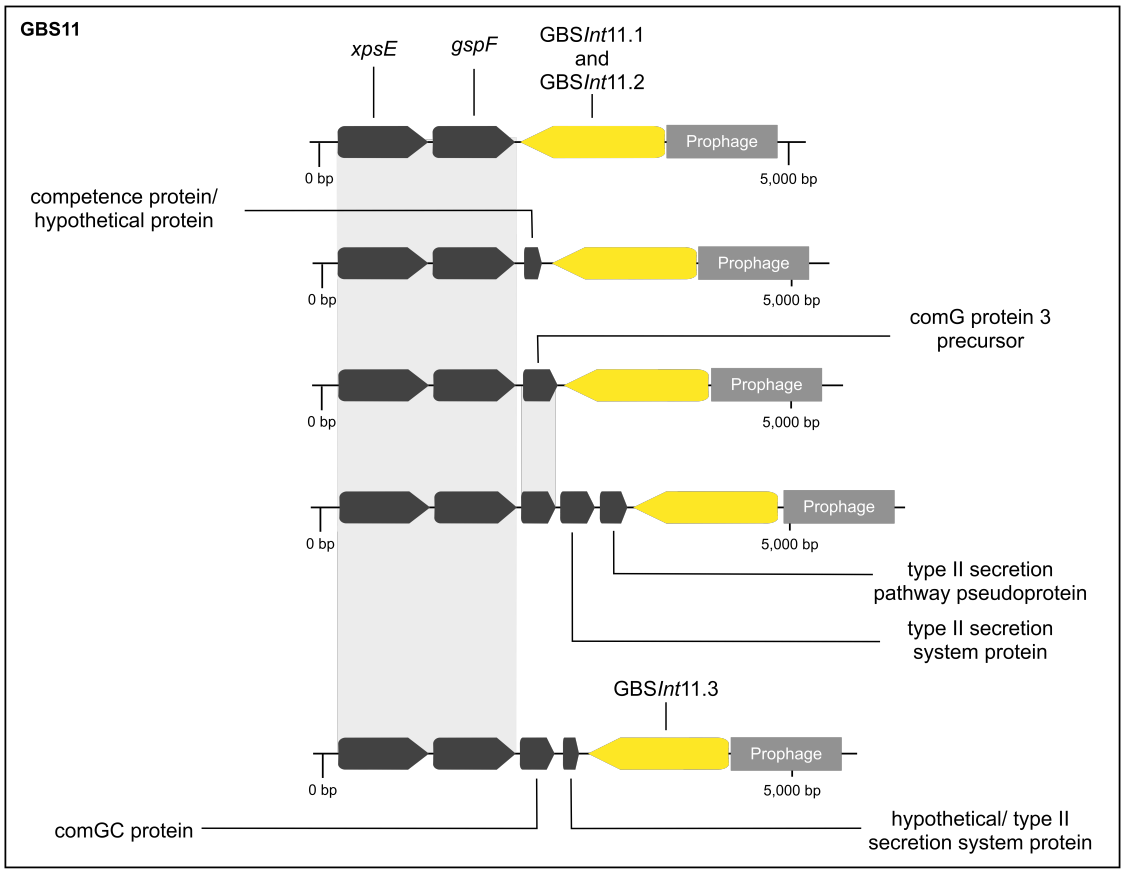


Figure A.5: GBS11 insertion site variations in group B *Streptococcus* (GBS): *xpsE* and *gspF* genes are always present, and may be followed directly by GBS*Int*11.1 or GBS*Int*11.2 and the rest of the prophage. In other cases, additional small genes for competence and secretion systems were inserted between *gpsF* and the prophage, regardless of integrase type, with GBS*Int*11.3 also found in this configuration. Genes are colour-coded based on function (black: chromosomal genes; yellow: site-specific integrase).

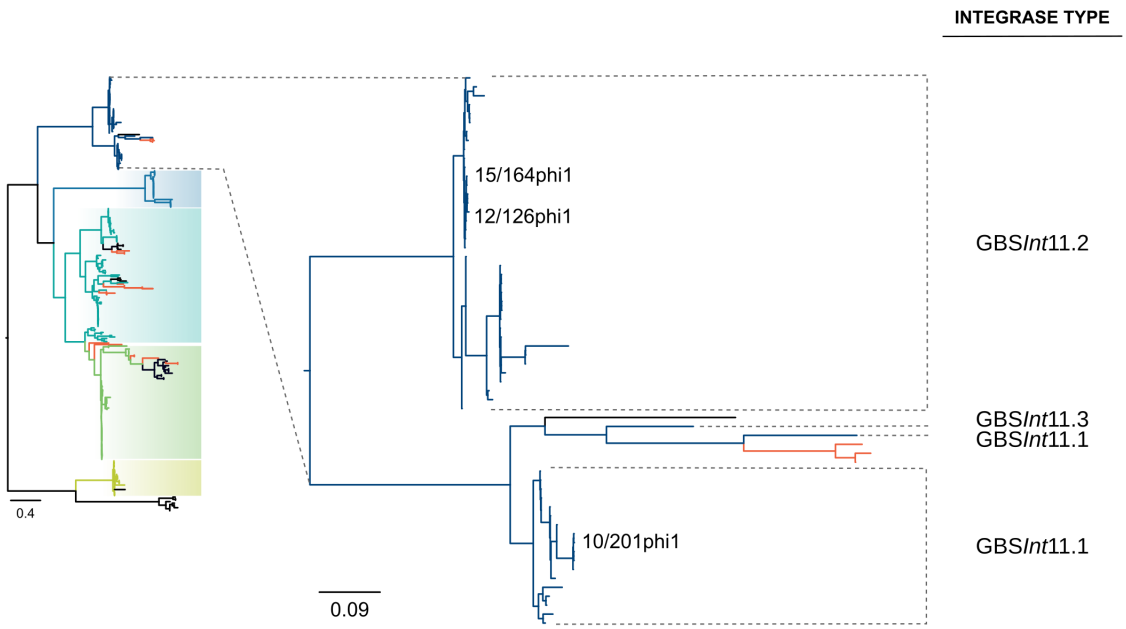


Figure A.6: Magnification of the approximate maximum-likelihood phylogenetic tree cluster of prophages with insertion site GBS11 (integrase types GBS*Int*11.1, GBS*Int*11.2 and one example of GBS*Int*11.3). Prophages F1 (GBS*Int*11.2) and F2 (GBS*Int*11.1) from van der Mee-Marquet et al. (2018) have been indicated (blue branches: phages with either GBS*Int*11.1, GBS*Int*11.2 or GBS*Int*11.3; black branch: prophage GBS*Int*5; red branches: prophages with GBS*Int*4). Tree was rooted at midpoint.

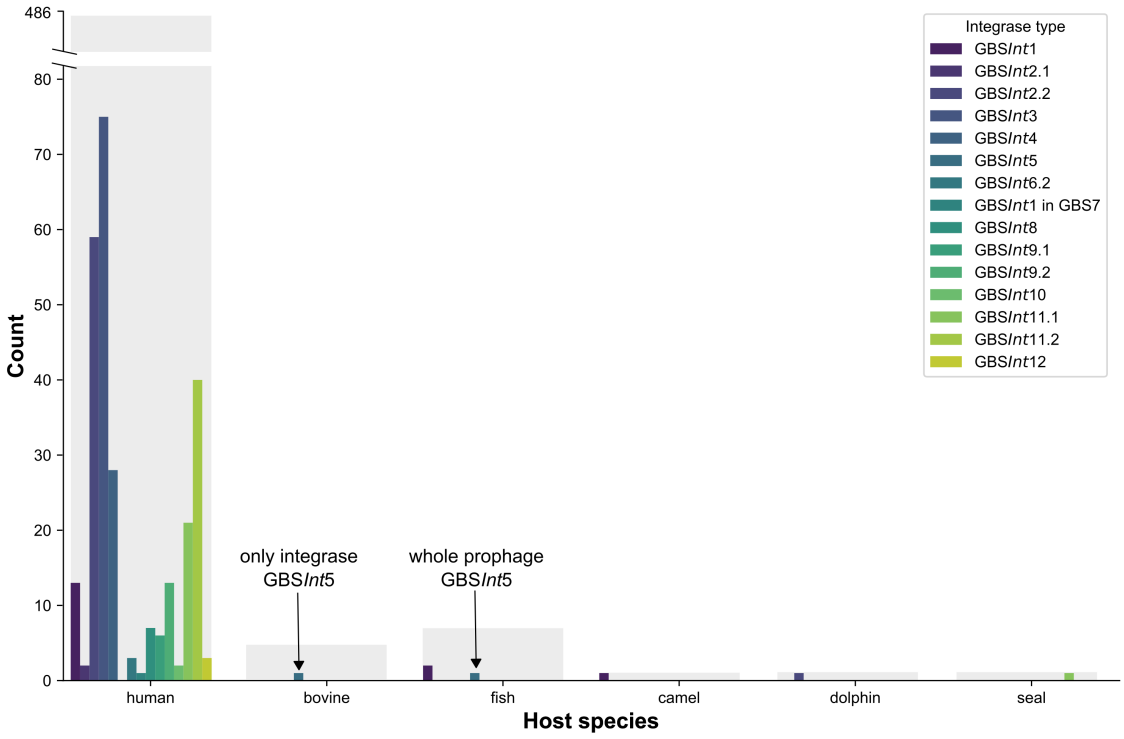


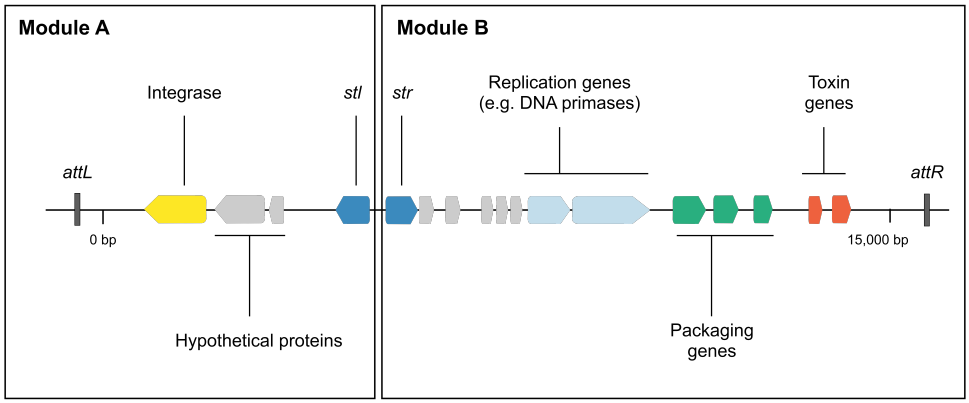
Figure A.7: Distribution of complete prophages classified based on their integrase types (GBS*Int*1 to GBS*Int*12) in a publicly available group B *Streptococcus* (GBS) dataset of 503 sequences comprising genomes from seven different host species and originating from different countries. Coloured bars refer to complete prophages, with the exception of the bovine blue bar, which refers to integrase GBS*Int*5, as a singleton, i.e. not associated with a full prophage. Grey bars show the total number of isolates per host species. No prophages or integrases were detected in the single canine GBS genome included in the study.

A.2 Phage-inducible chromosomal islands (PICI) manual detection method

1. Genomes were freshly annotated using Prokka v1.13.1.

2. Manual searches were performed with a text editor (Atom v1.48.0) for genes annotated as "site-specific integrase", "integrase" or "recombinase".

3. For every genome, after the identification of integrase genes, manual (i.e. non-automated) recognition of the genes at their 5' end and upstream was performed to look for the basic structure of phage inducible-chromosomal islands (PICI):



Module A (genes oriented in the same direction as the integrase)

- It starts with the **integrase** gene (which could also be annotated as site-specific integrase or sometimes recombinase); there could be a few genes, in variable number and with various functions, oriented in the same direction as the integrase before the next element;
- The next gene that is always present is named ***stl***; it is a **transcriptional regulator** and it is oriented in the same direction as the integrase (it could be annotated as HTH-domain family protein or DNA-binding protein);

Module B (genes oriented in the opposite direction)

- The next gene that is always present is named ***str***; it is a **transcriptional regulator** and it is oriented in the opposite direction compared to the integrase (it could be annotated as HTH-domain family protein or DNA-binding protein);

- The number of genes that can be found from here onward is variable and these genes can have different functions; usually there are a few **replication genes** (e.g. a DNA primase); they are all oriented in the opposite direction compared to the integrase;
- Sometimes, after the replication module, genes annotated as **packaging genes** can be present;
- **Toxin genes** can be found in different parts of the element, usually towards the end.

A.3 Database of prophage and phage-inducible chromosomal island (PICI) site-specific integrases identified in this study shown as amino acid sequences

GBS*Int*1 (identical to integrase at GBS7)

MIEKYTKKDGTTAYRLRAYLGVDPMTGKQVRTTRQGFKTEREAKRAEVKLIDDFQRQGAWKSNDK
TTFDDVAKLWFEQYRNTVKPSTFLVNQNYYKTILKPHLGQLQMTKITVMICQKFVNCLSRYSGYRLY
LSLANRIFKFAVNLGIIDNNPMSKTLRSKCTYKNMDTLTKKYYTKEELNAFLRIVEAEETLEMRLIYRL
LSYGGFRIGELIALKDTDFDFRNNTISITKTIAYTKEGWAVQSPKTKKSNRTISMDAETMTLAKLYIKQ
SIKPLHGSFKLFNFASDTVRKRLDRFILKHGLKRIPHGFRHTHASLLFEAGIPAKIAQERLGHAKIAITMD
LYTHLSKKSKDNVADKLAELVAI

GBS*Int*2.1

MASYRKLDSGWEYRIIYKDINGIRREKSKRGFSTKTLAKAAAVKAEREINSTDTELLDTTFYDYSIQWA
EVYKRPHVTAKTWQTYSKNFKHIKHYFGNMKVKDITHTFYQKVLNEFGEIVAQQTLDKFHYQVKGA
LKSAVRDGIIRYNVADGAIVKSQVAKKSKEEKFLEESDYLNLIEVSKDKIKYASYFTVYLIAVTGLRFA
EVQGLTWNDVDFDNGFLDINKSFDYSISQRFAPTKNEQSIRKVPIDLNTIDILKEYKDNYYQPNKLGRI
CYGASNNATNKAIKLTTGKPYPTNHTLRHTYASYLIMQGVDLISISQLLGHENLNITLKVYAHQLDKL
KEKNDKVIKDIFYNL

GBS*Int*2.2

MAFYRKLGSGWEYRITYRDSQGKKREKSKRGFKTKTLAKVAAQQAEIDLNTMTADLLDITVLDYNR
RWADIYKKPHITAKTWQTYTKNFKHIEHYFGTRKLKSITHTFYQQVLNDFGEKVAQQTLDKFHYQIK
GACKMAIRDGIIRDNFADGAIVRSQKPVKEESEKFMEESEYLTFIKVAKSKVKYPSYLTTYIIAVTGLRF
AEVQGLTWKDIDFDNGYIDINKTFDYSISQNFGPTKNEQSIRKVPIDKNSLELLRNFKSNYYQDNKLDR
ICFGASNNATNKVIKRVTGRNLTNHSLRHTYASYLIAQGVDLISVSKLLGHENLNITLKVYAHQIESLK
EKNDHQVKNIFQNLKFDG

GBS*Int*3

MRYKTMWIEELANGKFKYIERYTDPLTNKYKKVSVTLDKNSSQAQKKAGLILQEKIEDRLAIRNHSEM

TYGELKKEYLKQWIPTVKDSTKRGYLVSDSHIATVLPDDTIINKLTKRDIRLIIDKLLKHNSYHVTHKC
RKRLHAIFSYAIQMDYMTSNPTENVLVPKPKDDYKPEKVLYLTSNEVYDLCNRMIDNDEQTLADIVL
FMFLTGVRYGELACLTYDKIDFENKEILINATYDFNTREITTTKTKKSTRKISVSDNILDIVNKQKKTSS
FVFPNSNGVPILNAYINKRLKIYGDYHTHLFRHSHISFLAEKGIPLNAIMDRVGHSDPKTTLSIYSHTTV
NMKEIINKQTAPFVPFLKPE

GBS*Int*4

MRQKSMWSEKHKSGKVNFVERYKDPYTNKWKRTSVLMEKDTPRIRKEAQRILEAKIADIVRKLQTSD
MLFTNLIDEWWIFYQQEIKRSSIVTLKGNIREIRAEFGINIPVVNIDPRYVQNYLDNLDCSRNKKERNKS
MLNLIFDYAVSLDIIKDNPARRAKLPKIKKTLNDWKKIEEKYLEEEEIKRLLKELFRRPSTRRLGLLSEF
MSLNGCRIGEAISIEPDNIDFKNKTLQLHGTYDRTNGYINGEKTSPKTLASYRETIMTKREMEIIQELEF
INELEKNTNPRYRDMGYIFTTRNGVPIQINSFNLALKKANERLEQPINKNITSHIFRHTLVSRLAENNVP
LKAIMDRVGHADAKTTVQIYTHITKKMKSNIADIMENY

GBS*Int*5

MKDKIITQVVSIMAEQLTMEQLEQLERVLAANLANVVMTENVSKVDETSNPKLLHLFISAKRIEGCSE
KSLKYYKMVIEKMVAELDKPIRQISTSDLRTYLANYQKERQSSKVTIDNMRRIFSSFFSWLEDEDYILK
SPVRRIHKIKTDKVIKETLSDESLELLRDTCDNIRDLAMIDLLASTGMRVGELVRLNREDINFHERECLV
FGKGNSERIVYFDARTKIHLINYLDSRKDDSSALFVSLAYPYDRLMIGGVETRLREIGKRANLQKVHPH
KFRRTLATRAIDKGMPIEQVQHLLGHVKIDTTMHYAMVNQANVKNSHRKYIG

GBS*Int*6.1

MRIESYKKKNGTTAYKFLLYAGYVDGKRKYIRRSGFSTRQSARAALINLQAELEKPKSSMTFGMLTK
QWLKEYEKTVQGSTYLKTERNINKHILPKLDKVTIGDINPLLVQNLTEEWCSQLKYGGKILGLVRNILN
LAVRYGYISNNPALPITAPKIKRERKTGNNFYTLNQLKQFLELVEKTDNIEKIALFRLLAFTGIRKGELL
ALTWDDLNRNTLSINKAVTRTQTGLEIDVTKTKSSDRLISLDDETLEILQQLHETFPSSTFMFQSESGGI
MTPSLPRKWLLQIIKGTDLPQITVHGFRHTHASLLFESGLSLKQVQHRLGHGDLQTTMNVYTHITQSA
IDDIGTKFNQFVTNKQLN

GBS*Int*6.2

MQIESYQKKNGTTAYRFRIYIGVIDGKKKYIKRSGFTSKKIAKQALMNLQQEIENPESKSTMLFHELTN
LWLNNYEKTVQSSTYLKTKRNIENHILPSLGNYPIKDLTPLIIQKYADEWAVKLKYSSKIVGTVRNILN
YAVKFQYIPSNPSDPVTTPKIKRTINKKKDYYNKDELKEFMQLVYDTDNIDIIATFRLLAFTGLRKGEM
LALTWKDYRNGTIDVNKAIARDITGEYVGPTKNKSSERLISLDPETINILDELHETYPKTKYILESTAGR
WISPTQPRRWLLQILSNSKSRLEPIRIHGFRHTHASLLFESGLTLKQVQYRLGHEDLKTTMNTYVHITES
AKDDIGTKFSQYIDF

*No GBS*Int*7 is listed here as the integrase in insertion site GBS7 is the same as GBS*Int*1.

GBS*Int*8

MWHEEQANGNIKFIEYYKDPYTGKRQRAYVTLDRYTKQSETKARRLLNEIIECRIKSSGDQFVRFGQL
VEEWKTSHSKTVKARTMKVYRHPIEKIKDFIGDDVLVKNIDARLLQKFIDYLKDRYSDNTINLIKQPLN
MMLNYAVRMEYIMSNPMKNVVTPKRKKMSKKQFEDKYLETEQNQKIIEQLRDPIYGNHIANFSEIIFL
TGMRPGELLALRWDHIDFEKLKIKIEYTLDYTTNGHANAELGSVKNDGSYRTIDIPLRVKEMLVEELN
YQNTNDLRSDFVFITNKGKHLSINTINRRIKKTSEKLYGIVITSHSFRHAHITLLAELGIPLKSIMDRVGH
TDVNTTIKVYTHATDKIGKQMMDKINKFVPIQSL

GBS*Int*9.1

MASYRKRENGLWEYRISYKTIDGKYKRKEKGGFKTKKLAQAAAIEIEKKLTQNILTNDEVTLYDFVK
TWSEVYKRPYVKDKTWETYSKNFKHIKNYFQELKVKDITPLYYQKKLNEFGEKYAQETLEKFHYQIK
GAMKVAVREQVVTFNFAEGAKVKSQVEPKNEEEDFLEEREYKALLALTRENIQYVSYFTLYLLAVTG
LRFSEAMGLTWSDIDFKNGILDINKSFDYSNTQDFADLKNESSKRKVPIDSNTIDILREYKKNHWQANI
KNRVCFGVSNSACNKLIKKIVGRKVRNHSLRHTYASFLILNGVDIVTISKLLGHESPDITLKVYTHQME
ALAERNFEKIKNIFLVA

GBS*Int*9.2

MAYYRKRDNGWEYRISYKDESGKFRQKSKSGFKTKKLAQAAARDIEKKLSQNILTDGEVTLYDFVKT
WSEVYKRPYVKDKTWETYTKNFRHIKTYFKDIKVKDITPLYYQKRLNEFGEKYAQETLEKFHYQIKG
AMKVAVREQVIHFNFADDAKVKSQIESRAEENDFLEESEYKALLSLTRENIQYVSYFTLYLLSVTGLR
FSEVMGLTWNDVDFKNGILDINKAFDYSNTQDFCDLKNNPSERKVPIDRKTIEILYVYRQNYWQANIK
NRICFGVSNSACNKLIKKIIGRPVRNHTLRHTYASFLILNGVDIVTISKLLGHESPDITLKVYSHQMEALA

ERNFEKIKNIFLAS

GBS*Int*10

MRKVAIYSRVSTINQAEEGYSITGQIDSLTKYCDAMGWVIYKNYSDAGYSGGKLERPAISELIEDGKN
NKFDTVLVYKLDRLSRNVKDTLYLIKDIFTKNNIHFVSIKENIDTSSAMGNLFLTLLSAIAEFEREQIKER
MQFGVMNRAKSGKTTAWKTPPYGYTYDKENKVLLLNEFEATNVKQIFNMIVAGHSIMSITNYAKEH
FAGNTWTHVKIRRILENETYKGLVKYREQTFAGNHDAIIDEELFTKAQLALDKRTNSQNNTRPFQGKY
MLSHIAKCGYCGAPLKVCTGRPRVDGTRRQTYVCVNKTESGAKRGVNNYNNNKVCNSGRYEKSCV
EKYVINELSKIQHDKEYLEKMKNNSKKVDVSSLKKEIKSIDKKINRLNDLYVNDFISLSKLTEEIKKLN
KLKEGYHKTIKLNYVENKNEDVISTLVNNIDISKSSYDVQSRIVKQLVDRVEVTTDNIDIIFNF

GBS*Int*11.1

MNKVAIYVRVSTTMQAEEGYSIDEQIDKLKSYCKIKDWTVYDIYKDGGFSGGNIERPAMERLISDAKR
KKFDTVLVYKLDRLSRSQKDTLFLIEEVFDKNDISFLSLNESFDTSTAFGKAMIGILSVFAQLEREQIKE
RMLLGKIGRAKTGKSMMFSKVSFGYTYDKLKDELVVNQAESIIVRKIFDAYLGGLSLNKLRDYLNNN
GIYRGDKPWNYQGLRRILSNPVYIGMIRYREEIYPGNHKAIIDIDDYNKTQEEIKKRQIKALEFSNNPRP
FRSKYMLSGIAKCGYCGTPLQIILGSKRKDGTRNMRYQCINRFPRNTKGVTIYNDGKKCESGFYEKAD
IEEFVINEIRSLQINYNKLDAMFDRHPTVNSDDIKKQIITLDNKLKRLNDLYINNMIELDDLKKQTQSLR
KQKTILEDELLNNPAITQEKNKKHFKEMLATKDITKLDYETQKNIVNNLINKVFVKSGYIKIEWKIPFK
KA

GBS*Int*11.2

MITTNKVAIYVRVSTTNQAEEGYSIEEQKDKLKSYCNIKDWNVFNVYTDGGFSGSNTERPALEQLIKD
AKKKKFDTVLVYKLDRLSRSQKDTLYLIEDIFLENNIDFVSLLENFDTSTPFGKAMVGILSVFAQLERE
QIKERMQLGKLGRAKAGKSMMWAKVAYGYTYHKGSGEMTINELEAIVVREIFNSYLEGMSITKLRD
KINDTYPKTPAWSYRIIRQILDNPVYCGYNQYKGEVYKGNHEPIISEEDFNKTQDELKIRQRTAAEKFN
PRPFQAKYMLSGIAQCGYCKAPLKIIMGAVRKDGTRFIKYECYQRHPRTTRGVTTYNNNQKCHSSSYY
KQDVEDYVLREISKLQNDKKAIDELFENTNMDTIDRESIKKQIEAISSKIKRLNDLYIDDRITIDELRKKS
TEFTLSKTFLKEKLENDPILKQQESKDNIKKILSCDDILTMDYDQQKIIVKGLINKVQVTADKVIIKWKI

GBS*Int*11.3

MYIEELDDGKYKFIERYIDPLTGKKKRTSVTLDRKTKQAENKARSILQNRISKKINNVTKVELTYGELR
QEYLKQWLPTVKNNTIKNNTRYDEYISYLLDDDVLISNITKATIRNIANELGDKKSYNVVSKCMKRLS
AILNYAASLDYIQSNPAKSVKVIKPVENYDADEKIEFLTIDEMRELYVQMTSKSNKIRDLVVFMFLTGM
RYGEVVALTTDKIDFENKTIKINATYDYDGKELTTPKTENSVRVISVSDSILSIVNDFIVHNRMNNLITD
HIFVSRYGNPMSIRYVNKRLKDFMPEKQLKTHVFRHSHISYLAEKNVPLKAIMDRVGHKNAETTLKIY
THTTNNMKEYINGQTNINF

GBS*Int*12

MKYTKTKYPNIYFYETAKGKRYYVRRSFFFQGKKKEITKSGLSTIPQARAALTEIERQINEQELGINTQ
LTVDQYWEIFSAKRLSTGRWHESSYYLYDSMYRNHIKDEFGFVKLKNLDRNGYEVFIAEKLKKHTRH
TVHTINSSFMAILNDAVKNGNLAGNRLKGVYIGESAIPANNKKITLEQFKEWMDKAKEIMPKKFYALT
YMTIFGLRRGEVFGLRPMDVTKNEHGRAVLKLKDSRSNRTLNGKGSLKTKDSERYVCLDDVGTDYID
YLIDEADRIKRSLGIIKEQKKDYLSINEKGLLINPNQMNKHFGLVSEAIGIHVTPHMMRHFFTTQSIIAGV
PMEQLSQALGHTKIYMTDRYNQVEDELAEATTDMFLTRIR

PICI1*Int*

MERFMIMKITEVKKKDGTVIYRASIYLGTDKVTGKKVTTKITGRTKKEVREKAKQEAIEFIKNGSTRFK
ATSITSYQELATLWWDSYKHTVKYNTQLATEKLLTVHVIPIFGAYKLDKLTTPLIQSIINKLADKTNKG
ERKAYLHYDRIHALNKRILQYGVIMQAIPFNPAREVILPRNTKKANTKRVKHFENDELRTFFNYLNNL
DKNKYRYFYEVTLYKFLLATGCRINEALALNWSDIDLDNAVVHITKTLNYKQEINSPKSKSSYRDIDID
SRTVTMLKQYRRRQIQEAWKLGRSETVVFSDFIHKYPNNRTLQTRLRTHFKRANVSNIGFHGFRHTHA
SLLLNTGIPYKELQYRLGHSTLSMTMDIYSHLSKENAKKAVSFFETAINSI

PICI2*Int*

MERFMIMKITEVKKKNGATVYRASIYLGVDQVTGKKVKPKVTGRTHKEVKQKANQEKIAFQKDGYT
RFKATSIASYQELSNLWWESYKHTVKPNTQDNVKKLLDNHVIPLFGVYKLDKLTTPLIQSIVNKLADK
TNKGEPGAYLHYDKIHALNKRILQYGVTMQAISSNPARDVVLPRNTQKAKRKKVKHFENQDLKKFLD
YLGGLDLSKYRNLYEATLYKFLLATGCRINEALALSWSDIDLENATISITKTLNHLGQINSPKSKAIYRD
IDIDQATITMLKAYQLRQIQEAWKLGRTETVVFSDFIHDYPNNKTLGTRLKTRFKRAGVPNIGFHGFRH
THASLLLNSGIPYKELQYRLGHSTLSMTMDIYSHLSKENAKKAVSFYETALKAL

Appendix B

Supporting information Chapter 3

B.1 Tables and figures

Table B.1: Metadata and results of genomic analyses for 120 high-quality group B *Streptococcus* isolates from dairy cattle in Sweden. Columns show isolate ID, accession number, year of isolation, serotype (from whole genome sequencing), clonal complex (CC), sequence type (ST), antimicrobial resistance genes (AMR), presence and type of integrative conjugative elements (ICE) and lactose operon (Lac.2), Lac.2 insertion site, presence and type of plasmids, long-read MinION sequencing and farm of origin.

ISOLATE	ACCESSION	YEAR	SEROTYPE	CC	ST	AMR	ICE	LAC.2	PLASMIDS	MINION	FARM
MRI Z2-138	ERS1796664	2010	NT	12	10			Lac.2a			74
MRI Z2-139	ERS1796665	2010	III	23	722	*tet*(M)	Tn*916*	Lac.2a, Lac.2c			29
MRI Z2-140	ERS1796666	2010	IV	1	196	*tet*(M)	Tn*916*	Lac.2b			47
MRI Z2-141	ERS1796672	2010	V	1	1	*tet*(M)	Tn*916*	Lac.2b			10
MRI Z2-142	ERS1796673	2010	Ia	103/314	103			Lac.2b			106
MRI Z2-143	ERS1796674	2010	Ia	12	724			Lac.2c			19
MRI Z2-144	ERS1796680	2010	Ia	103/314	314	*tet*(M)	Tn*5801*	Lac.2b			59
MRI Z2-145	ERS1796681	2010	II	12	10			Lac.2a			49
MRI Z2-146	ERS1796682	2010	V	1	1	*tet*(M)	Tn*916*	Lac.2b			75
MRI Z2-147	ERS1796688	2010	III	23	723	*lsa*(C)		Lac.2c			45
MRI Z2-148	ERS1796689	2010	IV	1	1384	*tet*(M)	Tn*916*	Lac.2b			24
MRI Z2-149	ERS1796690	2010	IV	1	196			Lac.2b		yes	13
MRI Z2-150	ERS1796696	2010	Ia	103/314	103			Lac.2b			38
MRI Z2-152	ERS1796697	2011	NT	1	1387	*tet*(M)	Tn*5801*-like	Lac.2b			23
MRI Z2-153	ERS1796698	2011	Ia	103/314	103			Lac.2b			79
MRI Z2-154	ERS1796704	2011	V	1	1	*tet*(M)	Tn*916*	Lac.2d			50
MRI Z2-155	ERS1796705	2011	III	23	23			Lac.2c			105
MRI Z2-156	ERS1796706	2011	IV	1	726	*tet*(M)	Tn*916*	Lac.2b			72
MRI Z2-157	ERS1796712	2011	Ia	103/314	727	*tet*(M)	Tn*5801*	Lac.2c			20
MRI Z2-158	ERS1796713	2011	Ia	23	23			Lac.2d		yes	63
MRI Z2-159	ERS1796720	2011	Ia	103/314	103			Lac.2a			67
MRI Z2-160	ERS1796721	2011	III	23	23			Lac.2c			39
MRI Z2-161	ERS1796722	2011	II	12	10			Lac.2c			5

Table B.1 continued from previous page

ISOLATE	ACCESSION	YEAR	SEROTYPE	CC	ST	AMR	ICE	LAC.2	PLASMIDS	MINION	FARM
MRI Z2-162	ERS1796728	2011	NT	23	23			Lac.2c			26
MRI Z2-163	ERS1796729	2011	V	1	1	*tet*(M)	Tn*916*	Lac.2c			15
MRI Z2-164	ERS1796730	2011	V	1	1	*tet*(M)	Tn*916*	Lac.2b			37
MRI Z2-165	ERS1796752	2011	Ia	103/314	103			Lac.2b			85
MRI Z2-166	ERS1796753	2012	V	1	1	*tet*(M)	Tn*916*	Lac.2b			65
MRI Z2-167	ERS1796740	2012	V	1	1	*tet*(M)	Tn*916*	Lac.2a			56
MRI Z2-168	ERS1796741	2012	Ia	103/314	314	*tet*(M)	Tn*5801*	Lac.2b			48
MRI Z2-169	ERS1796748	2012	IV	1	1384	*tet*(M)	Tn*916*	Lac.2b			64
MRI Z2-170	ERS1796749	2012	Ia	103/314	103			Lac.2b			86
MRI Z2-171	ERS1796569	2012	Ia	103/314	103			Lac.2b			22
MRI Z2-172	ERS1796570	2012	IV	1	196	*tet*(M)	Tn*916*			yes	104
MRI Z2-173	ERS1796571	2012	Ia	103/314	314	*tet*(M)	Tn*5801*	Lac.2b			11
MRI Z2-174	ERS1796572	2012	Ia	103/314	728	*tet*(M)	Tn*5801*	Lac.2b	pZ2-174	yes	108
MRI Z2-175	ERS1796573	2012	Ia	103/314	103			Lac.2b			52
MRI Z2-177	ERS1796574	2012	V	1	1	*tet*(M)	Tn*916*	Lac.2c			9
MRI Z2-178	ERS1796575	2012	IV	1	726	*tet*(A), *tet*(M)	Tn*916*	Lac.2b			110
MRI Z2-179	ERS1796576	2012	V	1	1			Lac.2b		yes	90
MRI Z2-180	ERS1796577	2012	IV	1	196	*tet*(M)	Tn*916*	Lac.2b			81
MRI Z2-181	ERS1796579	2012	IV	1	726	*tet*(M)	Tn*916*	Lac.2b			73
MRI Z2-182	ERS1796580	2012	Ia	23	23	*tet*(M)	Tn*5801*-like	Lac.2b		yes	18
MRI Z2-183	ERS1796581	2012	V	1	1	*tet*(M)	Tn*916*	Lac.2a			41
MRI Z2-261	ERS1796582	1953	III	61/67	1386			Lac.2a			35
MRI Z2-262	ERS1796583	1953	III	61/67	1386			Lac.2a			34
MRI Z2-263	ERS1796585	1953	III	61/67	1386			Lac.2a			62
MRI Z2-264	ERS1796586	1953	III	23	23			Lac.2c			27
MRI Z2-265	ERS1796587	1954	III	61/67	1516			Lac.2a	pZ2-265	yes	111
MRI Z2-266	ERS1796589	1954	III	23	23			Lac.2c			8
MRI Z2-267	ERS1796590	1954	II	61/67	61			Lac.2a	pZ2-265		7

Table B.1 continued from previous page

ISOLATE	ACCESSION	YEAR	SEROTYPE	CC	ST	AMR	ICE	LAC.2	PLASMIDS	MINION	FARM
MRI Z2-269	ERS1796591	1954	Ib	12	6			Lac.2a			34
MRI Z2-270	ERS1796592	1954	Ib	12	1512			Lac.2a		yes	71
MRI Z2-271	ERS1796593	1954	Ia	23	23			Lac.2c			102
MRI Z2-272	ERS1796594	1954	Ia	23	23			Lac.2c			44
MRI Z2-273	ERS1796595	1955	III	23	23			Lac.2c			28
MRI Z2-274	ERS1796596	1961	Ib	12	1513			Lac.2d			98
MRI Z2-275	ERS1796597	1962	II	297	1510	*lsa*(C)		Lac.2d			70
MRI Z2-276	ERS1796598	1963	III	23	23			Lac.2c			3
MRI Z2-277	ERS1796599	1963	Ib	12	6			Lac.2c			70
MRI Z2-278	ERS1796600	1963	III	23	23			Lac.2c			88
MRI Z2-279	ERS1796601	1963	III	23	23			Lac.2c			99
MRI Z2-280	ERS1796602	1964	Ia	23	23			Lac.2c			4
MRI Z2-281	ERS1796603	1964	III	23	23			Lac.2c			31
MRI Z2-282	ERS1796604	1964	III	23	23			Lac.2c			89
MRI Z2-283	ERS1796605	1964	III	23	23			Lac.2c			6
MRI Z2-284	ERS1796606	1964	III	23	23			Lac.2c			53
MRI Z2-285	ERS1796607	1964	III	23	23			Lac.2c			82
MRI Z2-286	ERS1796609	1964	III	61/67	1386			Lac.2a		yes	84
MRI Z2-287	ERS1796610	1964	Ib	12	12			Lac.2c			1
MRI Z2-288	ERS1796611	1965	Ib	12	6			Lac.2d			84
MRI Z2-289	ERS1796612	1965	III	61/67	1394			Lac.2a	pZ2-265		61
MRI Z2-290	ERS1796614	1965	II	61/67	1515			Lac.2a	pZ2-265		36
MRI Z2-291	ERS1796616	1965	II	61/67	1392			Lac.2d			54
MRI Z2-293	ERS1796617	1966	Ib	12	12			Lac.2c			1
MRI Z2-294	ERS1796618	1966	Ib	61/67	1393			Lac.2d			6
MRI Z2-295	ERS1796619	1966	III	23	23			Lac.2c			93
MRI Z2-296	ERS1796620	1954	II	61/67	1392			Lac.2d			76
MRI Z2-297	ERS1796621	1967	Ia	297	1511	*lsa*(C)		Lac.2a			95

Table B.1 continued from previous page

ISOLATE	ACCESSION	YEAR	SEROTYPE	CC	ST	AMR	ICE	LAC.2	PLASMIDS	MINION	FARM
MRI Z2-298	ERS1796622	1967	Ib	12	6			Lac.2a			42
MRI Z2-299	ERS1796623	1967	II	61/67	1514			Lac.2a	pZ2-265	yes	33
MRI Z2-301	ERS1796624	1969	III	23	23			Lac.2c			94
MRI Z2-302	ERS1796625	1969	III	23	23			Lac.2c			1
MRI Z2-304	ERS1796626	1970	Ia	297	297	*lsa*(C)		Lac.2a		yes	103
MRI Z2-305	ERS1796627	1976	III	23	23			Lac.2c			101
MRI Z2-306	ERS1796628	1976	III	23	1509			Lac.2c			96
MRI Z2-307	ERS1796629	1977	III	23	23			Lac.2c		yes	109
MRI Z2-308	ERS1796630	1978	Ia	28	28			Lac.2a			unknown
MRI Z2-311	ERS1796631	1997	III	12	10			Lac.2a			46
MRI Z2-312	ERS1796632	1997	III	23	23	*lsa*(C)		Lac.2c			80
MRI Z2-313	ERS1796633	1997	III	23	23			Lac.2c			21
MRI Z2-314	ERS1796634	2004	III	23	23	*tet*(K)		Lac.2c			17
MRI Z2-315	ERS1796635	2004	III	23	23			Lac.2b			21
MRI Z2-316	ERS1796636	2004	Ia	23	1507			Lac.2c			57
MRI Z2-317	ERS1796637	2004	Ia	103/314	103			Lac.2b			69
MRI Z2-318	ERS1796638	2004	V	1	1	*tet*(M)	Tn*916*	Lac.2d		yes	16
MRI Z2-319	ERS1796639	2004	V	1	1517	*tet*(M)	Tn*916*	Lac.2b			107
MRI Z2-320	ERS1796640	2004	III	23	23	*lnu*(A)		Lac.2c			107
MRI Z2-321	ERS1796641	2004	IV	1	196	*tet*(M)	Tn*916*	Lac.2b			78
MRI Z2-322	ERS1796642	2004	Ia	103/314	103			Lac.2a		yes	68
MRI Z2-323	ERS1796643	2004	IV	1	1384	*tet*(M)	Tn*916*	Lac.2b			25
MRI Z2-324	ERS1796644	2004	III	23	23			Lac.2c			40
MRI Z2-325	ERS1796645	2004	III	23	23			Lac.2c			87
MRI Z2-326	ERS1796646	2004	III	23	23			Lac.2c			54
MRI Z2-327	ERS1796647	2004	II	12	10			Lac.2a			46
MRI Z2-328	ERS1796648	2005	V	1	1			Lac.2b		yes	58
MRI Z2-329	ERS1796649	2005	V	1	1	*tet*(M), *str*	Tn*916*	Lac.2b		yes	14

Table B.1 continued from previous page

ISOLATE	ACCESSION	YEAR	SEROTYPE	CC	ST	AMR	ICE	LAC.2	PLASMIDS	MINION	FARM
MRI Z2-330	ERS1796650	2005	IV	1	732	*tet*(M)	Tn*916*	Lac.2b			60
MRI Z2-331	ERS1796651	2006	V	1	1	*tet*(M)	Tn*916*	Lac.2b			77
MRI Z2-332	ERS1796652	2006	V	23	1385	*tet*(M)	Tn*916*	Lac.2b, Lac.2c		yes	83
MRI Z2-333	ERS1796653	2006	Ia	103/314	103			Lac.2c			55
MRI Z2-334	ERS1796654	2006	III	23	1508			Lac.2c			32
MRI Z2-335	ERS1796655	2006	Ia	103/314	314	*tet*(M), *cat*(pC221)	Tn*5801*	Lac.2b		yes	12
MRI Z2-336	ERS1796657	2006	Ib	12	8	*tet*(M), *cat*(pC221)	Tn*916*	Lac.2a	pZ2-336, IME	yes	30
MRI Z2-337	ERS1796658	2006	III	23	23			Lac.2c			43
MRI Z2-338	ERS1796659	2006	Ia	103/314	103	*tet*(K)		Lac.2b		yes	97
MRI Z2-339	ERS1796660	2007	IV	1	1384	*tet*(M)	Tn*916*	Lac.2b			25
MRI Z2-340	ERS1796661	2007	III	23	23	*tet*(K)		Lac.2c		yes	107
MRI Z2-342	ERS1796662	2009	II	12	10					yes	46
MRI Z2-343	ERS1796663	2009	V	1	1	*tet*(M), *erm*(B)	Tn*916*	Lac.2d			66

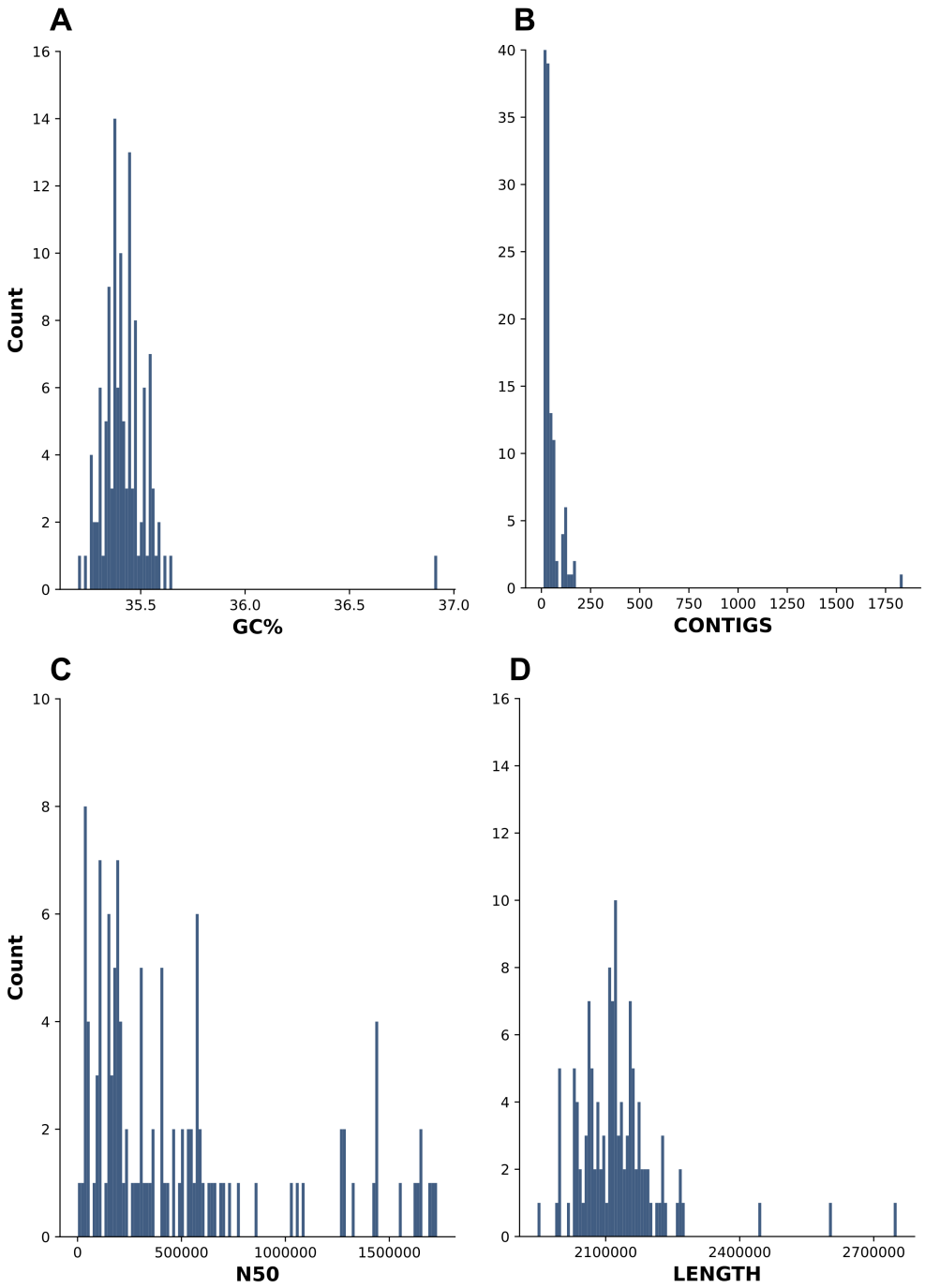


Figure B.1: Distribution, for the 122 group B *Streptococcus* genomes included in this study, of GC content (A), total number of contigs (B), N50 (C) and total genome length (D). Two genomes were excluded from the analyses: genome 23495_7#289 (MRI Z2-151) and 23495_7#366 (MRI Z2-309). The latter had a higher GC content compared to the rest of the dataset (A): 36.92%, with a dataset mean value of 35.42% ±0.32 2SD, and was identified as *Enterococcus thailandicus* based on KmerFinder. The former had a higher number of contigs (n = 1,837) (B) and total genome length (2,751,323 bp) (D) compared to the rest of the dataset, which probably indicates contamination.

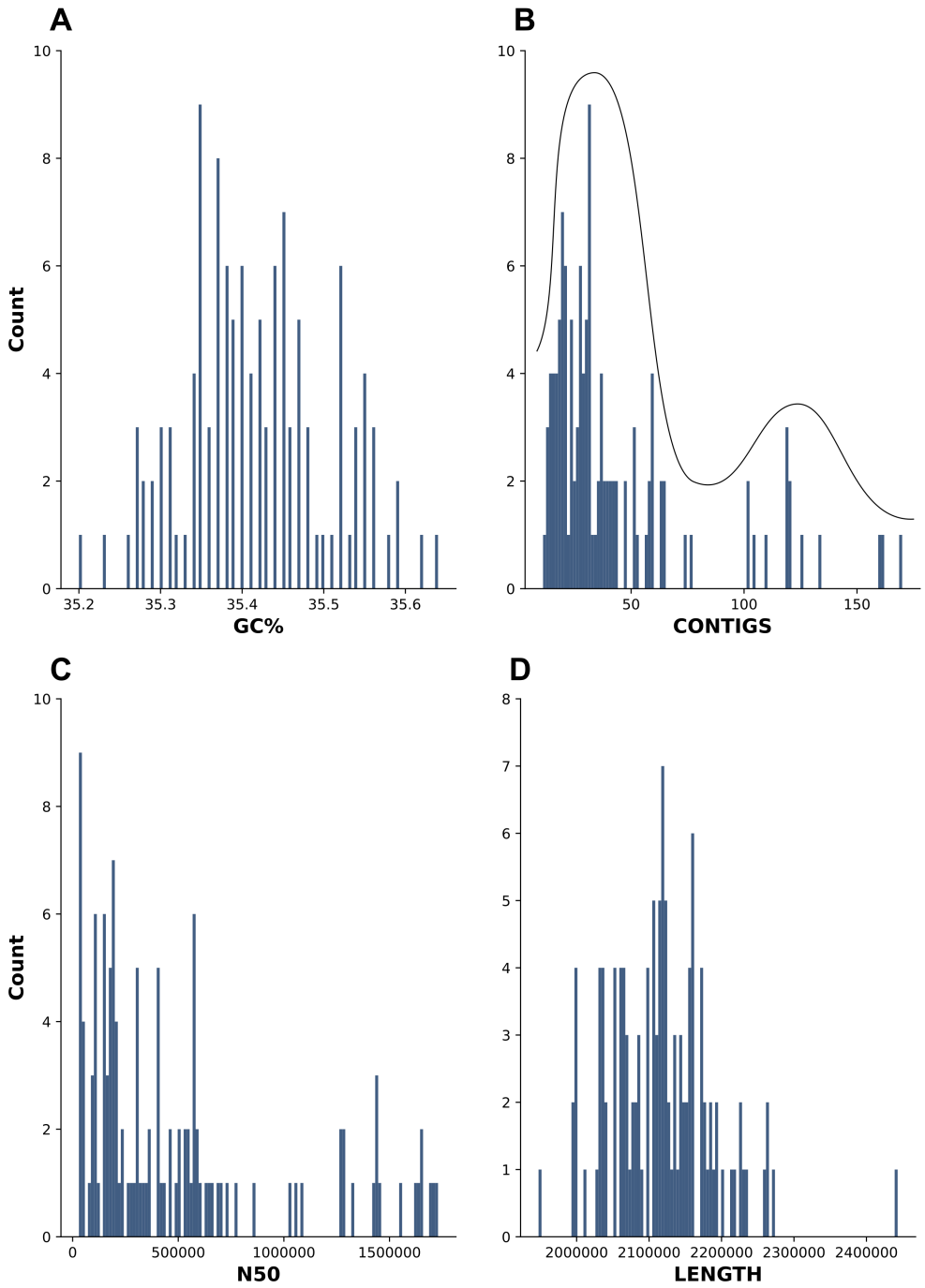


Figure B.2: Distribution, for 120 confirmed, non-contaminated group B *Streptococcus* genomes, of GC content (A), total number of contigs (B), N50 (C) and total genome length (D). Panel B shows a bi-modal distribution This is probably caused by differences in the number of mobile genetic elements, in particular insertion sequences, which incorporate repetitive motifs that impede assembly of short read sequences. Most isolates contributing to the right hand mode belong to bovine-adapted clonal complex 61/67 (n = 11).

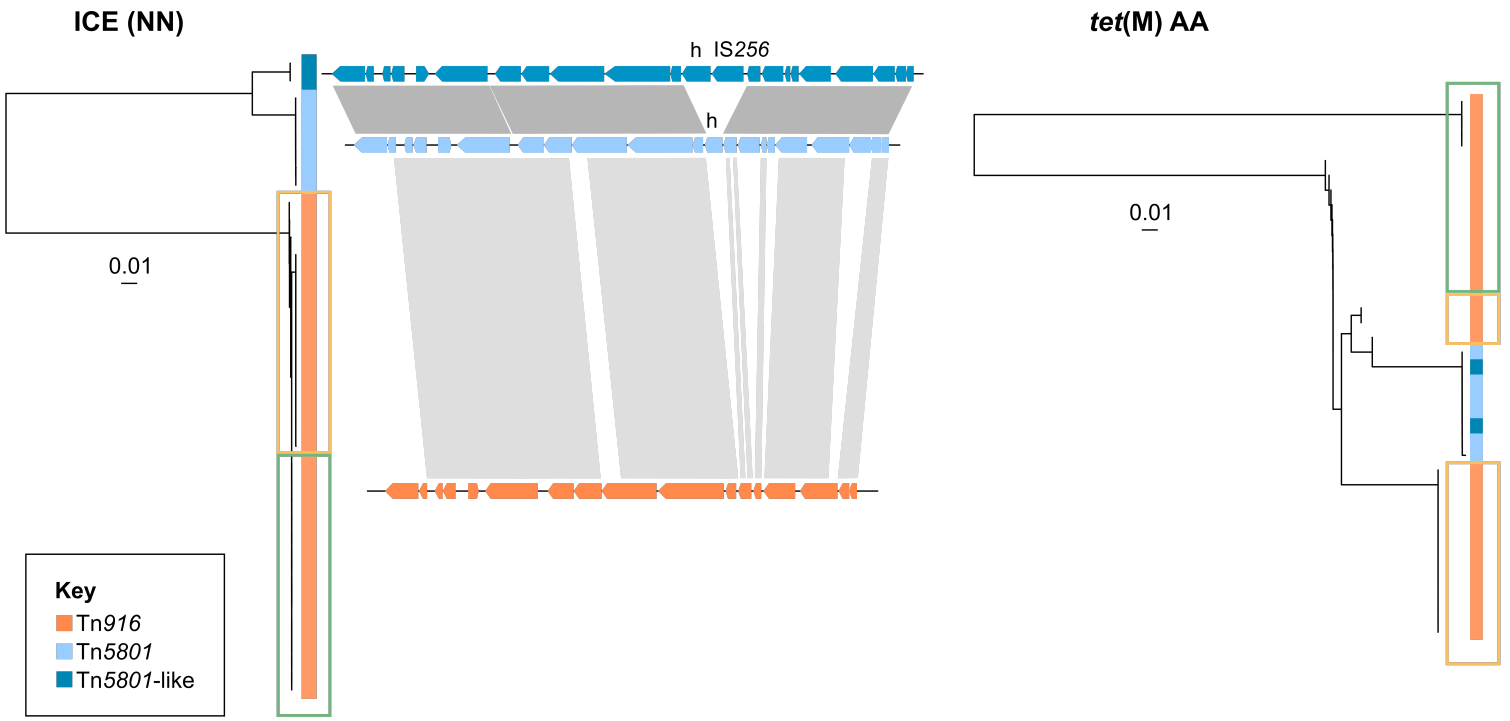


Figure B.3: Comparative Neighbor-Joining phylogenies of nucleotide sequences (NN) of integrative conjugative elements (ICE) (left) and of the *tet*(M) amino acid sequences (AA, right) encoded by *tet*(M) gene contained in those ICE. Among the genomes of 120 group B *Streptococcus* isolates from bovine milk, 37 whole ICE were identified, with gene compositions as visualised with Easyfig v2.2.2 (Sullivan et al., 2011). For sequences derived from Tn*916*, coloured boxes (yellow, green) indicate the distribution of *tet*(M) sequences relative to their ICE sequences. The *tet*(M) amino acid sequence tree shows clustering of sequences contained by Tn*5801* (n=6) or Tn*5801*-like (n=2) ICE within those contained by Tn*916* (n=29). *Tn*5801-like differed from *Tn*5801 because of the substitution of one hypothetical gene with a different hypothetical gene and one IS*256* family transposon in the Tn*5801*-like element, which is observable in the Easyfig comparison. Tree was rooted at midpoint.

B.2 List of commands for bioinformatic analyses

B.2.1 Genome assembly pipeline for short paired-end reads

This pipeline was designed to automate the trimming, filtering and assembling of Illumina paired-end reads. The script concatenates three different sub-scripts, two of which are part of the ConDeTri suite (Smeds & Künstner, 2011) and one of which is SPAdes assembler (Bankevich et al., 2012).

```
#!/bin/sh

# Create a list with input files (which are zipped fastq files)
   for the loop. The list of files is made only with read 1 of
   each pair of reads. Later the sed command will couple read 1
   and read 2.

ls *1.fastq.gz > list

# Create the work folder (where the files will be processed) and
   the output folder (where the files will be stored).

mkdir Work_folder
mkdir Output
mkdir Output/Contigs
mkdir Output/Scaffolds
mkdir Output/Trim_reads

# This is the main loop of the script and it will move each couple
   of paired end reads in the Work_folder before running three
   different scripts (for trimming, filtering and assembling).

while read fast;
do
```

```
# The sed command is used to name the variables.

varzip1=$(echo $fast)
varzip2=$(echo $fast | sed s/_1.fas/_2.fas/)

# Unzipping the couple of files of interest (read 1 and read 2).

   gzip -d $varzip1
   gzip -d $varzip2

# Define 2 new variables for unzipped files. They are var1 = read
   1 and var2 = read 2.

var1=$(echo $varzip1 | sed s/_1.fastq.gz/_1.fastq/ )
var2=$(echo $varzip2 | sed s/_2.fastq.gz/_2.fastq/ )

# Move unzipped files into Work_folder and define new variable
   "pref" that will be used in the scripts. The prefix variable
   will be the ID of the sequence, without any extension.

mv $var1 Work_folder && mv $var2 Work_folder
cd Work_folder
pref=$(echo $var2 | sed s/_2.fastq/""/)

# First script for trimming the reads (ConDeTri).
echo ""
echo "======Starting ConDeTri for $pref..."
echo ""

perl $HOME/anaconda3/envs/assembly/bin/condetri.pl -fastq1=$var1
   -fastq2=$var2 -prefix=$pref -hq=25 -lq=10 -frac=0.8
   -minle\textit{n}=50 -mh=5 -ml=1 -sc=33
```

```
echo ""
echo "======ConDeTri finished for $pref!"
echo ""

# Remove useless files.

rm *stats ; rm *unpaired.fastq

# Define variables for second script.

trim1=$(ls *trim1*)
trim2=$(ls *trim2*)

# Second script for filtering PCR duplicates (FilterPCRdupl).

echo ""
echo "======Starting FilterPCRdupl for $pref..."
echo ""

perl $HOME/anaconda3/envs/assembly/bin/filterPCRdupl.pl
   -fastq1=$trim1 -fastq2=$trim2 -prefix=$pref -cmp=50

echo ""
echo "======FilterPCRdupl finished for $pref!"
echo ""

# Trimmed reads are zipped and moved to the Trim_reads folder.
   Useless files are removed.

rm *hist
gzip -r *trim*
mv *trim* ../Output/Trim_reads
```

```
# Third script for assembling the reads (SPAdes). Run it and
   remove the useless files.

echo ""
echo "======Starting SPAdes assembly of $pref..."
echo ""

python $HOME/anaconda3/envs/assembly/bin/spades.py -t 12 --careful
   --only-assembler -1 *uniq1.fastq -2 *uniq2.fastq -o ./$pref

rm *uniq*

echo ""
echo "======SPAdes assembly of $pref is finished!"
echo ""

# Go into the folder created by the script, extract 2 files
   (scaffolds and contigs), remove the folder with all the other
   files and move the two files in the Output folder.

echo "======Compressing reads $pref and moving files..."
cd $pref
mv contigs.fasta $pref.fasta
mv $pref.fasta ../../Output/Contigs
mv scaffolds.fasta $pref.fasta
mv $pref.fasta ../../Output/Scaffolds && cd ../ && rm -r $pref

# rezip the input files and move them back to the original folder.

gzip -r *.fastq && mv ./* ../
cd ../
done < list
rm list
```

B.2.2 Quality control for genome assembly pipeline

This pipeline was designed to run QUAST (Gurevich et al., 2013) and to evaluate the overall quality of the genome assemblies. When the values for GC% or total number of contigs is higher than the mean plus twice the standard deviation for the dataset, the assembly is moved to a different folder for further inspection and its name is appended to a list of low-quality score files.

It has to be kept in mind that this pipeline is just an indication to investigate further assemblies that differ substantially from the average value of the dataset. As the mean is relative to the dataset, hence not an absolute value, the output does not have to be interpreted as an absolute indication of low-quality.

As an example, during these analyses it was noticed that the CC61/67 bovine-specific clade tends to have a significantly higher number of contigs compared to most CC, when sequencing with short-read technologies and assembling *de novo*. This is not an indication of overall poor quality, rather it is the result of a high number of MGE being harboured by this CC: as MGE contain a lot of repetitive sequences, *de novo* assembly can result in a highly fragmented genome.

```
#!/bin/sh

# Run the script from the directory containing the assembled
   genomes.
# Create directories where to store the output files.

mkdir ./quast_output
mkdir ./quast_output/lowQS

# Run QUAST

python $HOME/anaconda3/envs/assembly/bin/quast -o ./quast_output
   *.fasta

# Take QUAST output transposed_report.tsv (tab separated with
```

```
    columns = parameters, rows = genomes). Remove the first row
    (header).

sed '1d' ./quast_output/transposed_report.tsv >
    ./transposed_report_no_header.tsv

# Define variables and calculate the mean values for GC% and
    number of contigs.

tot_genomes=$(wc -l transposed_report_no_header.tsv | awk '{print
    $1}')
sum_contigs=$(awk '{s+=$14}END{print s}'
    transposed_report_no_header.tsv)
sum_GC=$(awk '{s+=$18}END{print s}'
    transposed_report_no_header.tsv)

mean_contigs=$(echo "scale=2; $sum_contigs / $tot_genomes" | bc)
mean_GC=$(echo "scale=2; $sum_GC / $tot_genomes" | bc)

# Define variables and calculate standard deviation (SD) and 2*SD
    (SD2).

SD_contigs=$(awk '{sum+=$14; sumsq+=$14*$14}END{print
    sqrt(sumsq/NR - (sum/NR)**2)}' transposed_report_no_header.tsv)
SD_GC=$(awk '{sum+=$18; sumsq+=$18*$18}END{print sqrt(sumsq/NR -
    (sum/NR)**2)}' transposed_report_no_header.tsv)

SD2_contigs=$(echo "$SD_contigs * 2" | bc)
SD2_GC=$(echo "$SD_GC * 2" | bc )

# Define variables and calculate the mean+2SD for contigs and GC.

sum_SD2_contigs=$(echo "$mean_contigs + $SD2_contigs" | bc)
```

```
sum_SD2_GC=$(echo "$mean_GC + $SD2_GC" | bc)

rounded_sum_SD2_contigs=$(printf "%.0f\n" "$sum_SD2_contigs")
rounded_sum_SD2_GC=$(printf "%.2f\n" "$sum_SD2_GC")

# If the total number of contigs or the GC% is > than 2SD, print
   the name of the sequences on a list.

awk '{if($14>'$rounded_sum_SD2_contigs' ||
   $18>'$rounded_sum_SD2_GC')print$1".fasta"}'
   transposed_report_no_header.tsv > list_lowQS.txt

# Remove unnecessary files

rm transposed_report_no_header.tsv
sed 's/_/#/2' list_lowQS.txt >list_names.txt

# Move low-quality score sequences from the list to another folder
   for further inspection.

for file in $(<list_names.txt);
do
   mv "$file" ./quast_output/lowQS;
done
```

Appendix C

Supporting information Chapter 4

C.1 Tables and figures

Table C.1: Table of 24 group B *Streptococcus* (GBS) genomes that were excluded from further analyses after the application of a quality control filter (based on reference ranges for total length and GC content calculated as mean ± 2SD, and for total number of contigs calculated as mean + 2SD). Values that fall outside the reference ranges are marked with an asterisk. Results for species confirmation with KmerFinder are shown.

ISOLATE	LENGTH (bp)	GC (%)	CONTIGS	KMERFINDER
SAMEA3888079	4098500*	34.33	1284*	*Staphylococcus* spp./GBS
DK-B-USS-084	546545*	34.99	1	GBS
SG-M450	644585*	35.04	1	GBS
SAMEA2168901	2116897	35.19	288*	GBS
SAMEA2168859	2180609	35.33	301*	GBS
GB001	2706205*	35.36	56	GBS
NA0054832	1952989	35.6	450*	GBS
SA0013062	1967828	35.64	441*	GBS
SA0012037	1936319	35.66	309*	GBS
M19	974690*	35.69	1	GBS
SAMEA3888133	2376281	35.75	450*	GBS/*Legionella* spp.
NA0061863	1908119	35.77	552*	GBS
SA0007345	1887823	35.9	673*	GBS
SA0022621	2801810*	36.16	2050*	*Staphylococcus* spp./GBS
SA0031308	2203434	36.29	1445*	*Staphylococcus* spp./GBS
BG014	5078690*	36.38	82	*Enterococcus* spp./GBS
SAMEA2168915	4841212*	36.4	317*	*Enterococcus* spp.
BG013	3069266*	37.23*	31	*Enterococcus* spp.
WSB3237	7506653*	37.81*	1204*	*Proteus* spp.
NGBS046	189868*	39.06*	5	GBS (low score)

NGBS1084	4953*	39.07*	1	none
NGBS080	6239*	49.53*	1	none
NGBS1074	na	na	na	na (not enough reads)
NA0068263	na	na	na	na (not enough reads)

Table C.2: List of the 850 high-quality group B *Streptococcus* (GBS) genomes included in this work. The following metadata are shown: isolate ID (for six isolates for which information on sample ID was not available, SRR numbers marked with an asterisk have been given instead), country of origin (international acronyms for countries and provinces/states have been used), host (food indicates food market fish samples), origin/clinical manifestation, year of isolation, sequence type (ST), serotype (from whole genome sequencing), genome assembly size in base pairs (bp), GC content (%) and total number of contigs. For origin/clinical manifestation, the following acronyms have been used: AB (adult blood), AID (adult invasive disease), AO (adult other), ASF (adult synovial fluid), AT (adult tissue), BOA (blood older adult), BTM (bulk tank milk), CSF (cerebrospinal fluid), EOD (early onset disease), ID (invasive disease), LOD (late onset disease), ME (meningoencephalitis), NC (neonatal carriage), NI (neonatal invasive), OASF (older adult synovial fluid), OAT (older adult tissue), OOA (other older adult), PA (periarticular abscess) and UTI (urinary tract infection). ST marked with an asterisk are single locus variants (SLV) of the indicated ST (n=9); we were not able to assign new ST to these SLV, as three of them were only available as genome assemblies (i.e. we could not extract the alleles from raw reads) and the remainder did not have sufficient coverage for all alleles to provide reliable ST data.

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
138P	USA	fish, tilapia	kidney	2007	261	Ib	1838701	35.47	1
138spar	USA	fish, tilapia	kidney	2011	261	Ib	1838126	35.47	1
1B13M	DK	bovine			1	V	2093307	35.39	67
2-22	IL	fish, tilapia		1988	261	Ib	1838867	35.47	1
200737_RR15000448	UK, SCT	human	blood		19	III	2153043	35.57	35
200738_RR15000449	UK, SCT	human	blood		196	IV	2136017	35.43	17
200739_RR15000450	UK, SCT	human	blood		23	Ia	2018182	35.2	9
200742_RR15000453	UK, SCT	human	blood		17	III	2050219	35.28	27
200743_RR15000454	UK, SCT	human	blood		104	II	2131285	35.43	29
200746_RR15000457	UK, SCT	human	blood		287	III	2103236	35.38	36
200748_RR15000459	UK, SCT	human	blood		3	II	2083686	35.32	28
221508_RR15000464	UK, SCT	human	blood		1	V	2076859	35.3	12
221514_RR16000008	UK, SCT	human	blood		452	IV	2049498	35.35	24
221515_RR16000009	UK, SCT	human	blood		19	V	2098492	35.32	32
3B11V	DK	bovine			23	III	2135617	35.42	41

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
3H1V	DK	human			23	Ia	2006790	35.21	55
3H3R	DK	human			23	III	2122644	35.41	57
4B14M	DK	bovine			130	IX	2080394	35.41	55
4H1O	DK	human			130	IX	2073413	35.45	88
5B15M	DK	bovine			1	V	2174789	35.42	55
5H2O	DK	human			1	V	2124013	35.4	53
5H4R	DK	human			1409*	II	2138583	35.42	72
5H6R	DK	human			23	Ia	1971688	35.19	47
6B19M	DK	bovine			1	V	2176886	35.41	39
7B18M	DK	bovine			314*	Ia	1967815	35.34	40
7H8O	DK	human			314	Ia	1966293	35.32	16
8B14M	DK	bovine			103	Ia	2047342	35.43	34
8H3O	DK	human			110	V	2153121	35.44	80
BE001	BE	human	carriage	2009	23	Ia	1970900	35.19	15
BE007	BE	human	carriage	2010	315	III	2029273	35.26	28
BG003	BG	human	NC	2009	8	Ib	2251831	35	26
BG005	BG	human	EOD	2009	23	Ia	2036609	35.35	15
BG006	BG	human	NC	2009	12	Ib	2180889	35.45	16
BG010	BG	human	NC	2009	12	II	2054403	35.25	18
BSE009	CN	human	amniotic fluid	2014	485	Ia	2148637	35.38	1
BSE010	CN	human	vaginitis	2014	485	Ia	2174543	35.51	37
BSU133	DE	human			6	Ib	2030979	35.24	16
BSU165	DE	human			28	II	2154567	35.29	19
BSU174	DE	human			41	V	1991831	35.28	27
BSU178	DE	human			7	Ia	2152788	35.4	57
BSU188	DE	human			23	Ia	2238435	35.29	80
BSU247	DE	human			26	V	2096318	35.18	11
BSU248	DE	human			12	Ib	2099392	35.13	15

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
BSU252	DE	human			1	V	1986507	35.11	14
BSU260	DE	human			88	Ia	2177888	35.31	21
BSU442	DE	human			22	II	2048445	35.29	83
BSU447	DE	human			19	III	2109784	35.52	37
BSU450	DE	human			10	Ib	2004235	35.19	31
BSU451	DE	human			103	Ia	2081745	35.44	9
BSU454	DE	human			8	Ib	2073083	35.35	21
BSU92	DE	human			196	IV	2061829	35.27	13
BSU96	DE	human			17	III	2070997	35.22	28
C001	CN	bovine	mastits	2013	103	Ia	2121372	35.65	1
CCH210801006	CA, ON	human	carriage	2010	23	Ia	2033594	35.17	32
CCUG_17336	SE	human			17	III	2029975	35.32	27
CCUG_19094	SE	human			19	III	2157615	35.29	42
CCUG_24810	SE	human	ear	1989	19	III	2135075	35.33	35
CCUG_25532	SE	human			26	V	2083826	35.36	14
CCUG_28551	SE	human			2	IV	2126496	35.28	28
CCUG_29376	SE	human	NI	1991	12	Ib	2181955	35.37	19
CCUG_30636	SE	human			1	V	2055665	35.29	20
CCUG_34230	USA, IA	bovine			23	III	2152364	35.47	11
CCUG_37430	SE	human			19	II	2093150	35.2	17
CCUG_37739	SE	human			23	Ia	1980127	35.17	12
CCUG_37740	SE	human	joint aspirate	1996	1	V	2133307	35.12	15
CCUG_38383	SE	human		1997	23	Ia	2050211	35.22	25
CCUG_39096_A	SE	human			9	Ib	2089836	35.19	20
CCUG_44110	SE	human	vagina	1999	88	Ia	2090695	35.32	24
CCUG_45061	NO	human			19	III	2147130	35.26	35
CCUG_49072	SE	human			524	V	2057457	35.29	16
CCUG_91	SE	human	CSF	1968	28	II	2114808	35.34	22

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
CU_GBS_08	HK	human	blood	2008	283	III	2084511	35.4	1
CU_GBS_98	HK	human	CSF	1998	283	III	2029669	35.38	1
CUGBS329	HK	human	joint aspirate	2002	283	III	2047867	35.21	24
CUGBS459	HK	human	joint aspirate	2003	19	III	2104810	35.53	34
CUGBS522	HK	human	blood	2005	283	III	2008949	35.17	23
CUGBS550	HK	human	blood	2006	283	III	2045093	35.23	17
CUGBS568	HK	human	blood	2007	283	III	2057520	35.24	19
CUGBS58	HK	human	blood	1998	283	III	2048736	35.21	26
CUGBS581	HK	human		2007	283	III	2060546	35.25	23
CUGBS587	HK	human	joint aspirate	2008	283	III	2029813	35.18	27
CZ002	CZ	human	NC	2008	19	III	2188010	35.52	36
CZ007	CZ	human	NC	2008	255	Ib	1991817	35.2	16
CZ009	CZ	human	NC	2008	1	V	2056436	35.29	15
CZ010	CZ	human	NC	2009	10	II	2091588	35.21	44
CZ013	CZ	human	NC	2009	1	V	2062023	35.29	14
CZ015	CZ	human	EOD	2009	1	V	2061168	35.29	14
DE012	DE	human	EOD	2009	88	Ia	2108481	35.28	9
DE033	DE	human	EOD	2010	23	Ia	2043187	35.17	14
DK004	DK	human	EOD	2009	17	III	2116330	35.49	32
DK005	DK	human	EOD	2009	23	Ia	1929726	35.27	16
DK008	DK	human	NC	2009	9	Ib	2054892	35.18	20
DK012	DK	human	EOD	2010	1	V	2051086	35.32	15
DK013	DK	human	NC	2010	1	V	2066647	35.28	17
DK014	DK	human	NC	2009	88	Ia	2102239	35.38	12
DK015	DK	human	NC	2008	10	V	2053782	35.23	21
DK022	DK	human	EOD	2011	1	V	2121886	35.39	13
FDAARGOS_254	USA	human		2014	22	II	2218541	35.75	1
FSL_C1-494	USA, NY	bovine	mastitis	2000	NF	II	2213073	35.49	80

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
FSL_F2-343	USA, NY	human	blood	2000	88	Ia	2132991	35.39	12
FSL_S3-001	USA, NY	human	blood	2000	1	V	2057459	35.3	16
FSL_S3-003	USA, NY	human	blood	2000	19	III	2095013	35.22	25
FSL_S3-005	USA, NY	human	blood	2000	22	II	2041668	35.29	73
FSL_S3-014	USA, NY	human	blood	2000	8	Ib	2142797	35.03	20
FSL_S3-023	USA, NY	human	blood	2000	1	V	2036310	35.28	15
FSL_S3-027	USA, NY	bovine	mastitis	2000	1617	II	2200370	35.42	138
FSL_S3-043	USA, NY	bovine	mastitis	2000	61	NT	2174085	35.58	151
FSL_S3-062	USA, NY	bovine	mastitis	2000	61	II	2227709	35.59	107
FSL_S3-077	USA, NY	bovine	mastitis	2000	415	II	2484758	35.44	217
FSL_S3-090	USA, NY	human	blood	2000	23	Ia	2076864	35.25	13
FSL_S3-102	USA, NY	human	blood	2000	31	III	1975219	35.19	31
FSL_S3-105	USA, NY	bovine	mastitis	2000	NF	III	2252156	35.65	79
FSL_S3-128	USA, NY	bovine	mastitis		1618	II	2234608	35.48	113
FSL_S3-170	USA, NY	bovine	mastitis	2000	1612	II	2260771	35.43	95
FSL_S3-222	USA, NY	bovine	mastitis	2000	1618	II	2234608	35.48	113
FSL_S3-229	USA, NY	bovine	mastitis	2000	415	II	2229321	35.43	104
FSL_S3-251	USA, NY	bovine	mastitis	2000	61	III	2305144	35.56	106
FSL_S3-268	USA, NY	human	blood	2000	22	II	2021871	35.26	65
FSL_S3-277	USA, NY	bovine	mastitis	2000	1621	III	2291898	35.74	82
FSL_S3-337	USA, NY	human	blood	2000	19	III	2095697	35.22	28
FSL_S3-501	USA, NY	bovine	mastitis	2001	1615	II	2211996	35.65	108
FSL_S3-568	USA, NY	bovine	mastitis	2001	415*	II	2225300	35.7	102
FSL_S3-586	USA, NY	bovine	mastitis	2001	67	II	2463419	35.51	103
FSL_S3-603	USA, NY	bovine	mastitis	2001	61	III	2217465	35.54	92
FSL_S3-608	USA, NY	bovine	mastitis	2001	490	II	2329351	35.48	101
FSL_S3-654	USA, NY	bovine	mastitis	2001	61	II	2220573	35.52	92
FWL1402	CN	frog		2014	739	III	2090292	35.44	1

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
GB00002	CA, AB	human			23	Ia	1972421	35.19	13
GB00018	CA, AB	human			444	Ia	1971658	35.18	19
GB00084	CA, AB	human	carriage		1	VIII	2038722	35.33	18
GB00111	CA, AB	human			32	III	2069293	35.28	31
GB00174	CA, AB	human			22	II	1991136	35.25	82
GB002	UK	human	EOD	2009	130	IX	2155495	35.5	37
GB00226	CA, AB	human			28	II	2102171	35.26	21
GB00535	CA, AB	human			8	Ib	2020398	35.2	18
GB00543	CA, AB	human			36	III	2100533	35.24	35
GB00548	CA, AB	human			88	Ia	2038909	35.28	11
GB00555	CA, AB	human	carriage		12	Ib	2078428	35.16	18
GB00561	CA, AB	human	carriage		19	V	2156242	35.46	39
GB00588	CA, AB	human			447	II	2059321	35.21	17
GB00601	CA, AB	human			24	Ia	2038109	35.22	11
GB00614	CA, AB	human			448	NT	2144698	35.25	32
GB007	UK	human	EOD	2010	19	III	2072450	35.23	38
GB00900	USA, MI	human			19	III	2174339	35.56	43
GB00919	USA, MI	human			12	II	2152132	35.2	34
GB00924	USA, MI	human			1	V	2060481	35.28	18
GB00940	USA, MI	human			17	III	1977355	35.22	32
GB00951	USA, MI	human			28	II	2064056	35.31	21
GB00954	USA, MI	human			22	II	2072897	35.24	113
GB00957	USA, MI	human			23	Ia	1982498	35.17	19
GB00965	USA, MI	human			88	Ia	2093731	35.35	46
GB010	UK	human	NC	2010	1	V	2104120	35.29	15
GBS ST-1	USA, CA	dog		2015	1	V	2108125	35.14	45
GBS-M002	CN	human	cervix	2014	1	VI	2092570	35.58	1
GBS1-NY	USA	human	blood	2012	22	II	2243708	35.93	1

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
GBS2-NM	USA	human		2012	22	II	2214297	35.89	1
GBS6	USA	human		2009	22	II	2231475	35.75	1
GBS85147	BR	human			103	Ia	1996151	35.49	1
GD201008-001	CN	fish, tilapia	brain	2010	7	Ia	2063112	35.65	1
GS16-0008	GH	fish, tilapia		2016	261	Ib	1801542	35.26	21
GX026	CN	fish, tilapia	brain	2011	261	Ib	1840649	35.49	1
H002	CN	human	vagina	2011	736	III	2147416	35.65	1
H0401	VT	human	vaginal fluid	2019	651	III	1998024	35.11	18
H0402	VT	human	sputum	2019	19	V	2156518	35.35	34
H0403	VT	human	amniotic fluid	2019	1	V	2060341	35.31	19
H0404	VT	human	urine	2019	19	V	2145274	35.38	37
H0405	VT	human	vaginal fluid	2019	19	V	2244078	35.48	63
H0406	VT	human	vaginal fluid	2019	485	Ia	2048394	35.31	18
H0407	VT	human	pus	2019	1	V	2138807	35.28	19
H0408	VT	human	blood	2019	862	III	2055234	35.22	23
HN016	CN	fish, tilapia	brain	2011	7	Ia	2064722	35.66	1
HN016	CN, GZAR	fish, tilapia	brain	2011	7	Ia	2000956	35.33	36
ILRI005	KY	camel	PA	2004	609	V	2109759	35.43	1
ILRI025	SO	camel	carriage		610	VI	2013384	35.25	29
ILRI030	SO	camel	carriage		617	VI	1999626	35.3	30
ILRI037	KY	camel	gingivitis	2004	612	IV	2020002	35.25	34
ILRI054	KY	camel	wound infection	2001	615	II	2021031	35.21	58
ILRI067	KY	camel	carriage	2003	614	V	1980133	35.12	52
ILRI112	KY	camel	PA	2002	617	VI	2029198	35.34	1
ILRI120	KY	camel	chronic cough	2002	618	IV	2049911	35.31	50
ILRI127	KY	camel	PA	2002	613	IV	1973342	35.26	33
IT007	IT	human	EOD	2008	17	III	2082763	35.46	35
IT016	IT	human	EOD	2009	130	IX	2075841	35.45	33

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
IT018	IT	human	EOD	2009	17	III	2113657	35.24	33
IT028	IT	human	EOD	2009	1	V	2057751	35.3	16
IT036	IT	human	EOD	2010	26	V	2181302	35.37	18
LDS_610	AR	bovine			1616	II	2106806	35.54	79
LDS_617	AR	bovine			1614	NT	2018953	35.51	130
LDS_623	AR	bovine			61	III	2189962	35.62	101
LG01	SG	food, bighead carp		2015	283	III	2071219	35.24	33
LMG_14609	BE	bovine			NF*	Ia	2244390	35.28	103
LMG_14838	BE	bovine			1622	III	2396395	35.33	126
LMG_15081		human			25	Ia	2043730	35.23	14
LMG_15083		human			7	Ia	2075588	35.35	27
LMG_15084	USA	human			19	II	2058284	35.27	15
LMG_15085	USA, TX	human			17	III	1967719	35.19	28
LMG_15089	BE	human			19	III	2155184	35.3	36
LMG_15090	BE	human	UTI	1994	8	Ib	2058852	35.27	21
LMG_15091	BE	human	UTI	1994	1166	IV	2049283	35.36	31
LMG_15092	BE	human			2	II	2012786	35.25	75
LMG_15093	BE	human	vagina	1994	110	V	2129987	35.4	39
LMG_15094	BE	human	UTI	1994	17	III	1965720	35.18	26
LMG_15095	BE	human			17	III	1967987	35.19	27
LZF004	CN	human	cervicitis	2014	103	Ia	2019092	35.37	22
LZF009	CN	human	cervical secretion	2014	485	Ia	2212820	35.26	33
LZF010	CN	human	cervicitis	2014	485	Ia	2041937	35.31	21
MRI Z1-004	DK	bovine	BTM	2009	10	Ib	2077195	35.23	19
MRI Z1-050	DK	bovine	BTM	2009	103	Ia	2119406	35.53	10
MRI Z1-053	DK	bovine	BTM	2009	103	Ia	2089337	35.48	10
MRI Z1-116	DK	bovine	BTM	2009	604	Ib	2025681	35.24	21
MRI Z1-158	DK	bovine	BTM	2009	23	Ia	2027911	35.28	14

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
MRI Z1-300	DK	bovine	BTM	2010	23	Ia	2063706	35.32	11
MRI Z1-354	DK	bovine	BTM	2010	7	V	2208028	35.59	20
MRI Z1-363	DK	bovine	BTM	2010	17	III	2004631	35.27	27
MRI Z1-368	DK	bovine	BTM	2012	8	Ib	2184011	35.41	22
MRI Z1-410	DK	bovine	BTM	2010	8	Ib	2032504	35.19	23
MRI Z1-529	DK	bovine	BTM	2010	8	Ib	2062959	35.29	21
MRI Z1-586	DK	bovine	BTM	2010	7	Ia	2192633	35.49	59
MRI Z1-597	DK	bovine	BTM	2011	19	III	2133138	35.35	30
MRI Z1-600	DK	bovine	BTM	2011	23	Ia	2068007	35.3	10
MRI Z1-602	DK	bovine	BTM	2011	2	II	2036955	35.35	25
MRI Z1-606	DK	bovine	BTM	2011	1	V	2125390	35.38	17
MRI Z1-705	DK	bovine	BTM	2011	2	IV	2086118	35.3	16
MRI Z1-707	DK	bovine	BTM	2011	1	V	2099562	35.36	10
MRI Z1-710	DK	bovine	BTM	2012	314	Ia	2035297	35.37	11
MRI Z1-715	DK	bovine	BTM	2012	103	Ia	2050669	35.45	6
MRI Z1-717	DK	bovine	BTM	2012	19	III	2113686	35.51	36
MRI Z1-719	DK	bovine	BTM	2012	23	III	2028502	35.27	11
MRI Z1-792	DK	bovine	BTM	2012	103	Ia	2016187	35.46	6
MRI Z1-802	DK	bovine	BTM	2012	2	IV	2173928	35.36	17
MRI Z1-803	DK	bovine	BTM	2012	314	Ia	2061813	35.45	7
MRI Z1-808	DK	bovine	BTM	2012	1	VIII	2132444	35.45	17
MRI Z1-811	DK	bovine	BTM	2012	19	III	2133206	35.35	32
MRI Z1-822	DK	bovine	BTM	2012	23	III	2097571	35.49	12
MRI Z1-851	DK	bovine	BTM	2012	196	IV	2233614	35.51	24
MRI Z1-858	DK	bovine	BTM	2012	314	Ia	1978154	35.3	11
MRI Z1-872	DK	bovine	BTM	2012	314	Ia	2034397	35.39	15
MRI Z1-890	DK	bovine	BTM	2012	196	IV	2149403	35.32	14
MRI Z2-001	FI	bovine	mastitis	2010	632	III	2163752	35.63	151

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
MRI Z2-002	FI	bovine	mastitis	2010	632	III	2164178	35.62	160
MRI Z2-007	FI	bovine	milk	2010	103	Ia	1988471	35.28	12
MRI Z2-025	FI	bovine	milk	2010	1	V	2091281	35.34	16
MRI Z2-039	FI	bovine	milk	2011	633	Ia	2053574	35.48	18
MRI Z2-040	FI	bovine	milk	2011	634	V	2091119	35.36	13
MRI Z2-041	FI	bovine	milk	2011	8	Ib	2109075	35.35	19
MRI Z2-044	FI	bovine	milk	2011	103	Ia	2006111	35.29	6
MRI Z2-045	FI	bovine	milk	2011	8	Ib	2124249	35.44	20
MRI Z2-046	FI	bovine	milk	2011	1	VII	2083012	35.45	12
MRI Z2-047	FI	bovine	milk	2011	635	V	2185964	35.54	18
MRI Z2-048	FI	bovine	milk	2011	183	V	2133357	35.35	15
MRI Z2-050	FI	bovine	milk	2011	633	Ia	2044189	35.4	18
MRI Z2-051	FI	bovine	milk	2011	636	V	2128609	35.49	13
MRI Z2-053	FI	bovine	milk	2011	8	Ib	2068617	35.27	18
MRI Z2-054	FI	bovine	milk	2011	2	II	2069711	35.43	28
MRI Z2-058	FI	bovine	milk	2012	10	Ib	2103051	35.27	18
MRI Z2-060	FI	bovine	milk	2012	196	IV	2178056	35.43	25
MRI Z2-062	FI	bovine	milk	2012	1	V	2143230	35.47	15
MRI Z2-064	FI	bovine	milk	2012	633	Ia	2070494	35.46	8
MRI Z2-065	FI	bovine	milk	2012	196	IV	2179205	35.39	23
MRI Z2-068	FI	bovine	milk	2012	10	II	2071418	35.22	31
MRI Z2-077	FI	human	UTI	2012	462	V	2065962	35.31	21
MRI Z2-080	FI	human	UTI	2012	27	III	2194503	35.48	47
MRI Z2-081	FI	human	SSTI	2012	23	Ia	1963332	35.18	16
MRI Z2-082	FI	human	SSTI	2012	8	Ib	2076666	35.12	20
MRI Z2-084	FI	human	UTI	2012	28	II	1989027	35.21	23
MRI Z2-089	FI	human	carriage	2012	1	II	2144609	35.5	17
MRI Z2-093	FI	human	UTI	2012	17	III	2180410	35.39	27

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
MRI Z2-094	FI	human	SSTI	2012	12	II	2184612	35.3	27
MRI Z2-096	FI	human	carriage	2012	23	Ia	2034094	35.19	14
MRI Z2-098	FI	human	carriage	2012	196	IV	2138991	35.39	33
MRI Z2-099	FI	human	carriage	2012	459	IV	2209874	35.37	17
MRI Z2-101	FI	human	SSTI	2012	19	III	2160452	35.51	37
MRI Z2-102	FI	human	SSTI	2012	19	III	2160368	35.51	38
MRI Z2-105	FI	human	SSTI	2012	12	II	2204559	35.47	30
MRI Z2-107	FI	human	SSTI	2012	144	Ia	2078589	35.38	15
MRI Z2-108	FI	human	UTI	2012	144	Ia	2078636	35.38	14
MRI Z2-115	FI	human	SSTI	2011	1	VI	2074185	35.29	8
MRI Z2-120	FI	human	UTI	2012	196	IV	2143474	35.38	27
MRI Z2-121	FI	human	UTI	2012	17	III	2007215	35.37	26
MRI Z2-130	FI	human	carriage	2012	196	IV	2111003	35.36	18
MRI Z2-137	FI	human	UTI	2012	8	Ib	2109944	35.37	17
MRI Z2-138	SE	bovine	milk	2010	10	NT	2165363	35.53	47
MRI Z2-139	SE	bovine	mastitis	2010	722	III	2040088	35.31	11
MRI Z2-140	SE	bovine	milk	2010	196	IV	2217059	35.41	25
MRI Z2-141	SE	bovine	mastitis	2010	1	V	2113594	35.38	18
MRI Z2-142	SE	bovine	mastitis	2010	103	Ia	2077349	35.54	10
MRI Z2-143	SE	bovine	mastitis	2010	724	Ia	2118546	35.44	62
MRI Z2-144	SE	bovine	milk	2010	314	Ia	2012491	35.36	13
MRI Z2-145	SE	bovine	milk	2010	10	II	2152500	35.4	47
MRI Z2-146	SE	bovine	milk	2010	1	V	2104283	35.4	21
MRI Z2-147	SE	bovine	milk	2010	723	III	2157126	35.53	18
MRI Z2-148	SE	bovine	milk	2010	1384	IV	2122724	35.39	25
MRI Z2-149	SE	bovine	milk	2010	196	IV	2120631	35.32	26
MRI Z2-150	SE	bovine	milk	2010	103	Ia	2076896	35.53	8
MRI Z2-152	SE	bovine	mastitis	2011	1387	NT	2148678	35.45	27

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
MRI Z2-153	SE	bovine	milk	2011	103	Ia	1996885	35.35	10
MRI Z2-154	SE	bovine	milk	2011	1	V	2116518	35.37	19
MRI Z2-155	SE	bovine	mastitis	2011	23	III	2161944	35.4	31
MRI Z2-156	SE	bovine	milk	2011	726	IV	2186227	35.43	31
MRI Z2-157	SE	bovine	milk	2011	727	Ia	2059906	35.41	13
MRI Z2-158	SE	bovine	milk	2011	23	Ia	2055495	35.29	20
MRI Z2-159	SE	bovine	milk	2011	103	Ia	2028585	35.44	14
MRI Z2-160	SE	bovine	milk	2011	23	III	2111811	35.51	13
MRI Z2-161	SE	bovine	milk	2011	10	II	2054652	35.2	21
MRI Z2-162	SE	bovine	milk	2011	23	NT	2107425	35.43	17
MRI Z2-163	SE	bovine	mastitis	2011	1	V	2105510	35.34	22
MRI Z2-164	SE	bovine	milk	2011	1	V	2113286	35.38	15
MRI Z2-165	SE	bovine	milk	2011	103	Ia	2076571	35.53	12
MRI Z2-166	SE	bovine	milk	2012	1	V	2114522	35.38	17
MRI Z2-167	SE	bovine	milk	2012	1	V	2117967	35.37	18
MRI Z2-168	SE	bovine	milk	2012	314	Ia	2059741	35.4	10
MRI Z2-169	SE	bovine	milk	2012	1384	IV	2122297	35.39	23
MRI Z2-170	SE	bovine	milk	2012	103	Ia	2068194	35.56	12
MRI Z2-171	SE	bovine	milk	2012	103	Ia	2115411	35.54	11
MRI Z2-172	SE	bovine	milk	2012	196	IV	2143072	35.4	26
MRI Z2-173	SE	bovine	milk	2012	314	Ia	2059466	35.4	11
MRI Z2-174	SE	bovine	milk	2012	728	Ia	1982147	35.3	15
MRI Z2-175	SE	bovine	mastitis	2012	103	Ia	2116048	35.55	13
MRI Z2-177	SE	bovine	milk	2012	1	V	2105863	35.34	17
MRI Z2-178	SE	bovine	milk	2012	726	IV	2185502	35.43	30
MRI Z2-179	SE	bovine	milk	2012	1	V	2083587	35.34	15
MRI Z2-180	SE	bovine	milk	2012	196	IV	2123052	35.38	27
MRI Z2-181	SE	bovine	milk	2012	726	IV	2187312	35.43	33

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
MRI Z2-182	SE	bovine	milk	2012	23	Ia	2065233	35.29	18
MRI Z2-183	SE	bovine	milk	2012	1	V	2118227	35.38	21
MRI Z2-184	SE	human	ID	2011	459	IV	2139137	35.25	16
MRI Z2-187	SE	human	ID	2011	8	Ib	2034303	35.19	18
MRI Z2-195	SE	human	ID	2011	196	IV	2178171	35.52	20
MRI Z2-197	UK	bovine	milk	2014	67	II	2068847	35.42	85
MRI Z2-198	UK	bovine	milk	2014	67	II	2195325	35.56	105
MRI Z2-200	UK	bovine	milk	2014	420	Ia	2163053	35.53	52
MRI Z2-202	UK	bovine	milk	2014	67	II	2126448	35.44	87
MRI Z2-261	SE	bovine	milk	1953	1386	III	2013367	35.38	112
MRI Z2-262	SE	bovine	milk	1953	1386	III	2012884	35.38	112
MRI Z2-263	SE	bovine	mastitis	1953	1386	III	2015941	35.38	113
MRI Z2-264	SE	bovine	mastitis	1953	23	III	2080520	35.33	14
MRI Z2-265	SE	bovine	mastitis	1954	1516	III	1955085	35.34	114
MRI Z2-266	SE	bovine	milk	1954	23	III	2110559	35.53	11
MRI Z2-267	SE	bovine	milk	1954	61	II	2046300	35.25	97
MRI Z2-269	SE	bovine	milk	1954	6	Ib	2072184	35.33	26
MRI Z2-270	SE	bovine	milk	1954	1512	Ib	2028297	35.27	22
MRI Z2-271	SE	bovine	milk	1954	23	Ia	2031105	35.27	12
MRI Z2-272	SE	bovine	milk	1954	23	Ia	2031194	35.27	13
MRI Z2-273	SE	bovine	milk	1955	23	III	2095855	35.46	11
MRI Z2-274	SE	bovine	milk	1961	1513	Ib	2031985	35.29	21
MRI Z2-275	SE	bovine	mastitis	1962	1510	II	2166159	35.54	20
MRI Z2-276	SE	bovine	mastitis	1963	23	III	2061566	35.49	13
MRI Z2-277	SE	bovine	mastitis	1963	6	Ib	2084416	35.43	25
MRI Z2-278	SE	bovine	mastitis	1963	23	III	2059329	35.46	10
MRI Z2-279	SE	bovine	mastitis	1963	23	III	2031448	35.28	8
MRI Z2-280	SE	bovine	milk	1964	23	Ia	2029620	35.27	15

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
MRI Z2-281	SE	bovine	milk	1964	23	III	2083050	35.36	12
MRI Z2-282	SE	bovine	milk	1964	23	III	2083058	35.36	11
MRI Z2-283	SE	bovine	milk	1964	23	III	2147039	35.55	10
MRI Z2-284	SE	bovine	milk	1964	23	III	2063266	35.29	12
MRI Z2-285	SE	bovine	milk	1964	23	III	2120720	35.53	17
MRI Z2-286	SE	bovine	milk	1964	1386	III	2009763	35.37	112
MRI Z2-287	SE	bovine	milk	1964	12	Ib	2170442	35.28	28
MRI Z2-288	SE	bovine	mastitis	1965	6	Ib	2031540	35.28	24
MRI Z2-289	SE	bovine	mastitis	1965	1394	III	2032072	35.36	100
MRI Z2-290	SE	bovine	mastitis	1965	1515	II	2102692	35.42	101
MRI Z2-291	SE	bovine	mastitis	1965	1392	II	2272147	35.63	97
MRI Z2-293	SE	bovine	milk	1966	12	Ib	2169690	35.27	31
MRI Z2-294	SE	bovine	milk	1966	1393	Ib	2068199	35.37	116
MRI Z2-295	SE	bovine	milk	1966	23	III	2181674	35.55	25
MRI Z2-296	SE	bovine	milk	1954	1392	II	2272644	35.63	97
MRI Z2-297	SE	bovine	milk	1967	1511	Ia	2073521	35.46	74
MRI Z2-298	SE	bovine	milk	1967	6	Ib	2114747	35.53	28
MRI Z2-299	SE	bovine	mastitis	1967	1516	II	2255067	35.53	69
MRI Z2-301	SE	bovine	mastitis	1969	23	III	2052215	35.48	12
MRI Z2-302	SE	bovine	mastitis	1969	23	III	2045474	35.44	10
MRI Z2-304	SE	bovine	milk	1970	297	Ia	2125348	35.35	37
MRI Z2-305	SE	bovine	milk	1976	23	III	2100198	35.42	11
MRI Z2-306	SE	bovine	milk	1976	1509	III	2065422	35.46	12
MRI Z2-307	SE	bovine	milk	1977	23	III	2095013	35.52	9
MRI Z2-308	SE	bovine	milk	1978	28	Ia	2133518	35.45	48
MRI Z2-311	SE	bovine	mastitis	1997	10	II	2150057	35.4	41
MRI Z2-312	SE	bovine	milk	1997	23	III	2128823	35.33	11
MRI Z2-313	SE	bovine	milk	1997	23	III	2206005	35.48	21

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
MRI Z2-314	SE	bovine	mastitis	2004	23	III	2210572	35.51	23
MRI Z2-315	SE	bovine	mastitis	2004	23	III	2189758	35.48	20
MRI Z2-316	SE	bovine	milk	2004	1507	Ia	2098210	35.49	17
MRI Z2-317	SE	bovine	milk	2004	103	Ia	2116452	35.55	14
MRI Z2-318	SE	bovine	milk	2004	1	V	2129057	35.39	16
MRI Z2-319	SE	bovine	milk	2004	1517	V	2149116	35.45	14
MRI Z2-320	SE	bovine	milk	2004	23	III	2162261	35.41	34
MRI Z2-321	SE	bovine	milk	2004	196	IV	2093602	35.45	25
MRI Z2-322	SE	bovine	milk	2004	103	Ia	2027939	35.43	14
MRI Z2-323	SE	bovine	mastitis	2004	1384	IV	2167311	35.4	27
MRI Z2-324	SE	bovine	milk	2004	23	III	2112616	35.52	17
MRI Z2-325	SE	bovine	milk	2004	23	III	2143096	35.55	11
MRI Z2-326	SE	bovine	milk	2004	23	III	2148108	35.58	16
MRI Z2-327	SE	bovine	mastitis	2004	10	II	2151301	35.4	44
MRI Z2-328	SE	bovine	milk	2005	1	V	2081819	35.32	17
MRI Z2-329	SE	bovine	milk	2005	1	V	2114889	35.39	19
MRI Z2-330	SE	bovine	milk	2005	732	IV	2221042	35.4	39
MRI Z2-331	SE	bovine	milk	2006	1	V	2114019	35.38	17
MRI Z2-332	SE	bovine	milk	2006	1385	Ia	2119785	35.24	44
MRI Z2-333	SE	bovine	milk	2006	103	Ia	2103084	35.53	16
MRI Z2-334	SE	bovine	milk	2006	1508	III	2197416	35.58	30
MRI Z2-335	SE	bovine	mastitis	2006	314	Ia	2138906	35.47	24
MRI Z2-336	SE	bovine	milk	2006	8	Ib	2254718	35.37	35
MRI Z2-337	SE	bovine	mastitis	2006	23	III	2144151	35.55	12
MRI Z2-338	SE	bovine	mastitis	2006	103	Ia	1997692	35.35	10
MRI Z2-339	SE	bovine	mastitis	2007	1384	IV	2122052	35.38	25
MRI Z2-340	SE	bovine	milk	2007	23	III	2162846	35.41	39
MRI Z2-342	SE	bovine	milk	2009	10	II	2147677	35.41	47

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
MRI Z2-343	SE	bovine	milk	2009	1	V	2134312	35.38	20
MRI Z2-344	AU	bovine	milk	2010	67	II	2178796	35.69	120
MRI Z2-352	AU	bovine	milk	2010	1613	II	2119048	35.54	121
MRI Z2-354	AU	bovine	milk	2010	1613	II	2120823	35.54	123
MRI Z2-363	AU	bovine	milk	2010	67	II	2176422	35.45	107
MRI Z2-364	AU	bovine	milk	2010	1552	II	2144176	35.64	158
MRI-Z2-366	VT	fish, tilapia	brain	2016	283	III	2049179	35.22	17
MRI-Z2-375	VT	fish, tilapia	brain	2016	283	III	2056521	35.23	26
MRI-Z2-379	VT	fish, tilapia	brain	2016	283	III	2049591	35.22	19
MRI-Z2-381	VT	fish, tilapia	brain	2016	283	III	2049580	35.22	19
MRI-Z2-388	VT	fish, tilapia	brain	2016	283	III	2053616	35.22	20
MRI_Z1-012	DK	bovine	BTM	2009	2	IV	2098167	35.31	30
MRI_Z1-022	DK	bovine	BTM	2009	121	NT	2254255	35.43	39
MRI_Z1-023	DK	bovine	BTM	2009	103	Ia	2100525	35.48	12
MRI_Z1-024	DK	bovine	BTM	2009	23	Ia	2105139	35.26	19
MRI_Z1-025	DK	bovine	BTM	2009	1	V	2118108	35.39	15
MRI_Z1-035	DK	bovine	BTM	2009	88	Ia	2154593	35.39	13
MRI_Z1-039	DK	bovine	BTM	2009	4	Ia	2162250	35.41	17
MRI_Z1-048	DK	bovine	BTM	2009	26	V	2083675	35.35	9
MRI_Z1-049	DK	bovine	BTM	2009	19	III	2272185	35.77	29
MRI_Z1-198	UK	sea mammals	trauma	1995	12	Ib	2044700	35.21	22
MRI_Z1-199	UK	sea mammals	storm damage	2002	23	Ia	2018141	35.44	19
MRI_Z1-200	UK	sea mammals	storm damage	2003	23	Ia	2061726	35.36	12
MRI_Z1-201	UK	sea mammals		2003	23	Ia	2003354	35.34	16
MRI_Z1-202	UK	sea mammals	lung emphysema	2003	23	Ia	1956289	35.42	15
MRI_Z1-203	UK	sea mammals	starvation	2006	23	Ia	2020089	35.44	14
MRI_Z1-204	UK, SCT	dog		1998	23	Ia	2001827	35.21	14
MRI_Z1-205	UK, SCT	dog		1999	1	V	2124649	35.28	18

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
MRI_Z1-206	UK, SCT	dog		2001	8	Ib	2047062	35.26	15
MRI_Z1-209	IT	bovine	mastitis	2003	1	V	2124201	35.39	19
MRI_Z1-211	IT	bovine	mastitis	2003	1	V	2121645	35.38	20
MRI_Z1-212	IT	bovine	mastitis	2003	1	V	2126132	35.39	20
MRI_Z1-213	IT	bovine	mastitis	2004	591	II	2274420	35.38	90
MRI_Z1-214	IT	bovine	mastitis	2004	591	II	2301833	35.42	74
MRI_Z1-215	IT	bovine	mastitis	2004	589	IV	2155185	35.32	21
MRI_Z1-216	IT	bovine	mastitis	2004	591	II	2273442	35.38	90
MRI_Z1-217	IT	bovine	mastitis	2004	591	II	2303583	35.43	79
MRI_Z1-218	IT	bovine	mastitis	2005	590	II	2087271	35.2	23
MRI_Z1-219	IT	bovine	mastitis	2008	590	II	2088346	35.2	20
NA0058901	TH	human	blood	2014	283	III	1986646	35.49	222
NGBS008	CA, ON	human		2009	1	V	2063465	35.29	18
NGBS009	CA, ON	human		2010	1	V	2134515	35.26	19
NGBS024	CA, ON	human	blood	2009	459	IV	2250267	35.4	51
NGBS031	CA, ON	human	LOD	2009	17	III	2011767	35.36	25
NGBS034	CA, ON	human	AB	2010	17	III	1967978	35.63	237
NGBS049	CA, ON	human	blood	2010	459	IV	2192386	35.36	44
NGBS065	CA, ON	human	blood	2010	452	IV	2106918	35.38	29
NGBS069	CA, ON	human	LOD	2010	17	III	2117740	35.32	35
NGBS089	CA, ON	human	EOD	2010	17	III	2053952	35.4	31
NGBS1006	CA, SK	human	BOA	2012	459	IV	2166094	35.28	23
NGBS1018	CA, SK	human	ASF	2012	459	NT	2167036	35.43	74
NGBS1021	CA, SK	human	AB	2012	459	IV	2245410	35.36	23
NGBS1041	CA, SK	human	AT	2012	459	IV	2168854	35.27	19
NGBS1042	CA, SK	human	AO	2012	459	IV	2248813	35.35	26
NGBS1045	CA, SK	human	OAT	2013	459	IV	2218762	35.37	18
NGBS1046	CA, SK	human	AT	2013	459	IV	2216136	35.34	19

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
NGBS1047	CA, SK	human	BOA	2013	459	IV	2184119	35.3	20
NGBS1049	CA, SK	human	AB	2013	459	IV	2181474	35.3	18
NGBS1050	CA, SK	human	BOA	2013	452	IV	2069620	35.36	30
NGBS1052	CA, SK	human	OOA	2013	459	IV	2184441	35.25	21
NGBS1056	CA, SK	human	BOA	2014	459	IV	2176201	35.3	18
NGBS1058	CA, SK	human	AT	2014	459	IV	2254958	35.38	26
NGBS1061	CA, MB	human	AT	2012	459	IV	2267683	35.39	26
NGBS1062	CA, MB	human	OAT	2012	459	IV	2187397	35.28	20
NGBS1064	CA, MB	human	AB	2012	459	IV	2181072	35.29	22
NGBS1065	CA, MB	human	AO	2012	459	IV	2182234	35.29	19
NGBS1066	CA, MB	human	BOA	2013	459	IV	2209245	35.33	45
NGBS1067	CA, MB	human	AB	2013	459	IV	2210279	35.33	24
NGBS1068	CA, MB	human	OOA	2013	459	IV	2182340	35.29	16
NGBS1072	CA, MB	human	AT	2013	459	IV	2180941	35.29	19
NGBS1079	CA, MB	human	BOA	2013	452	IV	2067590	35.37	25
NGBS1080	CA, MB	human	BOA	2014	1011	IV	2215378	35.32	18
NGBS1082	CA, MB	human	AO	2014	459	IV	2221347	35.31	34
NGBS1083	CA, MB	human	BOA	2014	459	IV	2203829	35.35	28
NGBS129	CA, ON	human	AB	2010	1063	III	1967777	35.18	29
NGBS151	CA, ON	human	blood	2010	3	IV	2205090	35.51	18
NGBS205	CA, ON	human	LOD	2011	17	III	2122395	35.34	28
NGBS258	CA, ON	human	blood	2011	459	IV	2223498	35.42	22
NGBS272	CA, ON	human		2011	1	V	2117550	35.35	269
NGBS299	CA, ON	human	AB	2011	17	III	2006717	35.36	27
NGBS314	CA, ON	human	blood	2011	452	IV	2014937	35.2	18
NGBS318	CA, ON	human	AB	2011	290	III	2049607	35.39	28
NGBS327	CA, ON	human	AB	2011	484	III	2188583	35.5	30
NGBS345	CA, ON	human	EOD	2011	874	III	2047671	35.16	24

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
NGBS361	CA, ON	human	EOD	2011	17	III	2015268	35.37	30
NGBS379	CA, ON	human	blood	2011	3	IV	2171538	35.44	23
NGBS386	CA, ON	human	adult synovial fluid	2011	17	III	2015886	35.38	28
NGBS400	CA, ON	human	blood	2011	852	IV	2233010	35.32	22
NGBS417	CA, ON	human	child synovial fluid	2011	17	III	2171746	35.32	31
NGBS447	CA, ON	human	blood	2011	196	IV	2038363	35.38	17
NGBS493	CA, ON	human	blood	2012	459	IV	2142292	35.36	18
NGBS499	CA, ON	human		2012	1	V	2039275	35.3	101
NGBS500	CA, ON	human	LOD	2012	17	III	2053309	35.39	28
NGBS501	CA, ON	human	AB	2012	17	III	2088603	35.38	32
NGBS502	CA, ON	human	AB	2012	95	III	2081507	35.47	33
NGBS516	CA, ON	human	EOD	2012	17	III	1993670	35.23	29
NGBS525	CA, ON	human	soft tissue	2011	459	IV	2219357	35.32	22
NGBS531	CA, ON	human	AB	2012	148	III	2152715	35.38	26
NGBS572	CA, ON	human	synovial fluid	2012	452	IV	2020792	35.21	25
NGBS588	CA, ON	human	blood	2012	682	IV	2043856	35.41	22
NGBS598	CA, ON	human	blood	2012	452	IV	2071681	35.39	27
NGBS608	CA, ON	human	AO	2012	17	III	2011423	35.36	30
NGBS622	CA, ON	human	LOD	2012	148	III	2106314	35.31	27
NGBS632	CA, ON	human	adult tissue	2012	17	III	1986430	35.28	28
NGBS680	CA, MB	human	AB	2010	459	IV	2180205	35.3	18
NGBS700	CA, MB	human	AO	2010	459	IV	2181904	35.32	19
NGBS762	CA, MB	human	OOA	2011	459	IV	2181860	35.31	14
NGBS767	CA, MB	human	BOA	2011	459	IV	2183052	35.3	16
NGBS768	CA, MB	human	BOA	2011	452	IV	2015671	35.2	15
NGBS788	CA, MB	human	AO	2011	459	IV	2256175	35.32	23
NGBS789	CA, MB	human	AT	2011	459	IV	2220877	35.33	22
NGBS791	CA, MB	human	AB	2011	459	IV	2254465	35.35	16

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
NGBS798	CA, MB	human	BOA	2011	710	IV	2181657	35.31	25
NGBS801	CA, MB	human	OAT	2011	459	IV	2180785	35.28	18
NGBS824	CA, MB	human	AB	2011	711	IV	2215945	35.31	21
NGBS830	CA, MB	human	OASF	2011	459	IV	2182927	35.31	19
NGBS855	CA, MB	human	BOA	2012	459	IV	2215628	35.34	13
NGBS899	CA, MB	human	BOA	2012	1078	IV	2236775	35.46	18
NGBS933	CA, SK	human	BOA	2010	3	IV	2144137	35.42	29
NGBS956	CA, SK	human	AT	2011	459	IV	2220484	35.31	23
NGBS960	CA, SK	human	AB	2011	3	IV	2142555	35.42	17
NGBS964	CA, SK	human	BOA	2011	459	IV	2185095	35.3	18
NGBS979	CA, SK	human	OAT	2011	459	IV	2199970	35.31	28
NGBS984	CA, SK	human	BOA	2011	459	NT	2177407	35.35	39
NGBS991	CA, SK	human	BOA	2011	196	IV	2229219	35.26	22
NGBS996	CA, SK	human	OASF	2012	459	IV	2219978	35.32	21
NNA019	CN	human	cervical secretion	2014	862	III	2220560	35.35	29
NNA020	CN	human	cervical secretion	2014	651	III	2010630	35.36	25
NNA036	CN	human	suppurative sinusitis	2014	485	Ia	2139737	35.43	36
NNB003	CN	human	vaginitis	2014	930	Ia	2048621	35.54	17
NNB008	CN	human	sputum	2014	485	Ia	2054443	35.38	20
NNP010101	VT	fish, tilapia	brain	2019	283	III	2056577	35.23	19
NNP010201	VT	fish, tilapia	brain	2019	283	III	2051059	35.21	16
NNP020101	VT	fish, tilapia	brain	2019	283	III	2052414	35.23	17
NNP020201	VT	fish, tilapia	brain	2019	283	III	2054643	35.24	27
NNP030101	VT	fish, tilapia	brain	2019	283	III	2048127	35.22	25
NNP040102	VT	fish, tilapia	brain	2019	1395	Ib	1803486	35.27	26
NTC110101	VT	bovine		2019	314	Ia	2079966	35.36	16
NTC110102	VT	bovine		2019	314	Ia	2106815	35.36	23
NTC110103	VT	bovine		2019	314	Ia	2105195	35.35	23

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
NTC110104	VT	bovine		2019	314	Ia	2106327	35.36	29
QMA0271	AU, QLD	fish, giant catfish		2009	261	Ib	1802470	35.28	1
QMA0499	HN	fish, tilapia		2014	260	Ib	1799426	35.24	25
S13	BR	fish, tilapia	eye	2015	552	Ib	1835156	35.43	1
SA0034366	TH	human	blood	2012	283	III	1993890	35.52	213
SA01	BR	fish, tilapia		2003	552	Ib	1841943	35.48	1
SA05	BR	fish, tilapia	brain	2003	552	Ib	1841945	35.48	1
SA09	BR	fish, tilapia		2005	552	Ib	1841929	35.48	1
SA102	BR	fish, tilapia	brain	2010	927	Ib	1849522	35.49	1
SA132	BR	fish, tilapia	brain	2011	260	Ib	1852032	35.49	1
SA136	BR	fish, tilapia	brain	2011	260	Ib	1849103	35.48	1
SA159	BR	fish, tilapia	brain	2011	552*	Ib	1841483	35.48	1
SA16	BR	fish, tilapia		2006	552	Ib	1841859	35.48	1
SA184	BR	fish, tilapia	brain	2011	552	Ib	1841893	35.48	1
SA191	BR	fish, tilapia	brain	2011	260	Ib	1848676	35.49	1
SA195	BR	fish, tilapia	brain	2011	552	Ib	1841715	35.48	1
SA20	BR	fish, tilapia	ME	2006	552	Ib	1841952	35.48	1
SA201	BR	fish, tilapia	brain	2012	552	Ib	1841835	35.48	1
SA209	BR	fish, tilapia	brain	2012	552	Ib	1841857	35.48	1
SA212	BR	fish, tilapia	brain	2012	552	Ib	1841962	35.48	1
SA245	BR	fish, tilapia	brain	2013	260	Ib	1848956	35.49	1
SA256	BR	fish, tilapia	brain	2013	260	Ib	1848972	35.48	1
SA33	BR	fish, tilapia	brain	2006	552	Ib	1841628	35.48	1
SA330	BR	fish, tilapia	brain	2013	552	Ib	1842081	35.47	1
SA333	BR	fish, tilapia	brain	2014	552	Ib	1842037	35.47	1
SA341	BR	fish, tilapia	brain	2014	552	Ib	1842113	35.47	1
SA374	BR	fish, tilapia	brain	2014	552	Ib	1842255	35.47	1
SA53	BR	fish, tilapia	brain	2007	260	Ib	1848971	35.48	1

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
SA623	BR	fish, tilapia	brain	2015	552	Ib	1842115	35.48	1
SA73	BR	fish, tilapia	brain	2008	260	Ib	1848839	35.49	1
SA75	BR	fish, tilapia		2008	260	Ib	1849016	35.48	1
SA79	BR	fish, tilapia	kidney	2009	552	Ib	1841946	35.48	1
SA81	BR	fish, tilapia		2009	552	Ib	1840363	35.48	1
SA85	BR	fish, tilapia	brain	2010	927	Ib	1849989	35.49	1
SA97	BR	fish, tilapia	brain	2010	927	Ib	1856410	35.49	1
Sag158	CN	human		2014	19	III	2096882	35.67	1
Sag37	CN	human		2014	12	Ib	2198785	35.77	1
SAMEA2168833	ES	human		2005	1	V	2081744	35.3	46
SAMEA2168834	ES	bovine	milk	2005	2	II	2132817	35.49	72
SAMEA2168835	CAF	human		2006	182	III	2189185	35.35	70
SAMEA2168836	FR	goat		1969	19	III	2125850	35.39	57
SAMEA2168837	FR	human	carriage	2011	19	III	2187391	35.54	79
SAMEA2168838	USA	human		1956	19	II	2049315	35.25	42
SAMEA2168839	FR	human	carriage	1958	19	III	2140798	35.24	82
SAMEA2168840	CAF	human		2006	28	II	2124158	35.48	40
SAMEA2168841	FR	human	ID	1959	28	II	2062909	35.26	35
SAMEA2168842	FR	human		1960	19	Ia	2081639	35.31	39
SAMEA2168843	DK	human	urine	1960	19	III	2168228	35.36	49
SAMEA2168844	MD	human		2006	28	II	2058563	35.32	42
SAMEA2168845	ES	bovine	milk	2005	19	V	2124558	35.41	63
SAMEA2168846	ES	human	carriage	2005	19	III	2082915	35.27	54
SAMEA2168847	CAF	human		2006	1	V	2118323	35.29	52
SAMEA2168848	FR	human	carriage	2008	1	V	2074181	35.24	34
SAMEA2168849	FR	human	urine	1998	1	V	2086232	35.34	43
SAMEA2168850	FR	human	urine	1998	1	VI	2069903	35.25	40
SAMEA2168851	DE	dog	carriage	1997	1	V	2029215	35.28	39

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
SAMEA2168852	FR	human		2007	196	IV	2189117	35.51	74
SAMEA2168853	FR	human	AID	2009	196	IV	2172278	35.36	39
SAMEA2168854	CAF	human	carriage	2007	196	IV	2036879	35.16	30
SAMEA2168855	FR	human	carriage	2008	459	IV	2225193	35.24	47
SAMEA2168856	FR	human	AID	2010	459	IV	2182022	35.25	38
SAMEA2168857	FR	human	LOD	2006	17	III	1984262	35.24	54
SAMEA2168860	USA	human	NI	<1989	17	III	2083055	35.2	95
SAMEA2168861	FR	human	LOD	2007	17	III	2012041	35.26	59
SAMEA2168864	FR	human	LOD	2008	17	III	1983219	35.32	57
SAMEA2168868	FR	human	LOD	2009	17	III	2011970	35.26	57
SAMEA2168870	FR	human	AID	2009	291	IV	1968211	35.23	40
SAMEA2168873	FR	human	carriage	2010	17	III	2099125	35.37	61
SAMEA2168880	FR	human	LOD	2010	17	III	2034272	35.29	58
SAMEA2168888	FR	human	EOD	2011	291	IV	2080638	35.3	41
SAMEA2168890	FR	human	ID	1961	17	III	2057424	35.21	61
SAMEA2168891	FR	human		1955	17	III	2031135	35.13	69
SAMEA2168893	CZ	bovine			355	III	2038932	35.33	194
SAMEA2168894	SENG	human	carriage	2007	17	III	2064532	35.3	66
SAMEA2168895	MD	human		2006	291	IV	2005201	35.28	42
SAMEA2168896	MD	human		2006	17	III	1988504	35.21	72
SAMEA2168897	MD	human	carriage	2007	17	III	1979847	35.24	60
SAMEA2168900	FR	human	LOD	1995	17	III	1947585	35.15	165
SAMEA2168902	ES	human		2005	291	IV	1974023	35.22	39
SAMEA2168903	UK	human	EOD	2007	17	III	2009414	35.25	191
SAMEA2168904	DE	dog		>2002	23	Ia	2019834	35.21	37
SAMEA2168905	FR	bovine	mastitis		23	III	2126297	35.42	47
SAMEA2168906	CAF	human		2006	23	Ia	2022918	35.18	50
SAMEA2168908	FR	human	urine	1998	23	Ia	2100989	35.14	33

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
SAMEA2168909	FR	human	AID	2005	380	III	2062978	35.23	35
SAMEA2168910	FR	human	carriage	1959	23	Ia	1965275	35.14	41
SAMEA2168911	FR	human	ID	1959	23	Ia	2116646	35.31	40
SAMEA2168912	FR	human	ID	1953	23	Ia	1951311	35.12	37
SAMEA2168913	MD	human	ID	1953	23	Ia	2011451	35.24	34
SAMEA2168914	MD	human		2006	23	V	1992189	35.18	36
SAMEA2168916	FR	human		2006	23	V	1993072	35.19	79
SAMEA2168917	CAF	human	EOD	1996	23	Ia	2066333	35.14	34
SAMEA2168918	CAF	human	carriage	2007	17	III	1984280	35.22	56
SAMEA2168920	CAF	human	carriage	2007	1	V	2053206	35.27	48
SAMEA2168922	FR	human	ID	2006	1	V	2132936	35.36	45
SAMEA2168923	FR	human	AID	2008	19	III	2182186	35.37	69
SAMEA2168924	FR	human	ID	2008	1	V	2136696	35.33	32
SAMEA2168925	FR	human	AID	2008	1	V	2107587	35.28	34
SAMEA2168927	CAF	human	carriage	2007	19	III	2123037	35.24	67
SAMEA2168929	NC	human	carriage	2003	19	III	2108746	35.23	47
SAMEA2168930	MY	bovine	milk	1971	17	III	1962935	35.22	56
SAMEA2168931	CAF	human	carriage	2006	8	Ib	2049196	35.25	33
SAMEA2168932	CAF	human	carriage	2006	26	V	2077276	35.22	94
SAMEA2168935	SENG	human	carriage	2006	26	V	2091241	35.38	89
SAMEA2168936	DE	bovine	mastitis	1971	7	Ia	2146158	35.33	110
SAMEA2168937	FR	human	arthritis	2011	291	IV	2080085	35.4	52
SAMEA2168938	FR	human	sperm	1998	10	II	2007593	35.16	52
SAMEA2168939	FR	human	urine	2005	248	Ia	2089265	35.2	50
SAMEA2168941	MD	human	carriage	2007	314	Ia	2172688	35.31	40
SAMEA2168942	MD	human	carriage	2007	103	Ia	2049680	35.19	73
SAMEA2168943	MD	human	carriage	2007	8	Ib	2176557	35.16	193
SAMEA2168944	ES	bovine	milk	2005	1574	II	2225365	35.66	134

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
SAMEA2168945	FR	bovine	mastitis	1983	10	Ib	2043049	35.17	55
SAMEA3888064	FR	bovine	mastitis	1996	61	II	2062134	35.41	94
SAMEA3888065	FR	bovine	mastitis	1997	416	III	2163965	35.45	84
SAMEA3888067	PT	bovine	mastitis	2011	61	II	2136295	35.36	104
SAMEA3888075	PT	bovine	mastitis	2011	61	II	2054135	35.44	100
SAMEA3888076	PT	bovine	mastitis	2011	61	II	2173738	35.42	103
SAMEA3888077	PT	bovine	mastitis	2011	61	II	2077615	35.34	95
SAMEA3888086	PT	bovine	mastitis	2012	61	II	2226149	35.51	127
SAMEA3888116	PT	bovine	mastitis	2012	61	II	1993996	35.36	76
SAMEA3888121	PT	bovine	mastitis	2012	61	II	2156239	35.45	110
SAMEA3888128	PT	bovine	mastitis	2012	1620	II	2164869	35.43	100
SAMEA3888149	PT	bovine	mastitis	2013	61	II	2188751	35.58	105
SAMEA3888166	PT	bovine	mastitis	2013	61*	II	2145213	35.49	152
SAMEA3888184	PT	bovine	mastitis	2013	61*	II	2221672	35.5	158
SAMEA3888204	PT	bovine	mastitis	2014	61	II	2225448	35.5	140
SAMEA3888205	PT	bovine	mastitis	2014	61	II	2142277	35.48	108
SAMEA3888207	PT	bovine	mastitis	2014	61	II	2140602	35.48	106
SAMEA3888210	PT	bovine	mastitis	2002	554	II	2110471	35.49	73
SAMEA3888211	PT	bovine	mastitis	2002	61	II	2156813	35.45	81
SAMEA3888213	PT	bovine	mastitis	2002	61	II	2029328	35.34	110
SAMEA3888215	PT	bovine	mastitis	2003	61	II	2183378	35.39	124
SAMEA3888220	PT	bovine	mastitis	2012	61	II	2141917	35.38	65
SAMEA3888222	PT	bovine	mastitis	2012	554	II	2175341	35.59	80
SAMEA3888224	PT	bovine	mastitis	2013	554	II	2094555	35.48	66
SAMEA3888238	PT	bovine	mastitis	2011	554	II	2096502	35.48	70
SAMEA3888242	PT	bovine	mastitis	2012	554	II	2097397	35.48	73
SAMEA3888243	PT	bovine	mastitis	2013	61	II	2141699	35.37	67
SAMEA3888245	PT	bovine	mastitis	2012	61	II	2141473	35.38	66

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
SAMEA3888272	PT	bovine	mastitis	2011	61	II	2202426	35.56	66
SAMEA3888273	PT	bovine	mastitis	2002	2	II	2149011	35.61	30
SAMEA3888274	PT	bovine	mastitis	2012	2	II	2190470	35.56	35
SAMEA3888275	ES	bovine	mastitis	2005	2	II	2128244	35.56	37
SG-M1003	VT	human	blood	2015	283	III	2048255	35.21	15
SG-M1011	VT	human	blood	2016	283	III	2051929	35.23	19
SG-M121	SG	human	blood	2015	283	III	2069748	35.24	22
SG-M249	SG	human	joint aspirate	2015	283	III	2069172	35.24	21
SG-M257	SG	human	synovial tissue	2015	283	III	2069219	35.24	21
SG-M29	SG	human	blood	2012	283	III	2068112	35.24	25
SG-M32	SG	human	CSF	2012	283	III	2070018	35.24	24
SG-M361	SG	human	blood	2013	283	III	2070080	35.25	28
SG-M406	SG	human	blood	2014	283	III	2070028	35.24	25
SG-M408	SG	human	blood	2014	167	III	2184286	35.48	18
SG-M423	SG	human	blood	2015	3	III	2212760	35.51	22
SG-M426	SG	human	blood	2015	335	III	2139821	35.36	33
SG-M474	TH	human	blood	2015	283	III	2039310	35.21	19
SG-M487	SG	food	muscles	2015	283	III	2069562	35.24	24
SG-M587	SG	human	blood	2003	283	III	2008142	35.18	16
SG-M613	SG	human	blood	2004	283	III	2018897	35.2	19
SG-M654	SG	human	blood	2006	283	III	2018815	35.19	21
SG-M821	SG	human	blood	2007	283	III	1999944	35.18	16
SG-M844	SG	human	blood	2008	283	III	2058969	35.22	22
SG-M870	SG	human	blood	2009	283	III	1999762	35.19	16
SG-M883	SG	human	blood	2010	283	III	2037442	35.2	23
SG-M900	SG	human	blood	2001	283	III	2019130	35.2	23
SG-M917	SG	human	blood	2002	283	III	2007259	35.17	17
SG-M961	LA	human	blood	2006	283	III	2056643	35.23	17

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
SG-M964	LA	human	blood	2016	283	III	2028170	35.19	19
SG-M970	LA	human	blood	2015	283	III	2030236	35.2	20
SG-M973	LA	human	blood	2010	283	III	2057433	35.23	21
SG-M976	LA	human	blood	2012	283	III	2056607	35.23	18
SG-M978	LA	human	blood	2013	283	III	2038853	35.2	22
SG-M980	LA	human	blood	2000	283	III	2056376	35.23	20
SG-M984	LA	human	blood	2007	283	III	2087653	35.27	26
SG-M990	LA	human	blood	2009	283	III	2057810	35.24	19
SG-M991	LA	human	blood	2003	283	III	2057856	35.24	20
SG-M992	LA	human	blood	2014	283	III	2039020	35.2	19
SGBS001	USA, TX	human		1993	1	V	2056342	35.29	17
SGBS003	USA, TX	human		1994	1	V	2072667	35.26	20
SGBS004	USA, TX	human		1995	1	V	2060112	35.3	18
SGBS005	USA, TX	human		1996	1	V	2057072	35.29	18
SGBS006	USA, TX	human		1998	1	V	2059366	35.3	17
SGBS007	USA, TX	human		1997	1	V	2057581	35.3	15
SGBS014	USA, TX	human		2002	1	V	2093278	35.34	17
SGBS015	USA, TX	human		2003	1	V	2135098	35.37	15
SGBS020	USA, TX	human		2006	1	V	2161341	35.42	17
SGBS021	USA, TX	human		2007	1	V	2094433	35.38	15
SGBS022	USA, TX	human		2008	1	V	2138582	35.31	18
SGBS025	USA, TX	human		2010	1	V	2063405	35.3	17
SGBS026	USA, TX	human		2011	1	V	2072389	35.26	19
SGBS031	USA, TX	human		1992	1	V	2066901	35.34	19
SGBS044	USA, TX	human		1999	1	V	2089738	35.29	20
SGBS047	USA, TX	human		2000	1	V	2072154	35.32	23
SGBS049	USA, TX	human		2000	873	V	2067902	35.33	23
SGBS054	USA, TX	human		2001	1	V	2070608	35.34	23

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
SGBS074	USA, TX	human		2004	1	V	2072764	35.33	21
SGBS079	USA, TX	human		2004	153	V	2072703	35.33	21
SGBS084	USA, TX	human		2005	1	V	2072039	35.33	21
SGBS108	USA, TX	human		2006	872	V	2105005	35.38	21
SGBS122	USA, TX	human		2009	1	V	2079205	35.29	23
SGBS150	USA, TX	human		2012	1	V	2132378	35.26	25
SGEHI2015-101	SG	food, black tilapia		2015	283	III	1982583	35.16	13
SGEHI2015-243_1	SG	food, mackerel tuna		2015	23	Ia	2059837	35.37	15
SGEHI2015-29	SG	food, red tilapia		2015	283	III	2020216	35.19	14
SGEHI2015-31	SG	food, snakehead		2015	283	III	2012018	35.17	19
SGEHI2015-49	SG	food, snakehead		2015	23	Ia	1978303	35.2	15
SGEHI2015-60	SG	food, black tilapia		2015	24	Ia	2110894	35.13	16
SGEHI2015-63	SG	food, snakehead		2015	7	Ia	2030468	35.34	39
SGEHI2015-77	SG	food, snakehead		2015	1611	II	2112236	35.29	25
SGEHI2015-II33	SG	food, snakehead		2015	335	III	2141395	35.35	38
SGEHI2015-II47	SG	food, grass carp		2015	652	II	2021234	35.2	32
SGEHI2015-II55_1	SG	food, wolf herring		2015	1	V	2047676	35.33	18
SGEHI2015-II56	SG	food, bighead carp		2015	1	V	2052213	35.33	15
SGEHI2015-IV100_1	SG	food, salmon		2015	485	Ia	2058564	35.23	10
SGEHI2015-IV100_2	SG	food, salmon		2015	1	V	2056537	35.29	14
SGEHI2015-IV118_1	SG	food, salmon		2015	861	III	2225513	35.65	34
SGEHI2015-IV132_1	SG	food, salmon		2015	24	V	2092506	35.38	6
SGEHI2015-IV170_1	SG	food, tilapia		2015	7	Ia	2037838	35.35	41
SGEHI2015-IV211_1	SG	food, tuna		2015	103	Ia	2072642	35.34	14
SGEHI2015-IV227	SG	food, wolf herring		2015	1	VI	2143319	35.31	7
SGEHI2015-IV232_1	SG	food, salmon		2015	103	Ia	2150509	35.2	9
SGEHI2015-IV45_2	SG	food, tilapia		2015	651	III	2044673	35.23	12
SGEHI2015-IV72_1	SG	food, salmon		2015	1	VII	2136950	35.51	14

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
SGEHI2015-IV87_1	SG	food, salmon		2015	7	Ia	2036080	35.34	38
SGEHI2015-NWC941_2	SG	food, grass carp		2015	283	III	2071016	35.24	28
SP550_GBS	AU	human		2008	1	V	2054773	35.29	13
SP551_GBS	AU	human		2008	17	III	1989065	35.1	26
SP552_GBS	AU	human		2008	28	II	2111797	35.31	17
SP554_GBS	AU	human		2008	335	III	2137664	35.35	36
SP555_GBS	AU	human		2008	23	III	1939382	35.19	11
SP556_GBS	AU	human		2008	23	Ia	2031220	35.35	17
SP580_GBS	AU	human		2008	19	III	2179678	35.53	34
SP581_GBS	AU	human		2008	2	IV	2037857	35.28	20
SP585_GBS	AU	human		2008	22	II	2039091	35.3	67
SP592_GBS	AU	human		2008	652	II	2015367	35.19	24
SP594_GBS	AU	human		2008	106	III	2169626	35.39	43
SP595_GBS	AU	human		2008	24	Ia	2033086	35.27	13
SRR5061187*	USA	human	ID	2015	103	Ia	2081201	35.44	14
SRR5061221*	USA	human	ID	2015	314	Ia	2106993	35.29	18
SRR5061604*	USA	human		2015	314	Ia	2058933	35.35	11
SRR7282039*	USA	human		2016	314	Ia	2166617	35.42	38
SRR7282212*	USA	human	ID	2016	103	Ia	2055901	35.33	13
SRR7283388*	USA	human	ID	2016	103	VII	2014891	35.21	14
SS1	USA	human	blood	1992	1	V	2092071	35.52	1
STIR-CD-01	KU	fish, mullet	brain	2001	7	Ia	2042744	35.35	32
STIR-CD-07	HN	fish, tilapia	heart	2008	260	Ib	1807476	35.28	23
STIR-CD-09	COL	fish, tilapia	kidney/brain	2008	260	Ib	1804941	35.27	25
STIR-CD-13	CR	fish, tilapia	eye	2008	260	Ib	1800362	35.25	30
STIR-CD-14	VT	fish, tilapia		2006	491	III	2058877	35.23	18
STIR-CD-17	HN	fish, tilapia	heart	2008	1623	Ib	1796870	35.26	71
STIR-CD-21	TH	frog	liver	2009	7	Ia	2107009	35.45	40

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
STIR-CD-22	TH	fish, tilapia	kidney		7	Ia	2106457	35.45	42
STIR-CD-25	TH	fish, tilapia	kidney		283	III	2060902	35.25	23
STIR-CD-26	TH	fish, tilapia	kidney		500	Ia	2031576	35.34	33
STIR-CD-29	BE	fish, tilapia	kidney/brain	2007	261	Ib	1801348	35.27	20
str._Gottschalk_1002A	CA, QC	human			23	Ia	2075685	35.3	21
str._Gottschalk_1003A	CA, QC	human			19	III	2128071	35.29	34
str._Gottschalk_1005B	CA, QC	human			288	IV	2128423	35.36	35
str._Gottschalk_13227	CA, QC	human			1	V	2129399	35.29	20
str._Gottschalk_998A	CA, QC	human			12	Ib	2106479	35.31	20
str._Gottschalk_999B	CA, QC	human			8	Ib	2132563	35.35	18
TCT1031K	MY	fish, tilapia		2008	283	III	1988265	35.18	12
TCT358K	MY	fish, tilapia		2007	283	III	1989011	35.18	15
WC1535	CN	fish, tilapia		2015	7	Ia	2212568	35.81	1
WSB3229	VT	fish		2018	283	III	2048195	35.21	16
WSB3235	VT	fish		2018	1395	Ib	1794211	35.24	24
WSB3236	VT	fish		2018	7	Ia	2086907	35.42	50
WSB3339	TH	fish, tilapia		2016	7	Ia	2035237	35.35	31
WSB3340	TH	fish, tilapia		2016	283	III	2041188	35.22	16
WSB3341	TH	fish, tilapia		2016	283	III	2088046	35.24	28
WSB3342	TH	fish, tilapia		2016	283	III	2088709	35.24	29
WSB3343	TH	frog		2018	283	III	2041541	35.22	18
WSB3344	TH	fish, tilapia		2018	283	III	2039611	35.21	14
WSB3345	TH	fish, tilapia		2013	7	Ia	2106770	35.46	46
WSB3346	TH	fish, tilapia		2012	7	Ia	2105104	35.46	42
WSB3347	TH	fish, tilapia		2013	7	Ia	2106666	35.46	42
WSB3348	TH	fish, tilapia		2013	283	III	2035140	35.23	20
WSB3349	TH	fish, tilapia		2013	283	III	2034604	35.23	16
WSB3350	TH	fish, tilapia		2013	283	III	2040458	35.21	18

Table C.2 continued from previous page

ISOLATE	COUNTRY	HOST	CLINICAL MAN.	YEAR	ST	SEROTYPE	SIZE (bp)	GC (%)	CONTIGS
WSB3351	TH	fish, tilapia		2013	283	III	2034773	35.23	17
WSB3352	TH	fish, tilapia		2011	7	Ia	2113137	35.46	43
WSB3353	TH	fish, tilapia		2012	283	III	2040731	35.21	15
WSB3354	TH	fish, tilapia		2012	7	Ia	2110447	35.47	42
WSB3355	TH	fish, tilapia		2012	7	Ia	2110509	35.47	46
WSB3356	TH	frog		2007	7	Ia	2217956	35.36	222
WSB3357	TH	fish, tilapia		2007	7	Ia	2110067	35.45	41
WSB3358	TH	fish, tilapia		2008	7	Ia	2112231	35.47	40
WSB3359	TH	fish, tilapia		2008	283	III	2041214	35.21	16
WSB3360	TH	fish, tilapia		2008	7	Ia	2113131	35.47	45
WSB3361	TH	fish, tilapia		2006	500	Ia	2036191	35.35	32
WSB3362	TH	fish, tilapia		2006	7	Ia	2108372	35.47	45
WSB3363	TH	fish, tilapia		2006	283	III	2032569	35.2	15
WSB3364	TH	fish, tilapia		2007	7	Ia	2111168	35.46	38
WSB3365	TH	fish, tilapia		2007	7	Ia	2113180	35.47	43
WSB3366	TH	fish, tilapia		2007	7	Ia	2111880	35.46	44
WSB3367	TH	fish, tilapia		2007	7	Ia	2110793	35.46	38
WSB3368	TH	fish, tilapia		2007	283	III	2041184	35.22	20
WSB3369	TH	fish, mekong giant catfish		2019	283	III	2046300	35.21	14
WSB3370	TH	fish, giant sea perch		2019	283	III	2029240	35.2	15
WSB3371	TH	fish, giant sea perch		2019	7	Ia	2144748	35.48	42
WSB3372	TH	fish, giant sea perch		2019	7	Ia	2108018	35.45	41
WSB3373	TH	fish, tilapia		2019	283	III	2044505	35.2	13

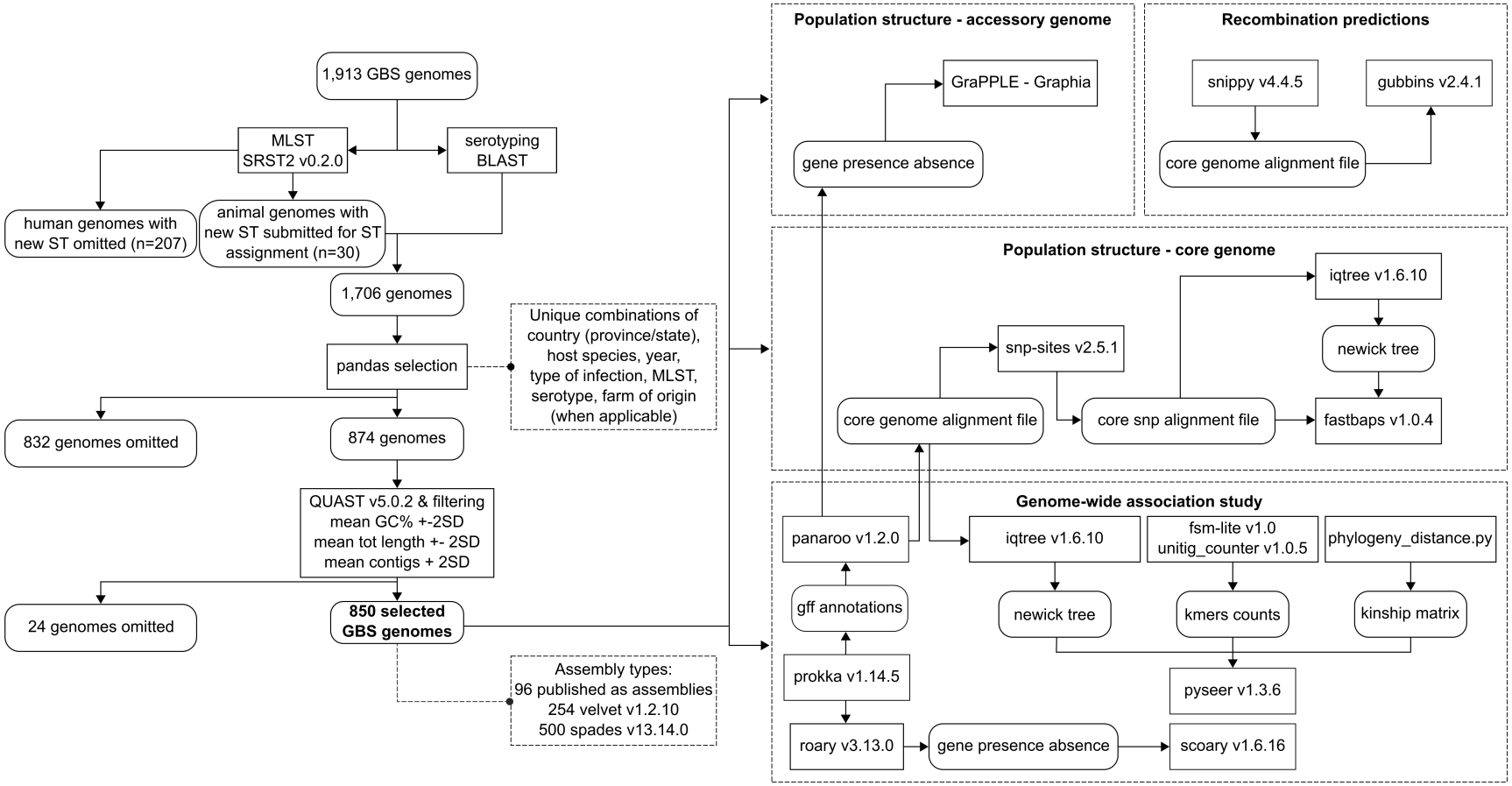


Figure C.1: Diagram illustrating the steps undertaken for the curation of a high-quality dataset representative of the global group B *Streptococcus* population and its subsequent analyses (analysis of population structure based on core and accessory genome, recombination predictions, which are described in this chapter, and GWAS, which is described in chapter 5). Programs used and their versions are illustrated within rectangles, whereas files are indicated within rounded rectangles.

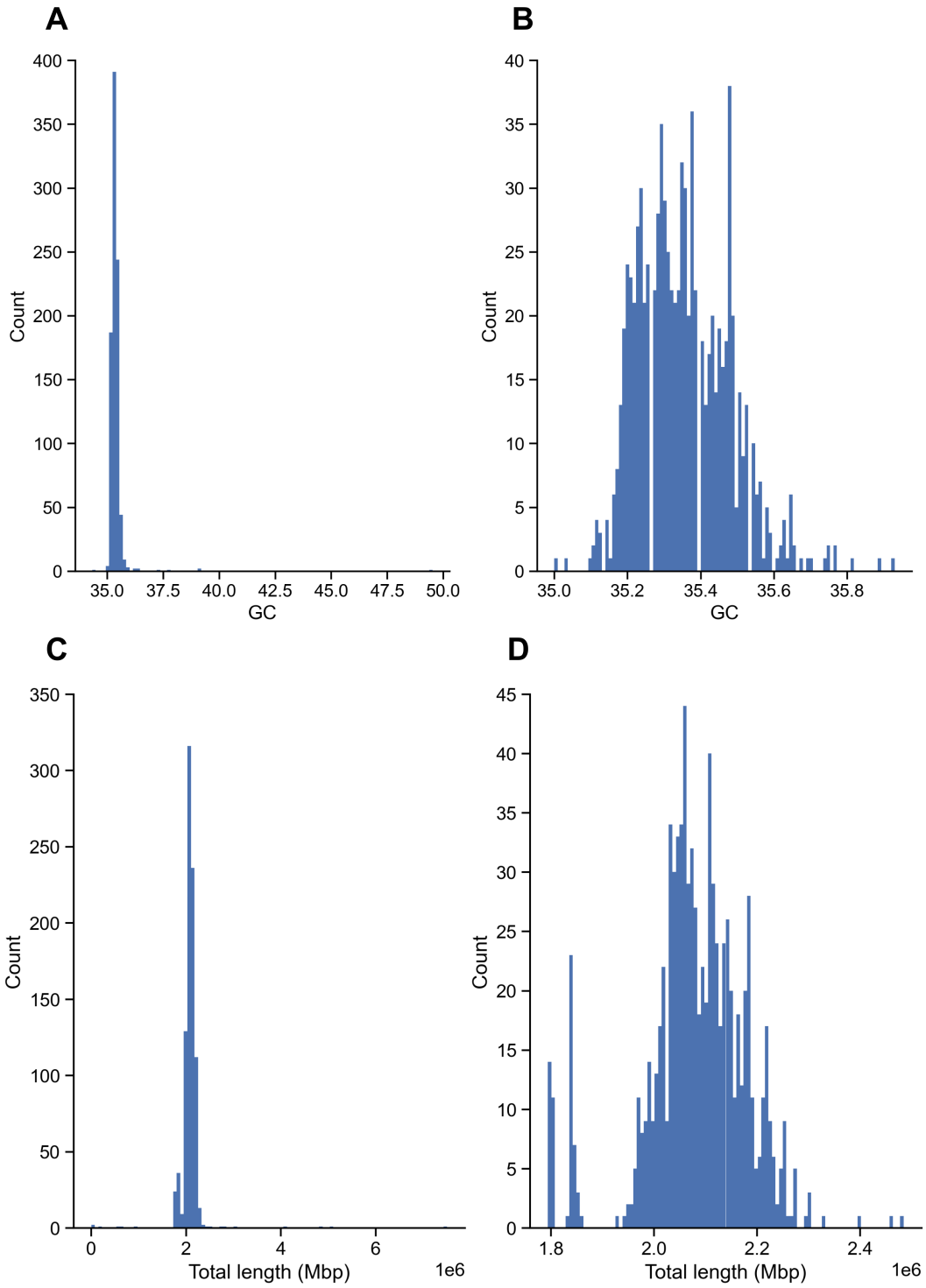


Figure C.2: Frequency plots for group B *Streptococcus* genomes included in this study. A) and C) show GC content (%) and total genome length, respectively, for 874 genomes, before the exclusion of assemblies that fell outside the reference ranges for at least one of the three parameters (GC content, total genome length and total number of contigs, calculated as mean ± 2SD). B) and D) show the same parameters for the 850 genome assemblies that passed the quality control filter. Plots were generated with matplotlib v3.3.2 (Barrett et al., 2005).

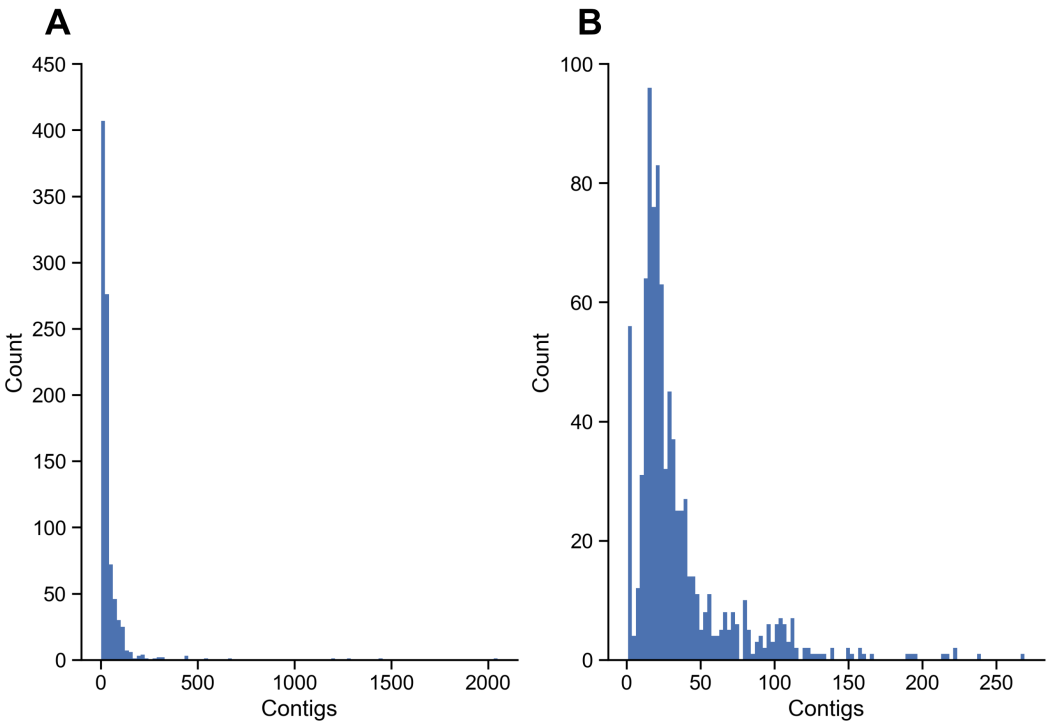


Figure C.3: Frequency plots for group B *Streptococcus* genomes included in this study. A) shows total number of contigs for 874 genomes, before the exclusion of assemblies that fell outside the reference ranges for at least one of the three parameters (GC content, total genome length and total number of contigs, calculated as mean $\pm$ 2SD). B) shows the same parameter for the 850 genome assemblies that passed the quality control filter. Plots were generated with matplotlib v3.3.2 (Barrett et al., 2005).

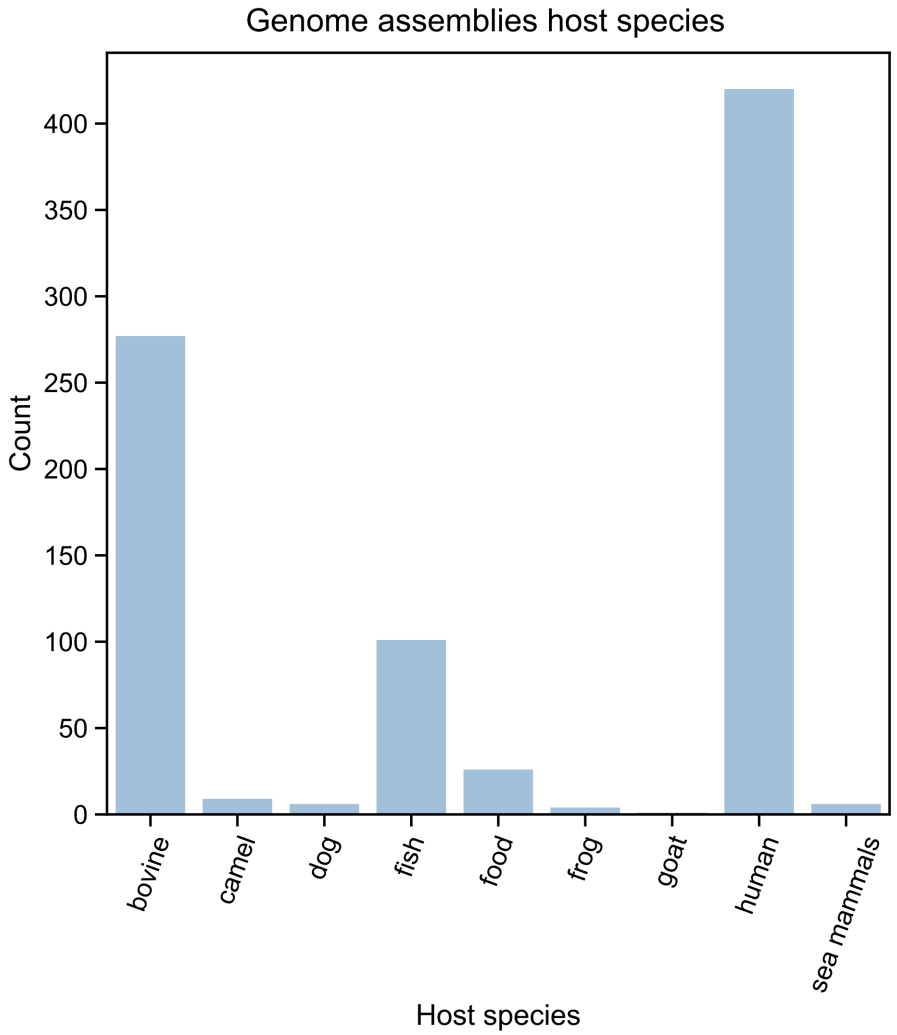


Figure C.4: Frequency plot for group B *Streptococcus* genomes included in this study based on host groups. Nine major host groups/sample types were identified: human (*n*=420), bovine (*n*=277), fish (*n*=101), food market fish samples (*n*=26), camel (*n*=9), dog (*n*=6), sea mammals (*n*=6), frog (*n*=4) and goat (*n*=1).

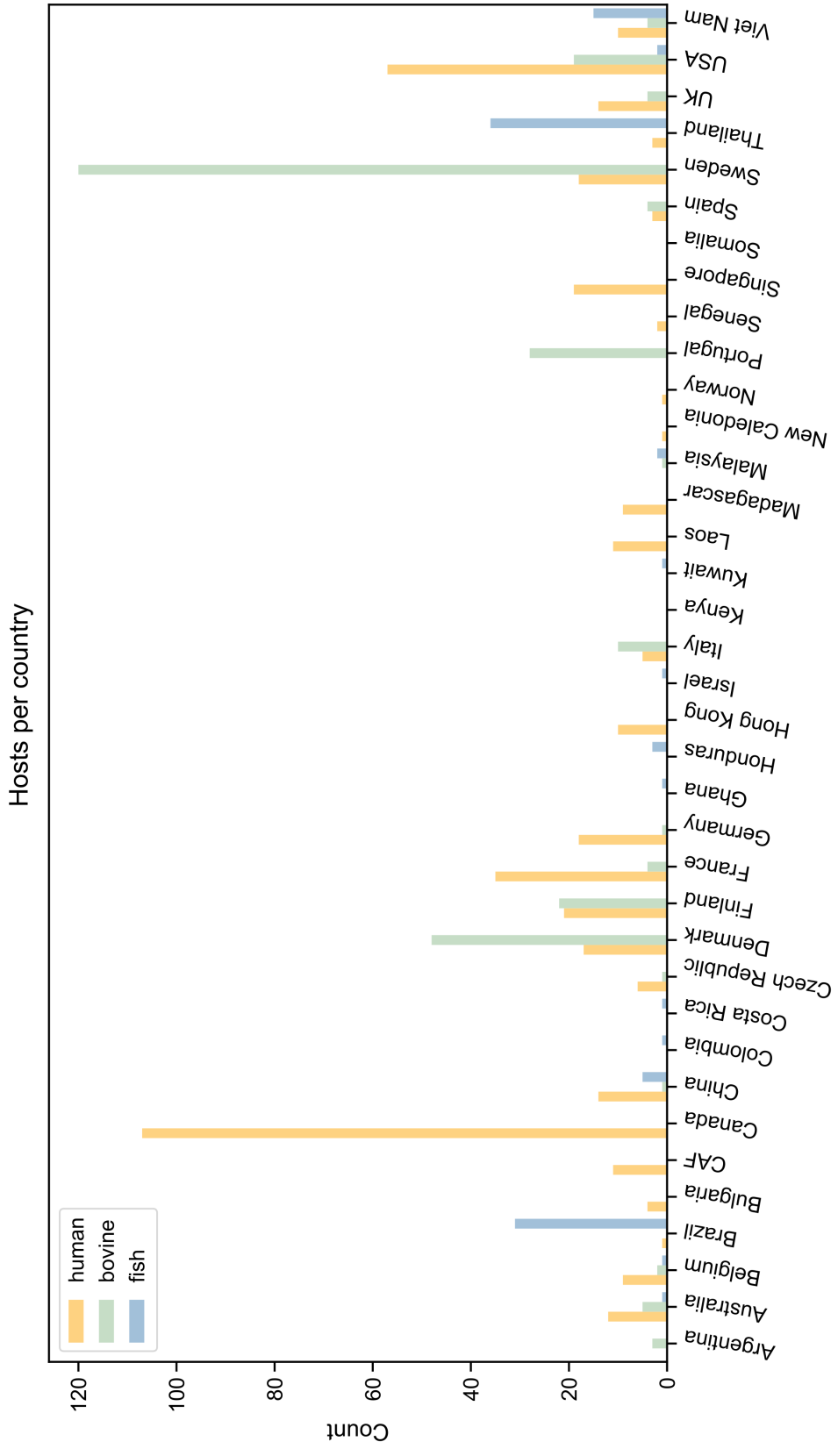


Figure C.5: Frequency plot for group B *Streptococcus* genomes included in this study based on countries of origin. The three major host species are indicated.

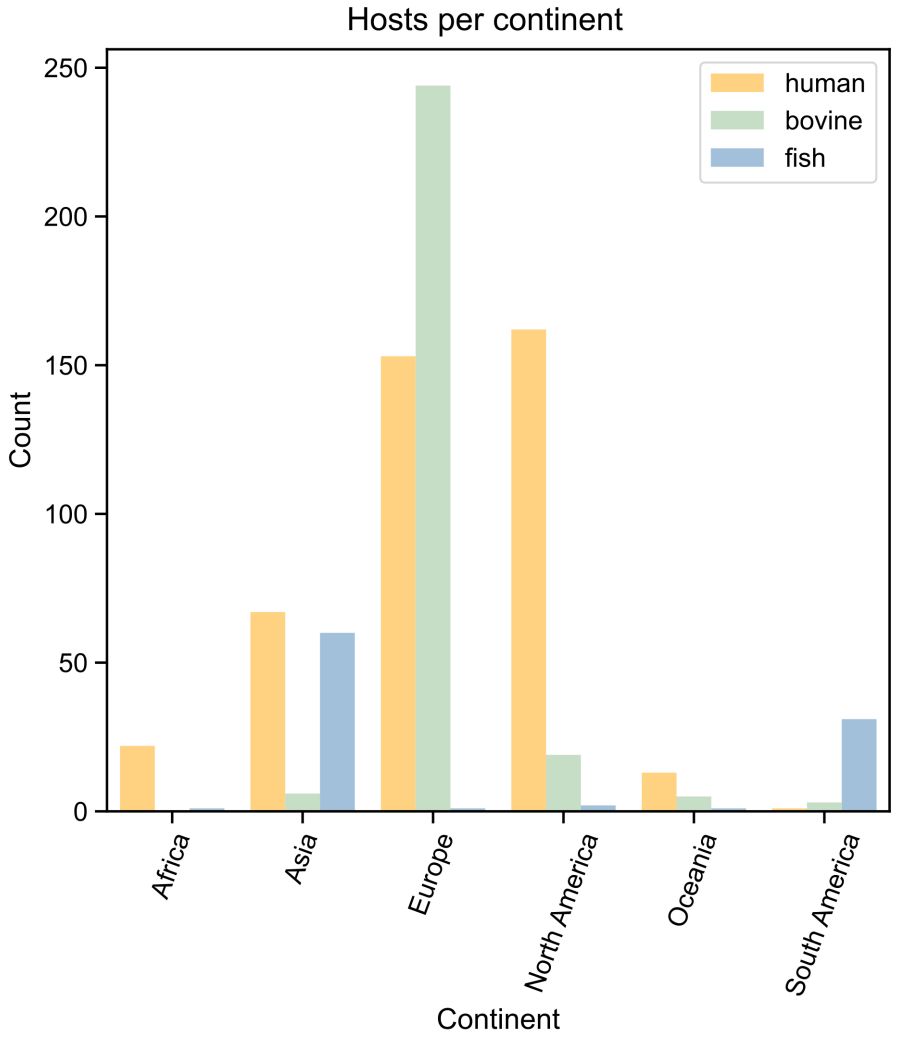


Figure C.6: Frequency plot for group B *Streptococcus* genomes included in this study based on continent of origin. North America comprises isolates from both North and Central America. The three major host species are indicated.

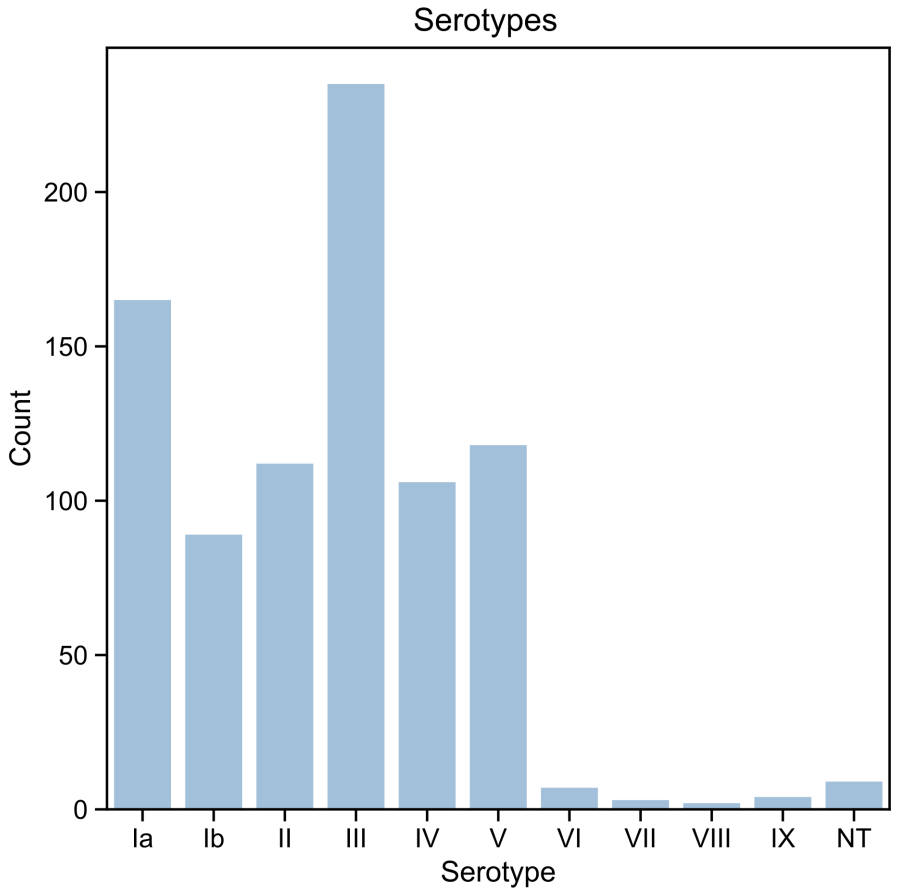


Figure C.7: Frequency plot for group B *Streptococcus* genomes included in this study based on serotype. The most well-represented serotype was serotype III (*n*=235), followed by serotype Ia (*n*=165), V (*n*=118), II (*n*=112), IV (*n*=106), Ib (*n*=89), VI (*n*=7), IX (*n*=4), VII (*n*=3), VIII (*n*=2) and nine non typeable isolates (NT).

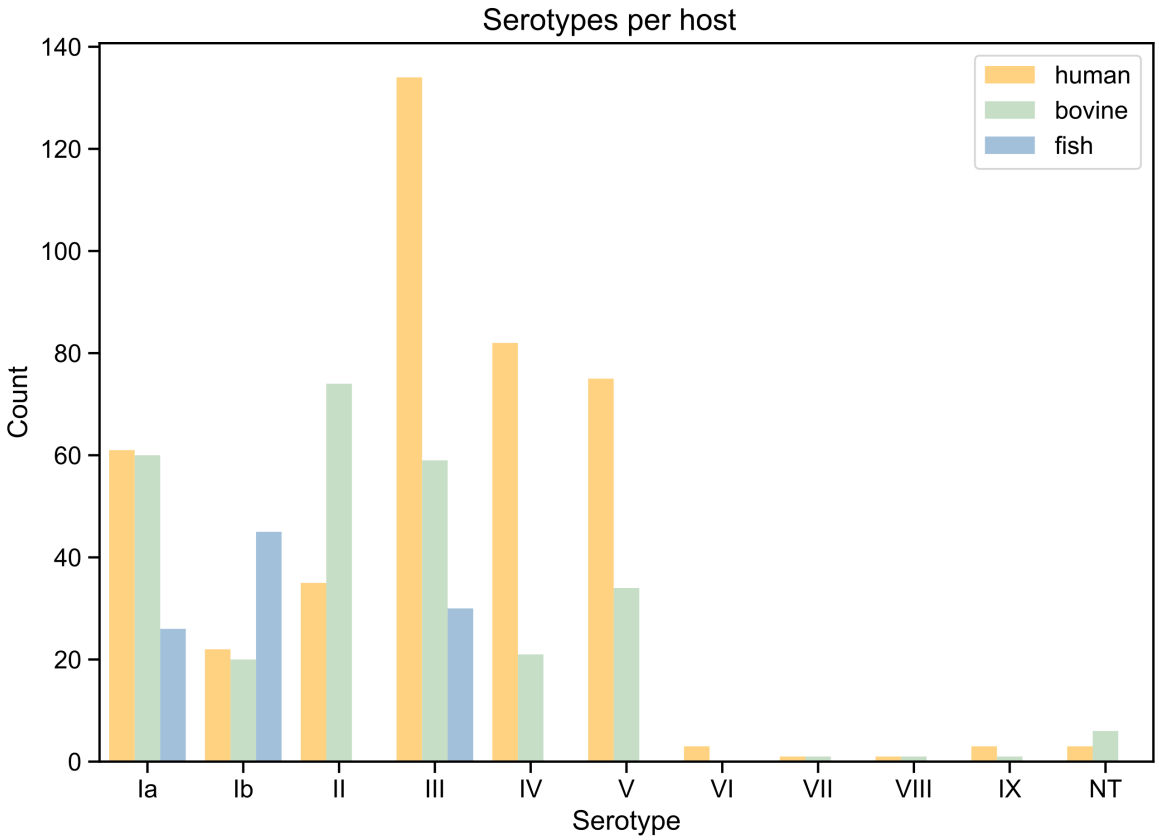


Figure C.8: Frequency plot for group B *Streptococcus* (GBS) genomes included in this study based on serotype. The three major host species are indicated. GBS isolates from humans belonged to all serotypes, and bovine isolates belonged to all types except for serotype VI. Isolates from fish belonged uniquely to three serotypes: Ia (CC7), Ib (CC552) and III (CC283).

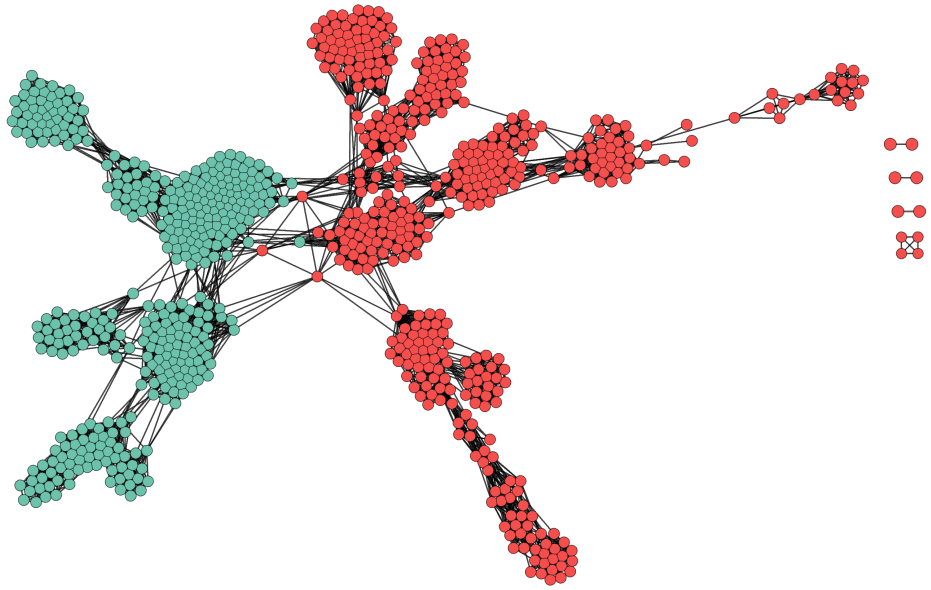


Figure C.9: Network graph of accessory genome distances between 850 group B *Streptococcus* (GBS) isolates. The distinction between host specialists (red) and host generalist (blue) lineages is shown.

C.2 List of commands for bioinformatic analyses

C.2.1 Fastbaps clustering

List of commands used for fastbaps clustering of 850 group B *Streptococcus* genomes from the core SNP alignment generated with snp_sites. Fastbaps was run in RStudio.

```
library(fastbaps)
library(ggtree)
library(ape)
library(phytools)
library(ggplot2)

#LOADING DATA
sparse.data <- fastbaps::import_fasta_sparse_nt("snp_sites.aln")
sparse.data <- optimise_prior(sparse.data, type =
   "optimise.symmetric")
#Result:
#[1] "Optimised hyperparameter: 0.008"

#RUNNING FASTBAPS
#To obtain a Bayesian hierarchical clustering of the data.
baps.hc <- fast_baps(sparse.data, k.init = 213)
#The k.init value should be calculated as number of sequences / 4
   (here 850/4 = 212.5).

#To obtain the partition of this hierarchy under Dirichlet Process
   Mixture model:
best.partition <- best_baps_partition(sparse.data, baps.hc)

#To plot the output of the algorithm directly in R with a
   pre-calculated tree using ggtree:
newick.file.name <- system.file("extdata",
   "snp_sites_midpoint_root.tre", package = "fastbaps")
```

```
iqtree <- phytools::read.newick("snp_sites_midpoint_root.tre")
all.plot.df <- data.frame(id = colnames(sparse.data$snp.matrix),
   fastbaps = best.partition, stringsAsFactors = FALSE)
gg <- ggtree(iqtree)
all.plot <- facet_plot(gg, panel = "fastbaps", data = all.plot.df,
   geom = geom_tile, aes(x = fastbaps), color = "blue")
all.plot

#To save the output of the clustering algorithm:
write.csv(all.plot.df, file="all.fastbaps.clusters.csv")
```

C.2.2 GraPPLE/Graphia

List of commands used for the creation of a network of accessory genome distances of 850 group B *Streptococcus* genomes. After preprocessing with GraPPLE, Graphia was used for visualisation, with the following options: k-MM using edge wight, k=12 and rank order = descending.

```
#To calculate the pairwise similarity between genomes from a
   binary presence/absence gene matrix (generated with panaroo)
python pw_similarity.py -i binary_presc_absc.tsv -o acc_gene_dist
   -r "isolates" -s "jaccard" -t 2 -f 0.8

#Add metadata from a table to a graph in .layout format as
   preprocessing step before Graphia visualisation
python metadata_to_layout.py -l acc_gene_dist_isols_pw_sim.layout
   -m gene_info.tsv -r "copy" -s headers.txt
```

Appendix D

Supporting information Chapter 5

D.1 Tables and figures

Table D.1: Reference genomes used for the annotation of significant *k*-mers and unitigs obtained from pyseer analyses, for each of the three phenotypes tested (human, bovine, fish). Complete genome assemblies were selected, when possible, to represent the diversity of the population belonging to each phenotype. Only two complete bovine group B *Streptococcus* (GBS) genomes are available from the NCBI database to date. No fish ST283 complete genomes are available, therefore a human ST283 was included in the annotation process, as these isolates show little genomic diversity due to the fact that they derive from cases of human invasive disease that develop shortly after a food-borne infection from raw fish consumption (Barkham et al., 2019).

Host species	**Reference genome**	**Sequence type (ST)**	Assembly
Human	CP010867	1	complete
	CP000114	7	complete
	HG939456	17	complete
	CP007570	22	complete
	NC_004116	110	complete
	CP010874	283	complete
Bovine	HF952104	1	complete
	CP008813	103	complete
Fish	NC_018646	7	complete
	CP019802	260	complete
	CP007482	261	complete

STIR-CD-25	283	draft
NZ_CP010874	283 (human)	complete
CP003919	552	complete

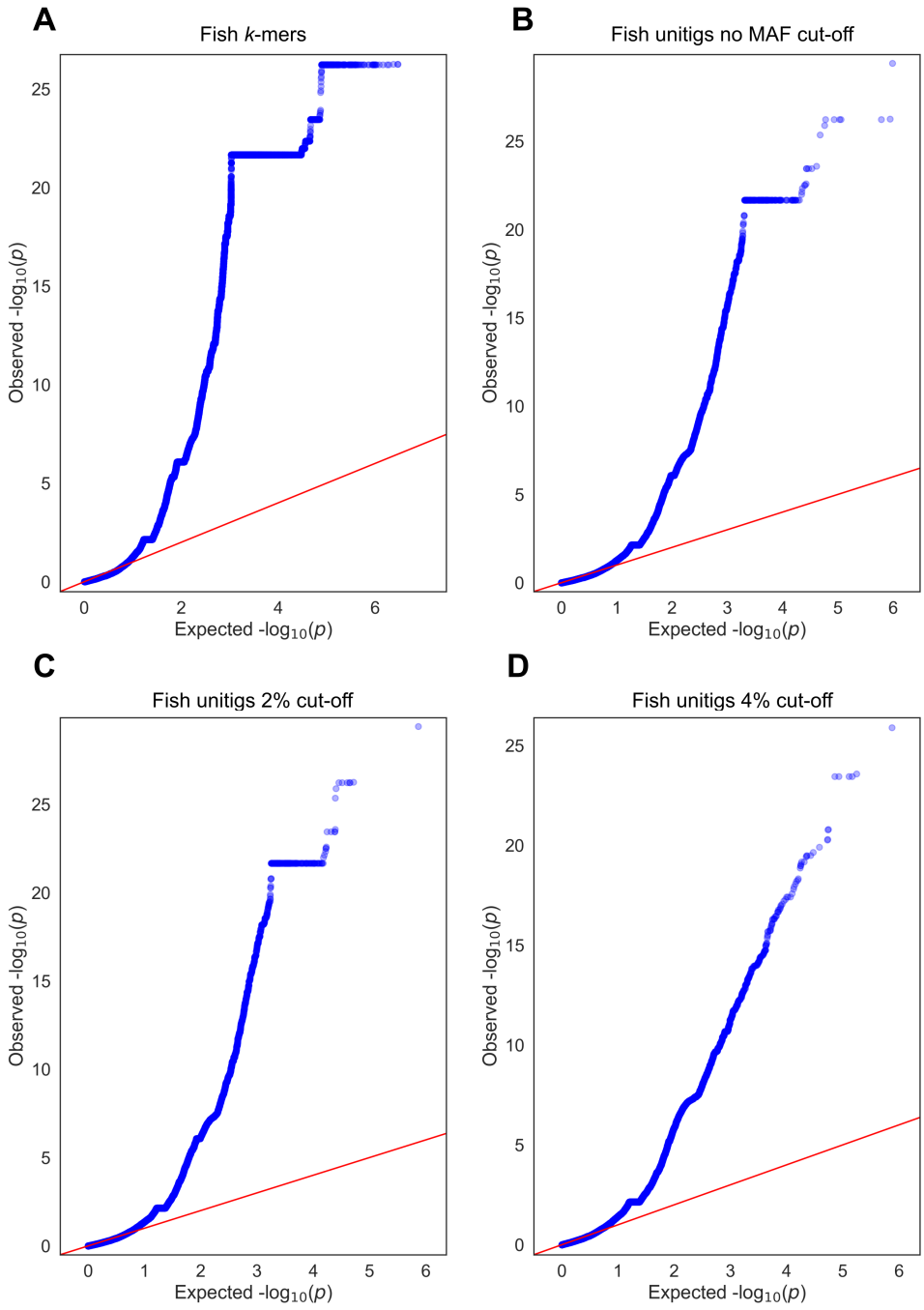


Figure D.1: Quantile-quantile (QQ)-plots comparing expected and observed -$\log_{10}$(p-value) for k-mers and unitigs for the fish phenotype (n=101 genomes). When an association is not present, these values will follow the null line (red). When association is present, points will follow the null line until about 1, and then they will draw a smooth curve above the line. When large 'shelves' are present (A-C), particularly around low p-values, this is symptomatic of poorly controlled confounding due to population structure. Increasing the minor allele frequency (MAF) cut-off to about 4-5% is recommended to try to control for this phenomenon (D). Plots were generated with matplotlib v3.3.2 (Barrett et al., 2005).

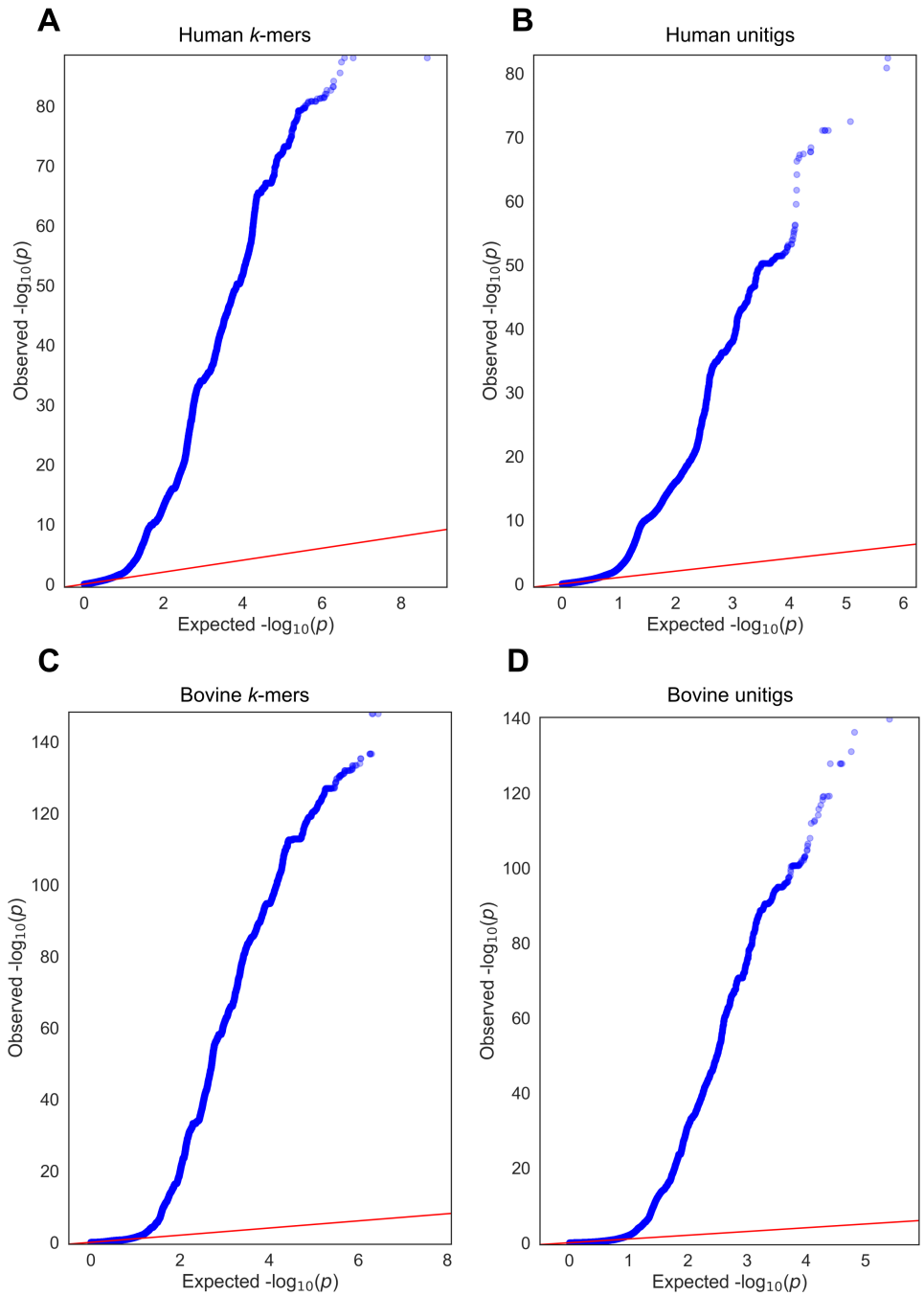


Figure D.2: Quantile-quantile (QQ)-plots comparing expected and observed -$\log_{10}$(p-value) for k-mers and unitigs for the human (n=420 genomes) and the bovine (n=277 genomes) phenotype. When an association is not present, these values will follow the null line (red). When association is present, points will follow the null line until about 1, and then they will draw a smooth curve above the line. Here, no large 'shelves' are present, which suggests that there is no confounding due to population structure. Plots were generated with matplotlib v3.3.2 (Barrett et al., 2005).

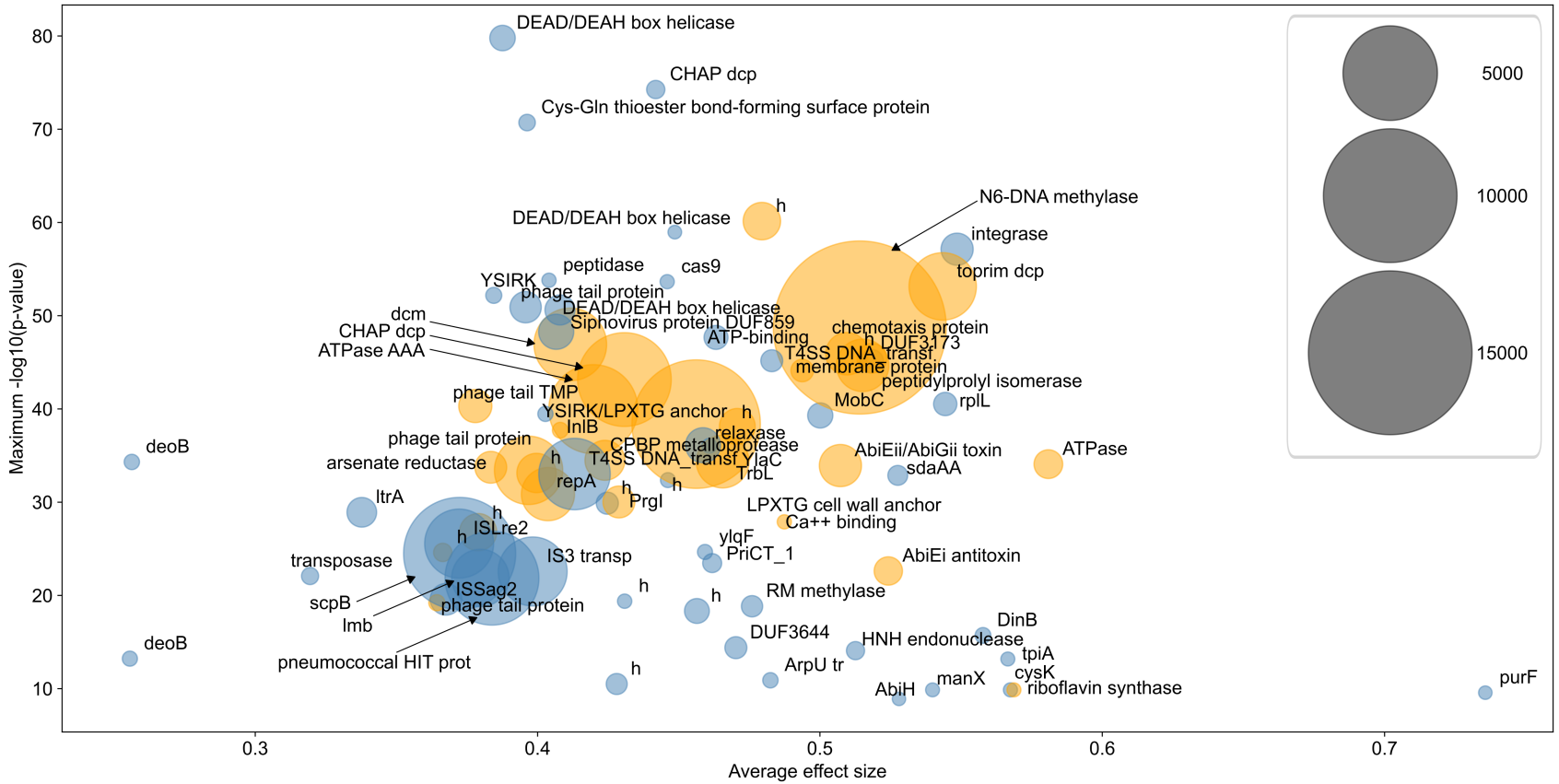


Figure D.3: Significant genes for the human phenotype based on the pyseer *k*-mer analysis. The -$\log_{10}$(*p*-value) has been plotted against the average effect size (beta coefficient). Circle sizes correspond to the number of hits per gene, which is influenced by the size of the gene (the longer the gene, the higher the number of hits). Blue genes are uniquely associated to this phenotype, whereas orange genes are also associated with the bovine host.

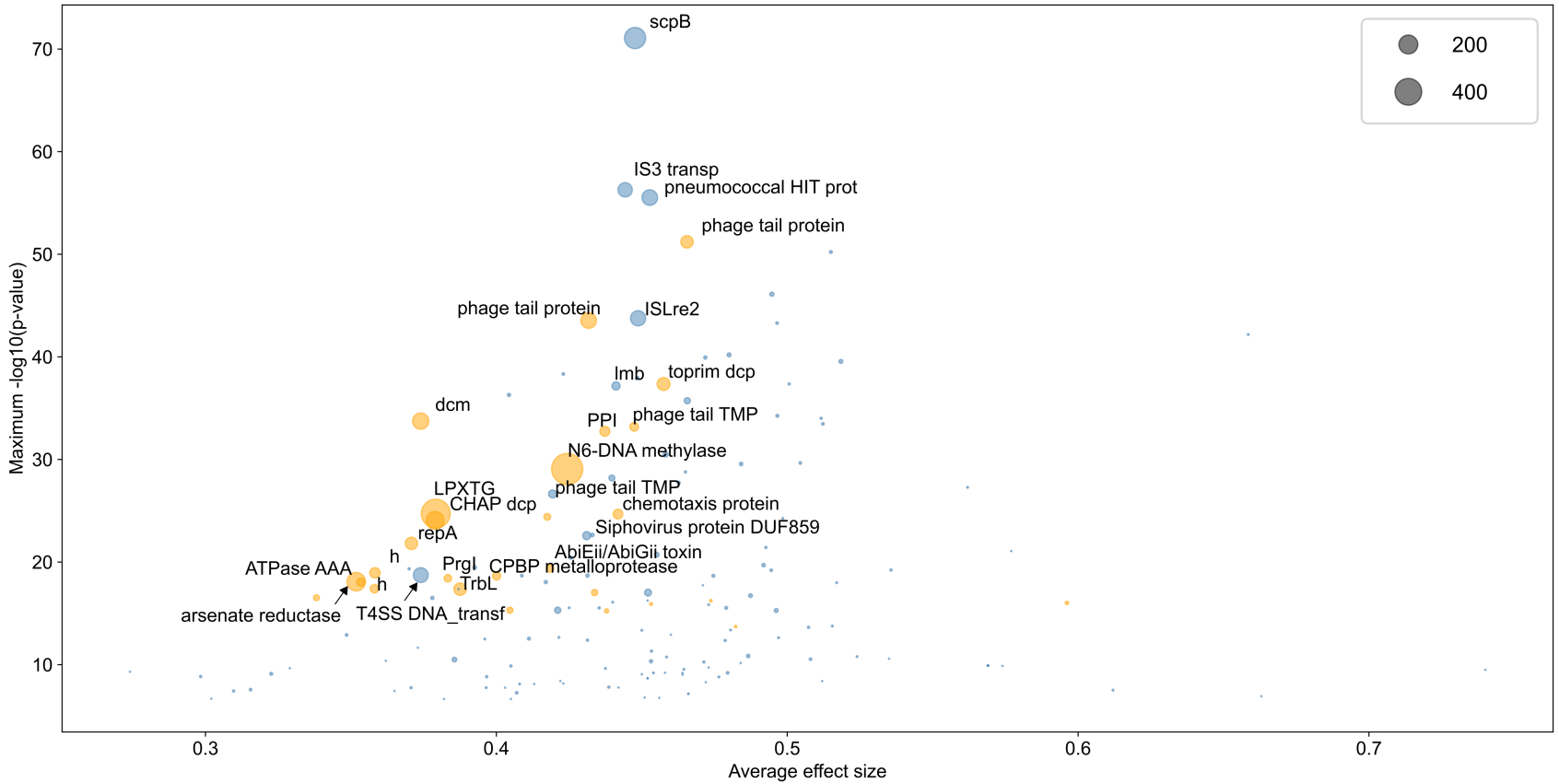


Figure D.4: Significant genes for the human phenotype based on the pyseer unitig analysis. The -$\log_{10}$(p-value) has been plotted against the average affect size (beta coefficient). Circle sizes correspond to the number of hits per gene, which is influenced by the size of the gene (the longer the genome, the higher the number of hits). Blue genes are uniquely associated to this phenotype, whereas orange genes are also associated with the bovine host.

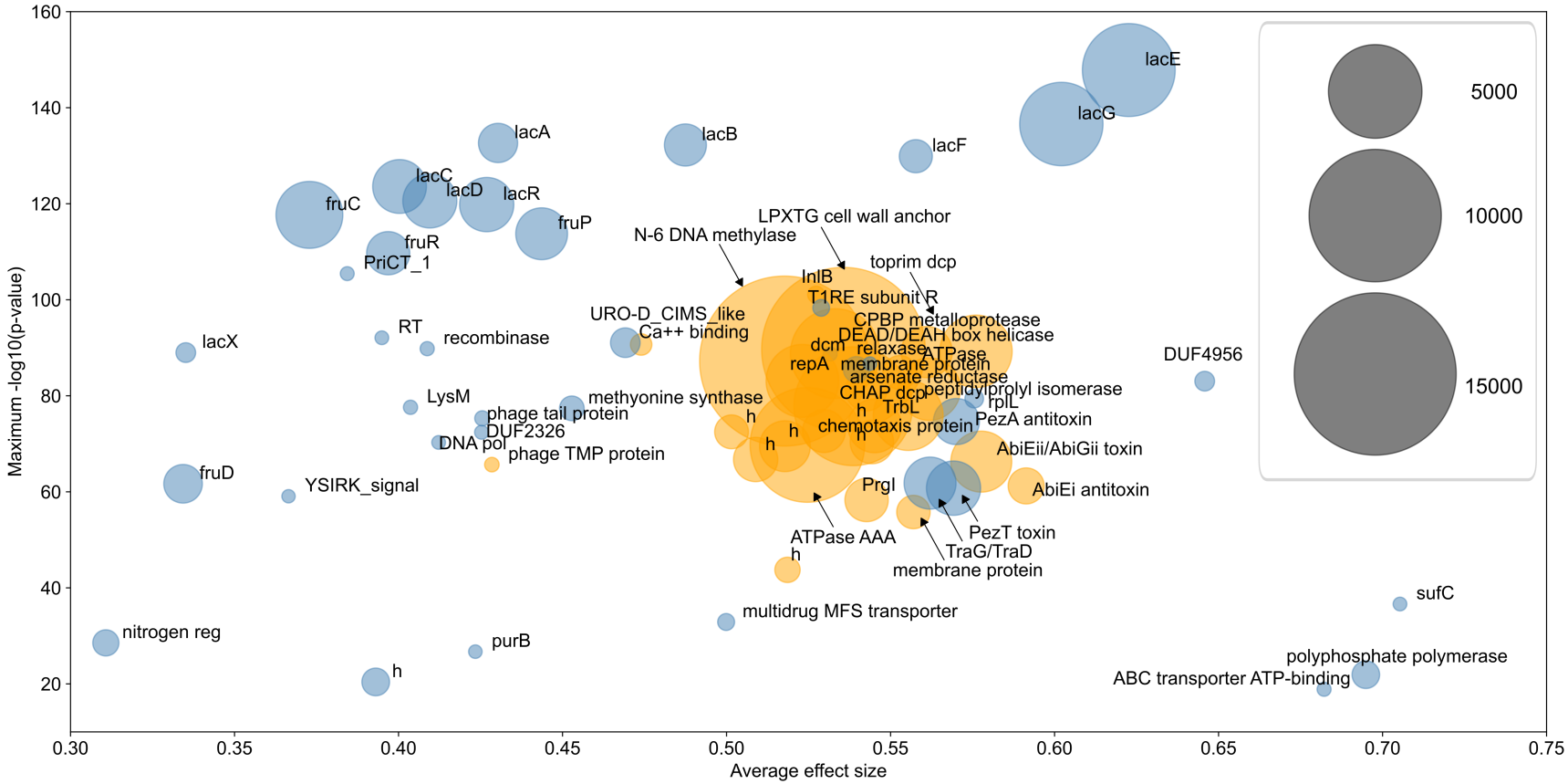


Figure D.5: Significant genes for the bovine phenotype based on the pyseer *k*-mer analysis. The -$\log_{10}$(*p*-value) has been plotted against the average effect size (beta coefficient). Circle sizes correspond to the number of hits per gene, which is influenced by the size of the gene (the longer the gene, the higher the number of hits). Blue genes are uniquely associated to this phenotype, whereas orange genes are also associated with the human host.

Figure D.6: Significant genes for the bovine phenotype based on the pyseer unitig analysis. The -$\log_{10}$(p-value) has been plotted against the average effect size (beta coefficient). Circle sizes correspond to the number of hits per gene, which is influenced by the size of the gene (the longer the gene, the higher the number of hits). Blue genes are uniquely associated to this phenotype, whereas orange genes are also associated with the human host.

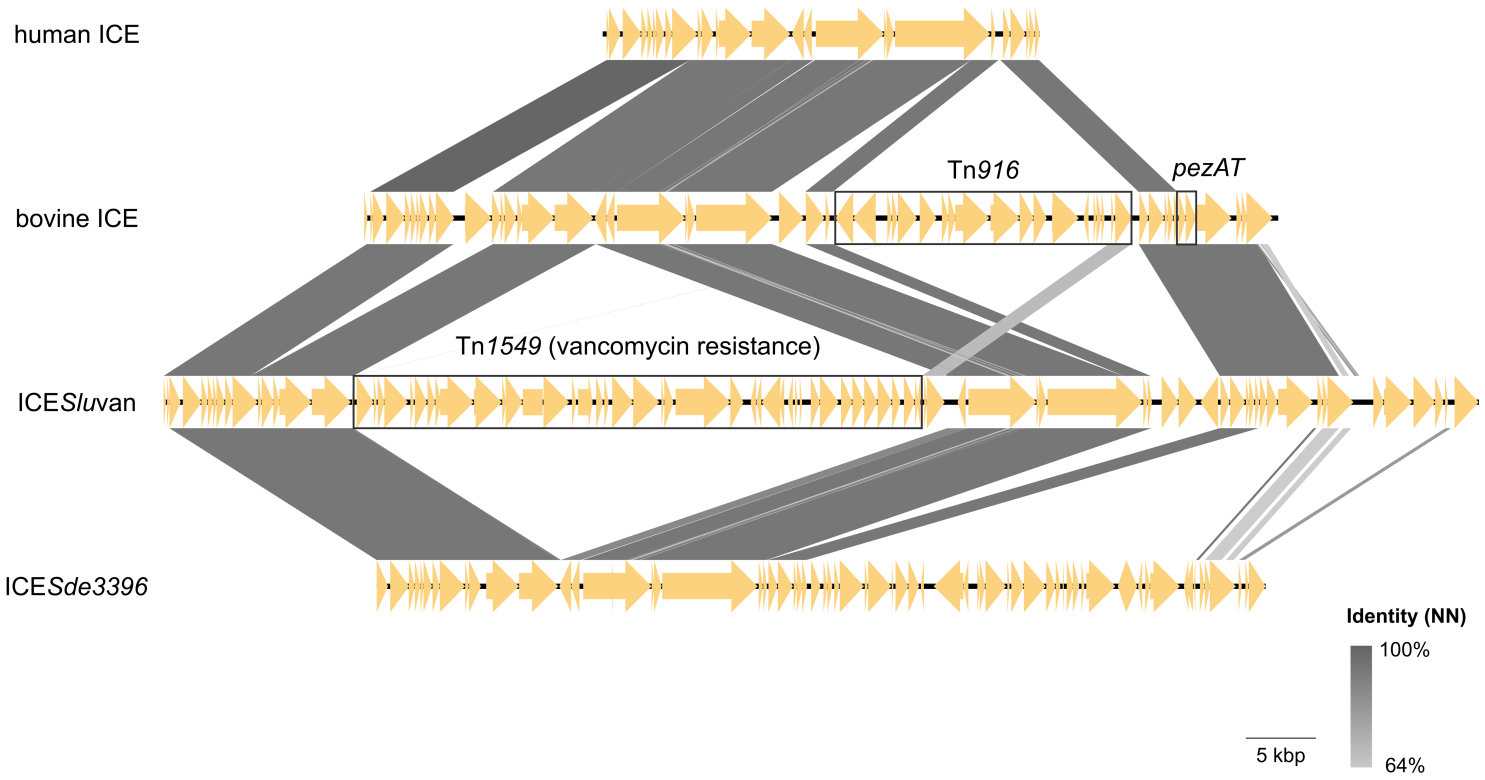


Figure D.7: Visualisation of the blastn comparison between the integrative conjugative elements (ICE) described in this work, obtained with Easyfig v2.2.2 (Sullivan et al., 2011). These comprise: the ICE found by pyseer as significantly associated with both human and bovine hosts, with the latter carrying *pezAT* in most cases, ICE*Slu*van from *Streptococcus lutetiensis*, which shares sequence similarities both with the human and bovine ICE (including *pezAT*), and ICE*Sde3396* from *Streptococcus dysgalactiae* subsp. *equisimilis*.

Figure D.8: Significant genes for the fish phenotype based on the pyseer unitig analysis. The -$\log_{10}$(p-value) has been plotted against the average effect size (beta coefficient). Circle sizes correspond to the number of hits per gene, which is influenced by the size of the gene (the longer the gene, the higher the number of hits). Blue genes are uniquely associated to this phenotype, whereas orange genes are also associated with the human host.

Figure D.9: Frequency distribution of IS*Stin5* (insertion sequence from *Streptococcus iniae*) variants in 44 group B *Streptococcus* genomes from fish isolates, all belonging to the CC552 lineage. Draft genomes are indicated with an asterisk.

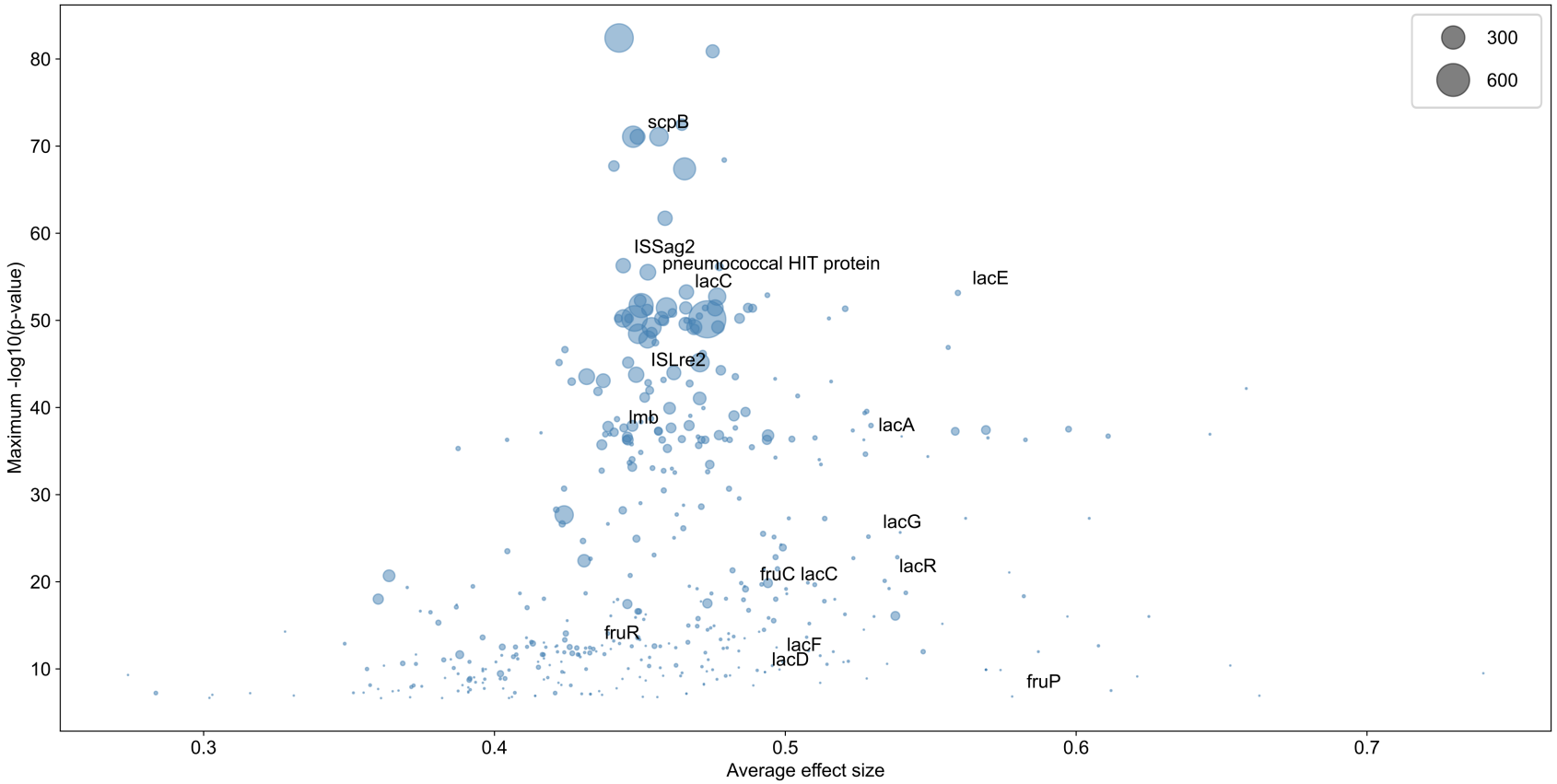


Figure D.10: Visualisation of significant genes for the human phenotype based on the pyseer unitig analysis, when unitigs were annotated with a set of reference genomes from all host species (all genomes in Tab. D.1). The -$\log_{10}$(p-value) has been plotted against the average affect size (beta coefficient). Circle sizes correspond to the number of hits per gene. When comparing this figure to Fig. D.4, it is evident how negatively associated genes (e.g. Lac.2 genes and genes from the ICE) are interfering with the interpretation of the output.

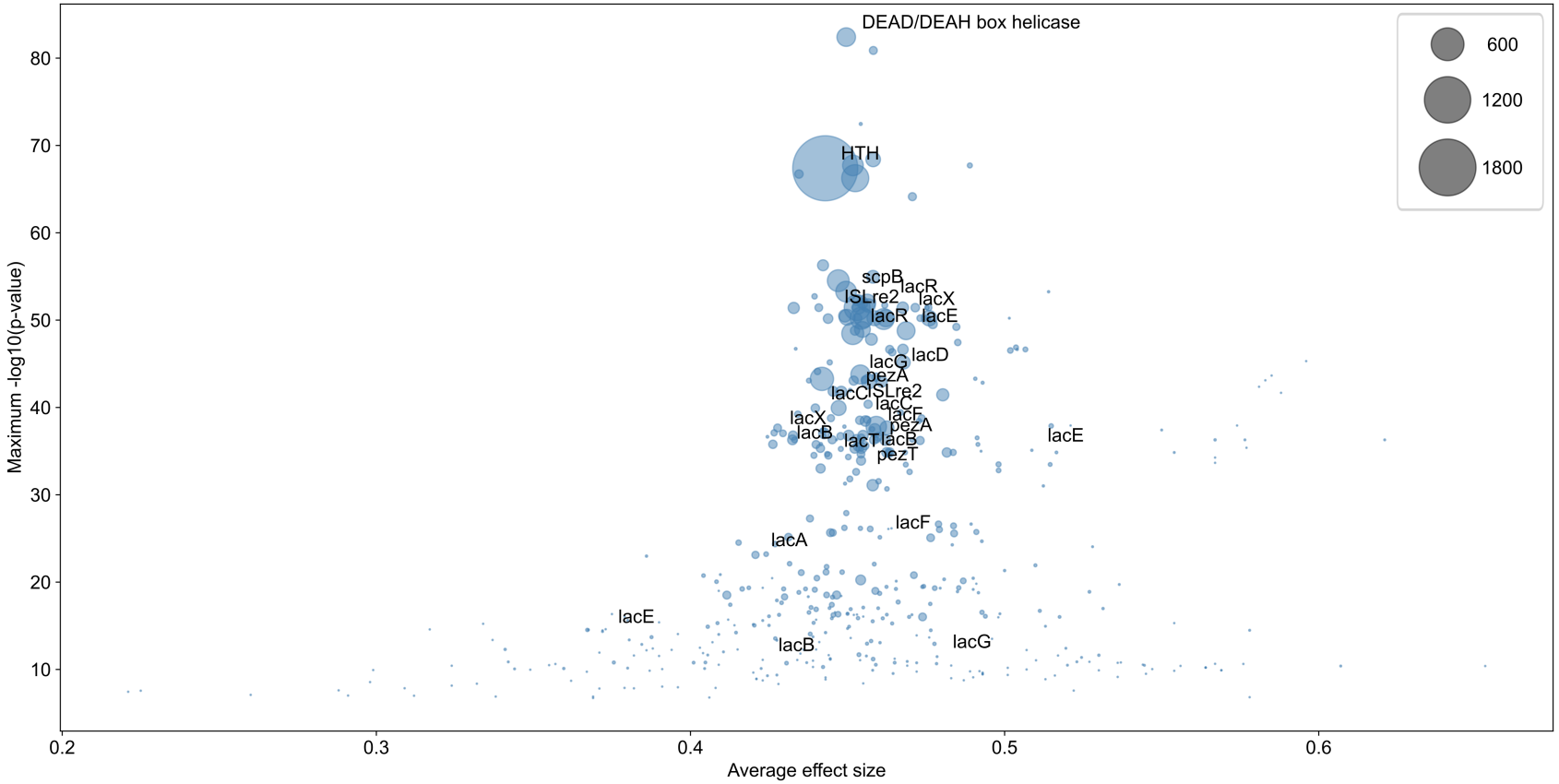


Figure D.11: Visualisation of significant genes for the human phenotype based on the pyseer unitig analysis, when unitigs were annotated with a pangenome generated with roary as reference genome. The -$\log_{10}$(p-value) has been plotted against the average affect size (beta coefficient). Circle sizes correspond to the number of hits per gene. When comparing this figure to Fig. D.4, it is evident how negatively associated genes and different variants of the same gene (e.g. Lac.2 genes, genes from the ICE such as DEAD/DEAH box helicase and HTH) are interfering with the interpretation of the output.

D.2 List of commands for bioinformatic analyses

D.2.1 *k*-mer-based GWAS with pyseer

List of commands used for the *k*-mer-based GWAS for the three different phenotypes (human, bovine, fish).

```
#Count k-mers on all assemblies
fsm-lite -l fsm_file_list.txt -s 6 -S 844 -v -t fsm_kmers
gzip fsm_kmers.txt

#Obtain a kinship matrix from midpoint-rooted core genome tree
   (obtained with IQtree)
phylogeny_distance.py --lmm core_genome_aln.tree > phylogeny_K.tsv

#Main pyseer command (phenotype.pheno corresponded each time to a
   different phenotype)
pyseer --lmm --phenotypes phenotype.pheno --kmers fsm_kmers.txt.gz
   --similarity phylogeny_K.tsv --output-patterns
   kmer_patterns.txt --cpu 24

#Counting k-mer patterns and defining p-value threshold
count_patterns.py kmer_patterns.txt
#prints: Patterns: 2939273
#prints: Threshold: 1.70E-08

#Extract only k-mers that are below the p-value threshold
cat <(head -1 phenotype_kmers.txt) <(awk '$4<1.70E-08 {print $0}'
   phenotype_kmers.txt) > significant_kmers.txt

#Annotation of significant k-mers to a list of reference genomes
   (both closed and draft)
annotate_hits_pyseer.py significant_kmers.txt references.txt
   annotated_kmers.txt
```

```
#Obtain a summary of significant genes
summarise_annotations.py annotated_kmers.txt > gene_hits.txt
```

D.2.2 Unitig-based GWAS with pyseer

List of commands used for the unitig-based GWAS for the three different phenotypes (human, bovine, fish).

```
#Count unitigs on all assemblies
unitig-counter -strains strain_list.txt -output unitigs_pyseer
   -nb-cores 24

#Main pyseer command (phenotype.pheno corresponded each time to a
   different phenotype); phylogeny_K.tsv is the same file obtained
   above
pyseer --lmm --phenotypes phenotype.pheno --kmers unitigs.txt.gz
   --similarity phylogeny_K.tsv --output-patterns
   unitigs_patterns.txt --cpu 24

#Counting unitig patters and defining p-value threshold
count_patterns.py unitigs_patterns.txt
#prints: Patterns: 221590
#prints: Threshold: 2.26E-07

#Extract only unitigs that are below the p-value threshold
cat <(head -1 phenotype_unitigs.txt) <(awk '$4<2.26E-07 {print
   $0}' phenotype_unitigs.txt) > significant_unitigs.txt

#Annotation of significant unitigs to a list of reference genomes
   (both closed and draft)
annotate_hits_pyseer.py significant_unitigs.txt references.txt
   annotated_unitigs.txt
```

```
#Obtain a summary of significant genes
summarise_annotations.py annotated_unitigs.txt > gene_hits.txt
```

Appendix E

Supporting information Chapter 6

E.1 Tables and figures

Table E.1: Dataset of 122 group B *Streptococcus* (GBS) isolates from Kenyan camels. Metadata including isolate names, source of isolation (milk, rectum, nose, mouth), age group (adult, calf), year of isolation, herd and county of origin are shown. Results for sequence type (ST), serotype, antimicrobial resistance genes (AMR), integrative element Tn*916* (presence=1, absence=0; 1* indicates that Tn*916* also carried a *tet*(L) gene), Lac.2 variant, prophage types (as per integrase type classification described in Crestani et al., 2020; 'R' stands for remnant, or incomplete prophage; ϕ1207.3 is not included in this count), PICI and Tn*916*-associated prophage ϕ1207.3 (presence=1, absence=0) are reported.

ISOLATE	SOURCE	AGE	YEAR	ST	SEROTYPE	AMR	Tn*916*	LAC.2	PROPHAGES	PICI	ϕ1207.3	HERD	COUNTY
A4	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	1	Isiolo
A5	milk	adult	2017	616	III	*tet*(M)	1		1 R		1	1	Isiolo
A6	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	1	Isiolo
A7	milk	adult	2017	1652	VI	*tet*(M)	1		1 R	PICI4	1	1	Isiolo
11N	nose	adult	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	2	Laikipia
13R	rectum	calf	2019	615	II		0		GBS11.1, 1 R		0	2	Laikipia
19M	milk	adult	2019	1652	VI	*tet*(M)	1		3 R	PICI4	1	2	Laikipia
19N	nose	adult	2019	612	IV		0		GBS11.1, 3 R		0	2	Laikipia
1M	milk	adult	2019	616	III	*tet*(M)	1	Lac.2d	GBS11.2, 1 R		1	2	Laikipia
23N	nose	adult	2019	612	IV		0		1 R		1	2	Laikipia
3N	nose	adult	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	2	Laikipia
5N	nose	adult	2019	615	II		0		GBS11.1, 1 R		0	2	Laikipia
9R	rectum	calf	2019	1652	VI	*tet*(M)	1		3 R	PICI4	1	2	Laikipia
B1	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	GBS10, 1 R		1	3	Meru
B10	milk	adult	2017	1652	VI	*tet*(M)	1	Lac.2d	1 R	PICI4	1	3	Meru
B5	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	GBS10, GBS11.2, 1 R		1	3	Meru
B6	milk	adult	2017	616	III	*tet*(M)	1	Lac.2b	1 R		1	3	Meru
B7	milk	adult	2017	616	III	*tet*(M)	1				1	3	Meru
27M	milk	adult	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	4	Laikipia
27R	rectum	calf	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	4	Laikipia
34N	nose	adult	2019	617	VI	*tet*(M)	1		GBS11.1, 2 R		0	4	Laikipia
39M	milk	adult	2019	1652	VI	*tet*(M)	1		GBS11.1 , 1 R		1	4	Laikipia
39O	mouth	calf	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	4	Laikipia

Table E.1 continued from previous page

ISOLATE	SOURCE	AGE	YEAR	ST	SEROTYPE	AMR	Tn*916*	LAC.2	PROPHAGES	PICI	ϕ1207.3	HERD	COUNTY
42N	nose	adult	2019	617	VI	*tet*(M)	1		GBS11.1, 1 R		0	4	Laikipia
44M	milk	adult	2019	1652	VI	*tet*(M)	1	Lac.2b	GBS11.1	PICI4	1	4	Laikipia
44N	nose	adult	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	4	Laikipia
46M	milk	adult	2019	617	VI	*tet*(M)	1		6 R		0	4	Laikipia
C34N	nose	calf	2019	617	VI	*tet*(M)	1		GBS11.1, 2 R		0	4	Laikipia
C36N	nose	calf	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	4	Laikipia
C39N	nose	calf	2019	1652	VI	*tet*(M)	1		GBS11.1, 1 R		1	4	Laikipia
C5	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e	1 R		1	5	Meru
C8	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e	1 R		1	5	Meru
51N	nose	adult	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	6	Laikipia
56N	nose	adult	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	6	Laikipia
60M	milk	adult	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	6	Laikipia
64N	nose	adult	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	6	Laikipia
C52N	nose	calf	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	6	Laikipia
C54N	nose	calf	2019	1652	VI	*tet*(M)	1		GBS11.1, 1 R		1	6	Laikipia
C57N	nose	calf	2019	1652	VI	*tet*(M)	1		GBS11.1	PICI4	1	6	Laikipia
C61N	nose	calf	2019	617	VI	*tet*(M)	1		1 R	PICI4	0	6	Laikipia
D4	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	GBS10, 1 R		1	7	Isiolo
D5	milk	adult	2017	1652	VI	*tet*(M)	1	Lac.2b	2 R		1	7	Isiolo
72N	nose	adult	2019	617	VI	*tet*(M), *tet*(L)	1*		2 R	PICI4	0	8	Laikipia
73M	milk	adult	2019	616	III	*tet*(M)	1	Lac.2d	1 R		1	8	Laikipia
73N	nose	adult	2019	617	VI	*tet*(M), *tet*(L)	1*		2 R	PICI4	0	8	Laikipia
78M	milk	adult	2019	616	III	*tet*(M)	1	Lac.2d	1 R		1	8	Laikipia
78N	nose	adult	2019	612	IV		0		GBS11.1, 3 R		0	8	Laikipia
79M	milk	adult	2019	616	III	*tet*(M)	1	Lac.2d	1 R		1	8	Laikipia
79N	nose	adult	2019	617	VI	*tet*(M), *tet*(L)	1*		2 R	PICI4	0	8	Laikipia
C74N	nose	calf	2019	617	VI	*tet*(M), *tet*(L)	1*	Lac.2d	3 R		0	8	Laikipia
E10	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e	1 R		1	9	Meru
E6	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e	1 R		1	9	Meru

Table E.1 continued from previous page

ISOLATE	SOURCE	AGE	YEAR	ST	SEROTYPE	AMR	Tn*916*	LAC.2	PROPHAGES	PICI	ϕ1207.3	HERD	COUNTY
82N	nose	adult	2019	612	IV		0		GBS11.1, 3 R		0	10	Laikipia
83O	mouth	calf	2019	612	IV		0		GBS11.1, 3 R		0	10	Laikipia
84N	nose	adult	2019	617	IV		0			PICI2	0	10	Laikipia
84Ob	mouth	calf	2019	617	IV		0			PICI3	0	10	Laikipia
87N	nose	adult	2019	612	IV		0		GBS11.1, 3 R		0	10	Laikipia
C84Na	nose	calf	2019	617	IV		0			PICI3	0	10	Laikipia
C84Nb	nose	calf	2019	617	VI		0		2 R	PICI4	0	10	Laikipia
C87N	nose	calf	2019	617	VI		0		2 R	PICI4	0	10	Laikipia
F7	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e	1 R		1	11	Meru
88N	nose	adult	2019	617	IV		0		GBS10, GBS11.1, 2 R	PICI2	0	12	Laikipia
89N	nose	adult	2019	617	IV		0		GBS10, GBS11.1, 1 R	PICI2	0	12	Laikipia
89O	mouth	calf	2019	615	II		0		GBS11.1, 1 R		0	12	Laikipia
90N	nose	adult	2019	617	IV		0		GBS10, GBS11.1, 2 R	PICI2	0	12	Laikipia
90O	mouth	calf	2019	616	III	*tet*(M)	1		GBS10, 1 R		1	12	Laikipia
91N	nose	adult	2019	615	II		0		GBS11.1, 1 R		0	12	Laikipia
91O	mouth	calf	2019	615	II		0		GBS11.1, 1 R		0	12	Laikipia
92N	nose	adult	2019	615	II		0		GBS11.1, 1 R		0	12	Laikipia
92O	mouth	calf	2019	615	II		0		GBS11.1, 2 R		0	12	Laikipia
C90N	nose	calf	2019	615	II		0		GBS11.1, 1 R		0	12	Laikipia
C91N	nose	calf	2019	615	II		0		GBS11.1, 1 R		0	12	Laikipia
C92N	nose	calf	2019	615	II		0		GBS11.1, 1 R		0	12	Laikipia
G6	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	13	Meru
G9	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	13	Meru
H10	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	GBS10, 1 R		1	14	Meru
I12	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e	1 R		1	15	Meru
I13	milk	adult	2017	616	III	*tet*(M)	1		1 R		1	15	Meru
I2	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e	1 R		1	15	Meru
I4	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e	1 R		1	15	Meru
I8	milk	adult	2017	616	III	*tet*(M)	1		GBS10, 1 R		1	15	Meru

Table E.1 continued from previous page

ISOLATE	SOURCE	AGE	YEAR	ST	SEROTYPE	AMR	Tn*916*	LAC.2	PROPHAGES	PICI	ϕ1207.3	HERD	COUNTY
J10	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	16	Meru
J2	milk	adult	2017	612	IV		0		1 R		0	16	Meru
J3	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	16	Meru
J5	milk	adult	2017	616	III	*tet*(M)	1		1 R		1	16	Meru
J6	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	16	Meru
J7	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	2 R		0	16	Meru
J9	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e			1	16	Meru
K1	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	GBS10, 2 R		1	17	Meru
K2	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e	1 R		1	17	Meru
K4	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	GBS10, 1 R		1	17	Meru
K6	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	17	Meru
K7	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	18	Meru
L3	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d			1	18	Meru
M10	milk	adult	2017	616	III	*tet*(M)	1	Lac.2b	1 R		1	19	Meru
M2	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e	1 R		1	19	Meru
M6	milk	adult	2017	616	III	*tet*(M)	1	Lac.2b	1 R		1	19	Meru
M7	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e			1	19	Meru
M9	milk	adult	2017	616	III	*tet*(M)	1	Lac.2b	GBS10, 1 R		1	19	Meru
N1	milk	adult	2017	616	III	*tet*(M)	1	Lac.2b	GBS10, 1 R		1	20	Meru
N5	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	20	Meru
N8	milk	adult	2017	612	IV		0		GBS11.1, 3 R		0	20	Meru
N9	milk	adult	2017	616	III	*tet*(M)	1		GBS10, 1 R		1	20	Meru
O1	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e			1	21	Meru
O11	milk	adult	2017	616	III	*tet*(M)	1		GBS10, 1 R		1	21	Meru
O4	milk	adult	2017	616	III	*tet*(M)	1		1 R		1	21	Meru
O5	milk	adult	2017	616	III	*tet*(M)	1		GBS10, 1 R		1	21	Meru
P1	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	GBS11.2, 2 R		1	22	Meru
P2	milk	adult	2017	1652	VI	*tet*(M)	1	Lac.2d	1 R	PICI4	1	22	Meru
P3	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	GBS10, GBS11.2, 1 R		1	22	Meru

Table E.1 continued from previous page

ISOLATE	SOURCE	AGE	YEAR	ST	SEROTYPE	AMR	Tn*916*	LAC.2	PROPHAGES	PICI	ϕ1207.3	HERD	COUNTY
P4	milk	adult	2017	1	V	*tet*(M)	1				1	22	Meru
P8	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	22	Meru
Q5	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e	1 R		1	23	Meru
R1	milk	adult	2017	616	III	*tet*(M)	1	Lac.2b	1 R		1	24	Meru
R10	milk	adult	2017	1652	VI	*tet*(M)	1		1 R	PICI4	1	24	Meru
R2	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	24	Meru
R3	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	1 R		1	24	Meru
R4	milk	adult	2017	616	III	*tet*(M)	1	Lac.2e			1	24	Meru
R5	milk	adult	2017	1652	VI	*tet*(M)	1		1 R	PICI4	1	24	Meru
R8	milk	adult	2017	1654	III	*tet*(M)	1		GBS10, 1 R		1	24	Meru
R9	milk	adult	2017	616	III	*tet*(M)	1	Lac.2d	2 R		1	24	Meru
T5	milk	adult	2017	1653	III	*tet*(M)	1	Lac.2b	1 R		1	25	Meru

Table E.2: Table comparing prevalences of tetracycline resistance (TcR) in group B *Streptococcus* from camel samples (milk vs non-milk) between the present study (sequence data from Kenya from Seligsohn et al., 2021a, and Seligsohn et al., 2021b) and Fisher et al., 2013 (data from Kenya and Somalia).

	Origin	**TcR prevalence (%)**	
		Milk	**Non-Milk**
This study	Kenya	97.3	48.9
Fischer et al., 2013	Kenya	72.4	20.4
	Somalia	0	na
	Overall	48.8	20.4

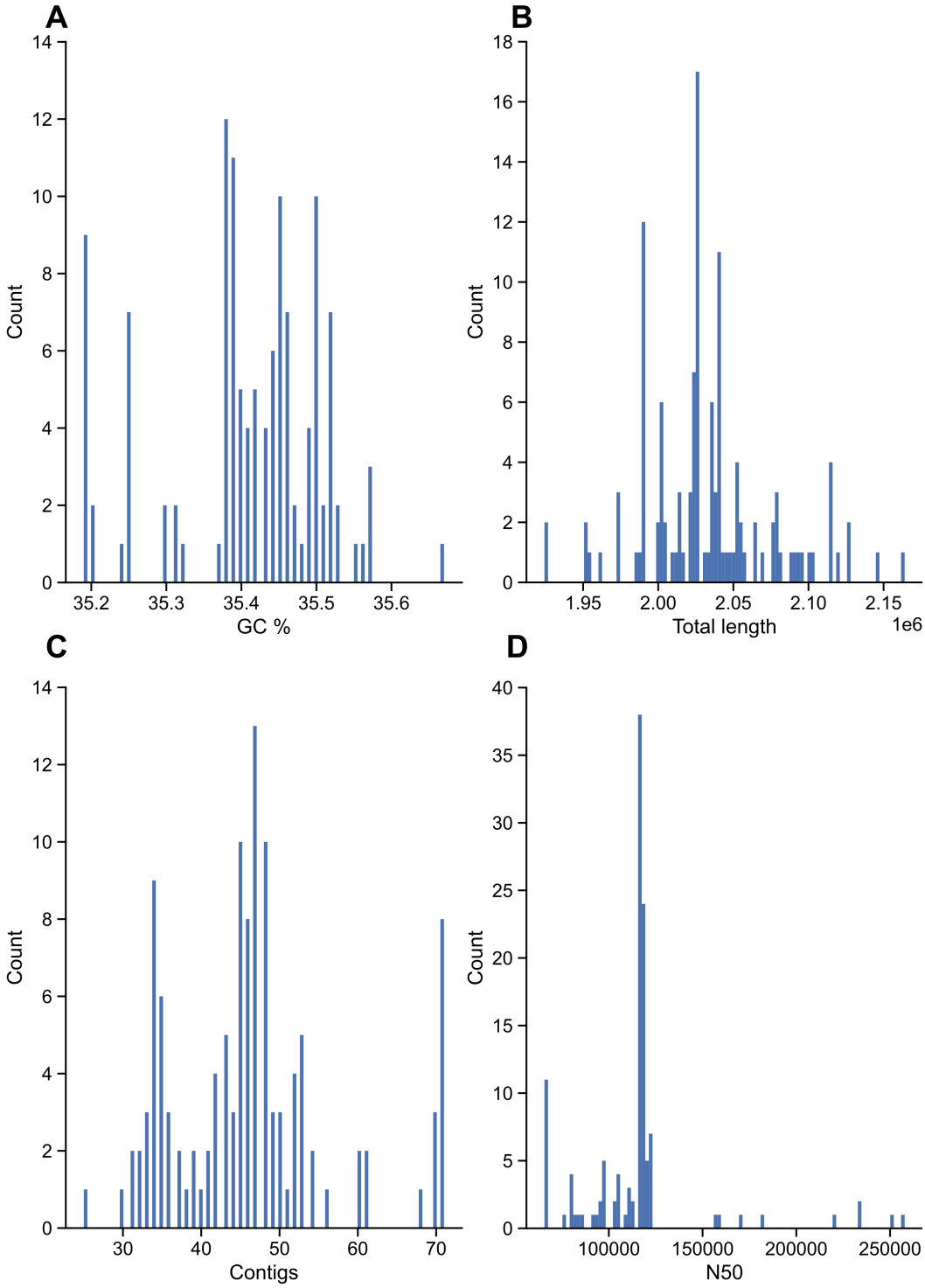


Figure E.1: Frequency distribution plots for 122 group B *Streptococcus* genomes from Kenyan camels. Panels (A-D) show the following data: GC content (%), total length of the genome (Mbp), total number of contigs and N50. Plots were generated with matplotlib v3.3.2 (Barrett et al., 2005).

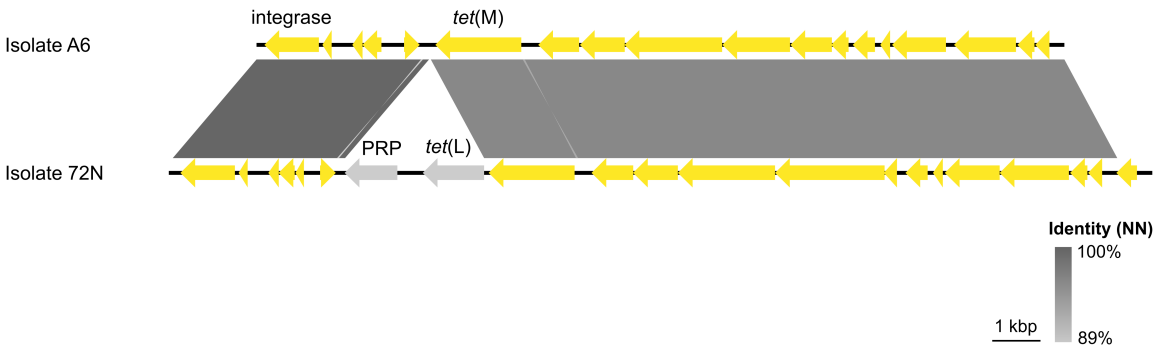


Figure E.2: Diagram of Tn*916* elements detected in 122 group B *Streptococcus* (GBS) genomes from Kenyan camels. The majority of Tn*916*-positive genomes coded for a variant of Tn*916* very similar to the reference (>99% ID and 100% QC), as shown in isolate A6. Four genomes carried a variant with two additional genes: a plasmid replication protein (PRP) and *tet*(L) gene, as shown in isolate 72N.

E.2 List of commands for bioinformatic analyses

E.2.1 Fastbaps clustering

List of commands used for fastbaps clustering of 122 group B *Streptococcus* genomes from the core alignment generated with snippy. Fastbaps was run in RStudio.

```
library(fastbaps)
library(ggtree)
library(ape)
library(phytools)
library(ggplot2)

#LOADING DATA
sparse.data <- fastbaps::import_fasta_sparse_nt("core.aln")
sparse.data <- optimise_prior(sparse.data, type =
   "optimise.symmetric")
#Result:
#[1] "Optimised hyperparameter: 0.002"

#RUNNING FASTBAPS
#To obtain a Bayesian hierarchical clustering of the data.
baps.hc <- fast_baps(sparse.data, k.init = 30)
#The k.init value should be calculated as number of sequences / 4
   (here 850/4 = 212.5).

#To obtain the partition of this hierarchy under Dirichlet Process
   Mixture model:
best.partition <- best_baps_partition(sparse.data, baps.hc)

#To plot the output of the algorithm directly in R with a
   pre-calculated tree using ggtree:
newick.file.name <- system.file("extdata",
   "core_genome_tree_camel.tre", package = "fastbaps")
```

```
iqtree <- phytools::read.newick("core_genome_tree_camel.tre")
all.plot.df <- data.frame(id = colnames(sparse.data$snp.matrix),
   fastbaps = best.partition, stringsAsFactors = FALSE)
gg <- ggtree(iqtree)
all.plot <- facet_plot(gg, panel = "fastbaps", data = all.plot.df,
   geom = geom_tile, aes(x = fastbaps), color = "blue")
all.plot

#To save the output of the clustering algorithm:
write.csv(all.plot.df, file="camel.fastbaps.clusters.csv")
```

References

Agnew, W., & Barnes, A. C. (2007). *Streptococcus iniae*: an aquatic pathogen of global veterinary significance and a challenging candidate for reliable vaccination. *Veterinary Microbiology*, *122*(1-2), 1–15.

Alexander, H. L., Richardson, A. R., & Stojiljkovic, I. (2004). Natural transformation and phase variation modulation in *Neisseria meningitidis*. *Molecular Microbiology*, *52*(3), 771–783.

Allaire, J. (2012). RStudio: integrated development environment for R. *Boston, MA*, *770*(394), 165–171.

Almeida, A., Alves-Barroco, C., Sauvage, E., Bexiga, R., Albuquerque, P., Tavares, F., ... Glaser, P. (2016). Persistence of a dominant bovine lineage of group B *Streptococcus* reveals genomic signatures of host adaptation. *Environmental Microbiology*, *18*(11), 4216–4229.

Alvarez-Martinez, C. E., & Christie, P. J. (2009). Biological diversity of prokaryotic type IV secretion systems. *Microbiology and Molecular Biology Reviews*, *73*(4), 775–808.

Amal, M. N. A., & Zamri-Saad, M. (2011). Streptococcosis in tilapia (*Oreochromis niloticus*): a review. *Pertanika Journal of Tropical Agricultural Science*, *34*(2), 195–206.

Ambroset, C., Coluzzi, C., Guédon, G., Devignes, M.-D., Loux, V., Lacroix, T., ... Leblond-Bourget, N. (2016). New insights into the classification and integration specificity of *Streptococcus* integrative conjugative elements through extensive genome exploration. *Frontiers in Microbiology*, *6*, 1483.

Amgarten, D. E., Braga, L. P. P., Da Silva, A. M., & Setubal, J. C. (2018). MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Frontiers in Genetics*, *9*(7), 304.

Andersen, H. J., Pedersen, L. H., Aarestrup, F. M., & Chriél, M. (2003). Evaluation of the

surveillance program of *Streptococcus agalactiae* in Danish dairy herds. *Journal of Dairy Science*, *86*(4), 1233–1239.

Argemi, X., Matelska, D., Ginalski, K., Riegel, P., Hansmann, Y., Bloom, J., … Prévost, G. (2018). Comparative genomic analysis of *Staphylococcus lugdunensis* shows a closed pan-genome and multiple barriers to horizontal gene transfer. *BMC Genomics*, *19*(1), 1–16.

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research*, *44*(W1), W16–W21.

Asakura, Y., Kojima, H., & Kobayashi, I. (2011). Evolutionary genome engineering using a restriction–modification system. *Nucleic Acids Research*, *39*(20), 9034–9046.

Asencios, Y. O., Sánchez, F. B., Mendizábal, H. B., Pusari, K. H., Alfonso, H. O., Sayán, A. M., … Chaupe, N. S. (2016). First report of *Streptococcus agalactiae* isolated from *Oreochromis niloticus* in Piura, Peru: Molecular identification and histopathological lesions. *Aquaculture Reports*, *4*, 74–79.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., … Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, *9*(1), 75.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevznernal, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477.

Bapteste, E., Boucher, Y., Leigh, J., & Doolittle, W. F. (2004). Phylogenetic reconstruction and lateral gene transfer. *Trends in Microbiology*, *12*(9), 406–411.

Barkham, T., Zadoks, R. N., Azmai, M. N. A., Baker, S., Bich, V. T. N., Chalker, V., … Chen, S. L. (2019). One hypervirulent clone, sequence type 283, accounts for a large proportion of invasive *Streptococcus agalactiae* isolated from humans and diseased tilapia in Southeast Asia. *PLoS Neglected Tropical Diseases*, *13*(6), e0007421.

Barony, G. M., Tavares, G. C., Pereira, F. L., Carvalho, A. F., Dorella, F. A., Leal, C. A. G., & Figueiredo, H. C. P. (2017). Large-scale genomic analyses reveal the population structure and evolutionary trends of *Streptococcus agalactiae* strains in Brazilian fish farms. *Scientific Reports*, *7*(1), 13538.

Barrangou, R., & Dudley, E. G. (2016). CRISPR-based typing and next-generation tracking technologies. *Annual Review of Food Science and Technology*, *7*, 395–411.

Barrett, P., Hunter, J., Miller, J. T., Hsu, J. C., & Greenfield, P. (2005). matplotlib–a portable Python plotting package. In *Astronomical Data Analysis Software and Systems XIV* (Vol. 347, p. 91).

Behrooz, S. K., Lida, L., Ali, S., Mehdi, M., Rasoul, M., Elnaz, O., … Gholamreza, I. (2018). Study of MazEF, sam, and phd-doc putative toxin–antitoxin systems in *Staphylococcus epidermidis*. *Acta Microbiologica et Immunologica Hungarica*, *65*(1), 81–91.

Bekele, T., & Molla, B. (2001). Mastitis in lactating camels (*Camelus dromedarius*) in Afar Region, north-eastern Ethiopia. *Berliner und Münchener Tierärztliche Wochenschrift*, *114*(5-6), 169–172.

Belhocine, K., Plante, I., & Cousineau, B. (2004). Conjugation mediates transfer of the Ll.LtrB group II intron between different bacterial species. *Molecular Microbiology*, *51*(5), 1459–1469.

Bellais, S., Six, A., Fouet, A., Longo, M., Dmytruk, N., Glaser, P., … Poyart, C. (2012). Capsular switching in group B *Streptococcus* CC17 hypervirulent clone: a future challenge for polysaccharide vaccine development. *The Journal of Infectious Diseases*, *206*(11), 1745–1752.

Bellanger, X., Payot, S., Leblond-Bourget, N., & Guédon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiology Reviews*, *38*(4), 720–760.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Bennett, S. (2004). Solexa ltd. *Pharmacogenomics*, *5*(4), 433–438.

Bergman, R., Nerlich, A., Chhatwal, G., & Nitsche-Schmitz, D. (2014). Distribution of small native plasmids in *Streptococcus pyogenes* in India. *International Journal Medical Microbiology*, *304*(3-4), 370–378.

Bisharat, N., Crook, D. W., Leigh, J., Harding, R. M., Ward, P. N., Coffey, T. J., … Jones, N. (2004). Hyperinvasive neonatal group B *Streptococcus* has arisen from a bovine

ancestor. *Journal of Clinical Microbiology*, *42*(5), 2161–2167.

Bishop, E. J., Shilton, C., Benedict, S., Kong, F., Gilbert, G., Gal, D., … Currie, B. J. (2007). Necrotizing fasciitis in captive juvenile *Crocodylus porosus* caused by *Streptococcus agalactiae*: an outbreak and review of the animal and human literature. *Epidemiology and Infection*, *135*(8), 1248–1255.

Bjørkeng, E. K., Hjerde, E., Pedersen, T., Sundsfjord, A., & Hegstad, K. (2013). ICESluvan, a 94-kilobase mosaic integrative conjugative element conferring interspecies transfer of VanB-type glycopeptide resistance, a novel bacitracin resistance locus, and a toxin-antitoxin stabilization system. *Journal of Bacteriology*, *195*(23), 5381–5390.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Boonyayatra, S., Wongsathein, D., & Tharavichitkul, P. (2020). Genetic relatedness among *Streptococcus agalactiae* isolated from cattle, fish, and humans. *Foodborne Pathogens and Disease*, *17*(2), 137–143.

Bose, M., & Barber, R. D. (2006). Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biology*, *6*(3), 223–227.

Bossi, L., Fuentes, J. A., Mora, G., & Figueroa-Bossi, N. (2003). Prophage contribution to bacterial population dynamics. *Journal of Bacteriology*, *185*(21), 6467–6471.

Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau, M. E. R., Nesbø, C. L., … Doolittle, W. F. (2003). Lateral gene transfer and the origins of prokaryotic groups. *Annual Review of Genetics*, *37*(1), 283–328.

Bowater, R., Forbes-Faulkner, J., Anderson, I., Condon, K., Robinson, B., Kong, F., … Blyde, D. (2012). Natural outbreak of *Streptococcus agalactiae* (GBS) infection in wild giant Queensland grouper, *Epinephelus lanceolatus* (Bloch), and other wild fish in northern Queensland, Australia. *Journal of Fish Diseases*, *35*(3), 173–186.

Bradley, A. J., Leach, K. A., Breen, J. E., Green, L. E., & Green, M. J. (2007). Survey of the incidence and aetiology of mastitis on dairy farms in England and Wales. *Veterinary Record*, *160*(8), 253–258.

Breeding, K. M., Ragipani, B., Lee, K.-U. D., Malik, M., Randis, T. M., & Ratner, A. J. (2016). Real-time PCR-based serotyping of *Streptococcus agalactiae*. *Scientific Reports*, *6*, 38523.

Brochet, M., Couvé, E., Glaser, P., Guédon, G., & Payot, S. (2008). Integrative conjugative elements and related elements are major contributors to the genome diversity of *Streptococcus agalactiae*. *Journal of Bacteriology*, *190*(20), 6913–6917.

Brochet, M., Couvé, E., Zouine, M., Vallaeys, T., Rusniok, C., Lamy, M.-C., … Glaser, P. (2006). Genomic diversity and evolution within the species *Streptococcus agalactiae*. *Microbes and Infection*, *8*(5), 1227–1243.

Brochet, M., Da Cunha, V., Couvé, E., Rusniok, C., Trieu-Cuot, P., & Glaser, P. (2009). Atypical association of DDE transposition with conjugation specifies a new family of mobile elements. *Molecular Microbiology*, *71*(4), 948–959.

Brooks, M. R., Padilla-Vélez, L., Khan, T. A., Qureshi, A. A., Pieper, J. B., Maddox, C. W., & Alam, M. T. (2020). Prophage-mediated disruption of genetic competence in *Staphylococcus pseudintermedius*. *mSystems*, *5*(1), e00684–19.

Brueggemann, A. B., Pai, R., Crook, D. W., & Beall, B. (2007). Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. *PLoS Pathogens*, *3*(11), e168.

Brynildsrud, O., Bohlin, J., Scheffer, L., & Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology*, *17*(1), 1–9.

Budroni, S., Siena, E., Hotopp, J. C. D., Seib, K. L., Serruto, D., Nofroni, C., … Medini, D. (2011). *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proceedings of the National Academy of Sciences*, *108*(11), 4494–4499.

Burrus, V., Pavlovic, G., Decaris, B., & Guédon, G. (2002). The ICE*St1* element of *Streptococcus thermophilus* belongs to a large family of integrative and conjugative elements that exchange modules and change their specificity of integration. *Plasmid*, *48*(2), 77–97.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421.

Cambray, G., Guerout, A.-M., & Mazel, D. (2010). Integrons. *Annual Review of Genetics*, *44*, 141–166.

Campbell, A. M. (1992). Chromosomal insertion sites for phages and plasmids. *Journal of Bacteriology*, *174*(23), 7495.

Campisi, E., Rosini, R., Ji, W., Guidotti, S., Rojas-López, M., Geng, G., … Rinaudo, C. D. (2016). Genomic analysis reveals multi-drug resistance clusters in group B *Streptococcus* CC17 hypervirulent isolates causing neonatal invasive disease in southern mainland China. *Frontiers in Microbiology*, *7*, 1265.

Carattoli, A. (2009). Resistance plasmid families in *Enterobacteriaceae*. *Antimicrobial Agents and Chemotherapy*, *53*(6), 2227–2238.

Carattoli, A. (2013). Plasmids and the spread of resistance. *International Journal of Medical Microbiology*, *303*(6-7), 298–304.

Carattoli, A., Zankari, E., Garcìa-Fernandez, A., Larsen, M. V., Lund, O., Villa, L., … Hasman, H. (2014a). PlasmidFinder and pMLST: *in silico* detection and typing of plasmids. *Antimicrobial Agents and Chemotherapy*, AAC–02412.

Carattoli, A., Zankari, E., García-Fernández, A., Larsen, M. V., Lund, O., Villa, L., … Hasman, H. (2014b). *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy*, *58*(7), 3895–3903.

Carraro, N., Richard, X., Sulser, S., Delavat, F., Mazza, C., & van der Meer, J. R. (2020). An analog to digital converter controls bistable transfer competence development of a widespread bacterial integrative and conjugative element. *eLife*, *9*, e57915.

Carvalho-Castro, G. A., Silva, J. R., Paiva, L. V., Custódio, D. A., Moreira, R. O., Mian, G. F., … Costa, G. M. (2017). Molecular epidemiology of *Streptococcus agalactiae* isolated from mastitis in Brazilian dairy herds. *Brazilian Journal of Microbiology*, *48*(3), 551–559.

Caudell, M. A., Quinlan, M. B., Subbiah, M., Call, D. R., Roulette, C. J., Roulette, J. W., … Quinlan, R. J. (2017). Antimicrobial use and veterinary care among agro-pastoralists in Northern Tanzania. *PLoS One*, *12*(1), e0170328.

Centers for Disease Control and Prevention. (2005). Early-onset and late-onset neonatal group B streptococcal disease - United States, 1996-2004. *Morbidity and Mortality Weekly Report*, *54*(47), 1205.

Chain, P. S. G., Carniel, E., Larimer, F. W., Lamerdin, J., Stoutland, P. O., Regala, W. M., …

Garcia, E. (2004). Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proceedings of the National Academy of Sciences*, *101*(38), 13826–13831.

Chaiwarith, R., Jullaket, W., Bunchoo, M., Nuntachit, N., Sirisanthana, T., & Supparatpinyo, K. (2011). *Streptococcus agalactiae* in adults at Chiang Mai University Hospital: a retrospective study. *BMC Infectious Diseases*, *11*(1), 149.

Chan, W. T., & Espinosa, M. (2016). The *Streptococcus pneumoniae pezAT* toxin–antitoxin system reduces β-lactam resistance and genetic competence. *Frontiers in Microbiology*, *7*, 1322.

Chan, W. T., Yeo, C. C., Sadowy, E., & Espinosa, M. (2014). Functional validation of putative toxin-antitoxin genes from the Gram-positive pathogen *Streptococcus pneumoniae*: phd-doc is the fourth bona-fide operon. *Frontiers in Microbiology*, *5*, 677.

Chen, J., Quiles-Puchalt, N., Chiang, Y. N., Bacigalupe, R., Fillol-Salom, A., Chee, M. S. J., ... Penadés, J. R. (2018). Genome hypermobility by lateral transduction. *Science*, *362*(6411), 207–212.

Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M., & Corander, J. (2013). Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular Biology and Evolution*, *30*(5), 1224–1228.

Chideroli, R. T., Amoroso, N., Mainardi, R. M., Suphoronski, S. A., de Padua, S. B., Alfieri, A. F., ... Pereira, U. P. (2017). Emergence of a new multidrug-resistant and highly virulent serotype of *Streptococcus agalactiae* in fish farms from Brazil. *Aquaculture*, *479*, 45–51.

Christensen, S. K., Pedersen, K., Hansen, F. G., & Gerdes, K. (2003). Toxin–antitoxin loci as stress-response-elements: ChpAK/MazF and ChpBK cleave translated RNAs and are counteracted by tmRNA. *Journal of Molecular Biology*, *332*(4), 809–819.

Clausen, P. T., Aarestrup, F. M., & Lund, O. (2018). Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics*, *19*(1), 307.

Clewell, D. B., Flannagan, S. E., & Jaworski, D. D. (1995). Unconstrained bacterial promiscuity: the Tn*916*-Tn*1545* family of conjugative transposons. *Trends in Microbiology*, *3*(6), 229–236.

Cobo-Ángel, C. G., Jaramillo-Jaramillo, A. S., Lasso-Rojas, L. M., Aguilar-Marin, S. B.,

Sanchez, J., Rodriguez-Lecompte, J. C., … Zadoks, R. N. (2018). *Streptococcus agalactiae* is not always an obligate intramammary pathogen: Molecular epidemiology of GBS from milk, feces and environment in Colombian dairy herds. *PLoS One*, *13*(12), e0208990.

Cobo-Ángel, C. G., Jaramillo-Jaramillo, A. S., Palacio-Aguilera, M., Jurado-Vargas, L., Calvo-Villegas, E. A., Ospina-Loaiza, D. A., … Ceballos-Marquez, A. (2019). Potential group B *Streptococcus* interspecies transmission between cattle and people in Colombian dairy farms. *Scientific Reports*, *9*(1), 1–9.

Cohan, F. M. (2002). What are bacterial species? *Annual Reviews in Microbiology*, *56*(1), 457–487.

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., … Barrell, B. G. (2001). Massive gene decay in the leprosy bacillus. *Nature*, *409*(6823), 1007–1011.

Collin, S. M., Lamb, P., Jauneikaite, E., Le Doare, K., Creti, R., Berardi, A., … Lamagni, T. (2019). Hospital clusters of invasive group B streptococcal disease: a systematic review. *Journal of Infection*, *79*(6), 521–527.

Compain, F., Hays, C., Touak, G., Dmytruk, N., Trieu-Cuot, P., Joubrel, C., & Poyart, C. (2014). Molecular characterization of *Streptococcus agalactiae* isolates harboring small *erm*(T)-carrying plasmids. *Antimicrobial Agents and Chemotherapy*, *58*(11), 6928–6930.

Consuegra, J., Gaffé, J., Lenski, R. E., Hindré, T., Barrick, J. E., Tenaillon, O., & Schneider, D. (2021). Insertion-sequence-mediated mutations both promote and constrain evolvability during a long-term experiment with bacteria. *Nature Communications*, *12*(1), 1–12.

Corander, J., & Marttinen, P. (2006). Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology*, *15*(10), 2833–2843.

Corander, J., Marttinen, P., Sirén, J., & Tang, J. (2008). Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics*, *9*(1), 1–14.

Corander, J., Waldmann, P., & Sillanpää, M. J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics*, *163*(1), 367–374.

Cresawn, S. G., Bogel, M., Day, N., Jacobs-Sera, D., Hendrix, R. W., & Hatfull, G. F. (2011). Phamerator: a bioinformatic tool for comparative bacteriophage genomics. *BMC Bioinformatics*, *12*(1), 395.

Crestani, C., Forde, T. L., Lycett, S. J., Holmes, M. A., Fasth, C., Persson-Waller, K., & Zadoks, R. N. (2021). The fall and rise of group B *Streptococcus* in dairy cattle: reintroduction due to human-to-cattle host jumps? *Microbial Genomics*, *7*(9), 000648.

Crestani, C., Forde, T. L., & Zadoks, R. N. (2020). Development and application of a prophage integrase typing scheme for Group B *Streptococcus*. *Frontiers in Microbiology*, *11*, 1993.

Curcio, M. J., & Derbyshire, K. M. (2003). The outs and ins of transposition: from mu to kangaroo. *Nature Reviews Molecular Cell Biology*, *4*(11), 865.

Da Cunha, V., Davies, M. R., Douarre, P.-E., Rosinski-Chupin, I., Margarit, I., Spinali, S., … Ma, L. (2014). *Streptococcus agalactiae* clones infecting humans were selected and fixed through the extensive use of tetracycline. *Nature Communications*, *5*, ncomms5544.

Dahlberg, C., & Chao, L. (2003). Amelioration of the cost of conjugative plasmid carriage in *Eschericha coli* K12. *Genetics*, *165*(4), 1641–1649.

Darling, C. L. (1975). Standardization and evaluation of the CAMP reaction for the prompt, presumptive identification of *Streptococcus agalactiae* (Lancefield group B) in clinical material. *Journal of Clinical Microbiology*, *1*(2), 171–174.

Datta, N., & Hedges, R. (1971). Compatibility groups among *fi-* R factors. *Nature*, *234*(5326), 222.

Davies, M. R., Shera, J., Van Domselaar, G. H., Sriprakash, K. S., & McMillan, D. J. (2009). A novel integrative conjugative element mediates genetic transfer from group G *Streptococcus* to other β-hemolytic streptococci. *Journal of Bacteriology*, *191*(7), 2257–2265.

Davies, M. R., Tran, T. N., McMillan, D. J., Gardiner, D. L., Currie, B. J., & Sriprakash, K. S. (2005). Inter-species genetic movement may blur the epidemiology of streptococcal diseases in endemic regions. *Microbes and Infection*, *7*(9-10), 1128–1138.

de Aguiar, E. L., Mariano, D. C. B., Viana, M. V. C., de Jesus Benevides, L., de Souza Rocha, F., de Castro Oliveira, L., … Azevedo, V. (2016). Complete genome sequence of

Streptococcus agalactiae strain GBS85147 serotype of type Ia isolated from human oropharynx. *Standards in Genomic Sciences*, *11*(1), 39.

De Gieter, S., Konijnenberg, A., Talavera, A., Butterer, A., Haesaerts, S., De Greve, H., … Garcia-Pino, A. (2014). The intrinsically disordered domain of the antitoxin Phd chaperones the toxin Doc against irreversible inactivation and misfolding. *Journal of Biological Chemistry*, *289*(49), 34013–34023.

Delannoy, C. M., Crumlish, M., Fontaine, M. C., Pollock, J., Foster, G., Dagleish, M. P., … Zadoks, R. N. (2013). Human *Streptococcus agalactiae* strains in aquatic mammals and fish. *BMC Microbiology*, *13*(1), 1–9.

Delannoy, C. M., Zadoks, R. N., Crumlish, M., Rodgers, D., Lainson, F. A., Ferguson, H., … Fontaine, M. C. (2016). Genomic comparison of virulent and non-virulent *Streptococcus agalactiae* in fish. *Journal of Fish Diseases*, *39*(1), 13–29.

DeNap, J. C., & Hergenrother, P. J. (2005). Bacterial death comes full circle: targeting plasmid replication in drug-resistant bacteria. *Organic and Biomolecular Chemistry*, *3*(6), 959–966.

Desjardins, C. A., Cohen, K. A., Munsamy, V., Abeel, T., Maharaj, K., Walker, B. J., … Pym, A. S. (2016). Genomic and functional analyses of *Mycobacterium tuberculosis* strains implicate ald in D-cycloserine resistance. *Nature Genetics*, *48*(5), 544–551.

de Sousa, A. L., Negrão, D. M., Lobato, A. R. F., de Los Santos, F., Franklin, E., Pinheiro, K. d. C., … Ramos, R. T. J. (2018). Phageweb-web interface for rapid identification and characterization of prophages in bacterial genomes. *Frontiers in Genetics*, *9*, 644.

de Sousa, J. A. M., Buffet, A., Haudiquet, M., Rocha, E. P. C., & Rendueles, O. (2020). Modular prophage interactions driven by capsule serotype select for capsule loss under phage predation. *The ISME Journal*, *14*(12), 2980–2996.

De Vos, P., Garrity, G., Jones, D., Krieg, N., Ludwig, W., Rainey, F., … Whitman, W. (2009). Volume three: *The Firmicutes*. In *Bergey's Manual of Systematic Bacteriology* (pp. 4597–4602). Springer.

Diedrick, M. J., Flores, A. E., Hillier, S. L., Creti, R., & Ferrieri, P. (2010). Clonal analysis of colonizing group B *Streptococcus*, serotype IV, an emerging pathogen in the United States. *Journal of Clinical Microbiology*, *48*(9), 3100–3104.

Dini, M., Shokoohizadeh, L., Jalilian, F. A., Moradi, A., & Arabestani, M. R. (2019). Geno-

typing and characterization of prophage patterns in clinical isolates of *Staphylococcus aureus*. *BMC Research Notes*, *12*(1), 1–6.

DiPersio, L. P., DiPersio, J. R., Beach, J. A., Loudon, A. M., & Fuchs, A. M. (2011). Identification and characterization of plasmid-borne *erm*(T) macrolide resistance in group B and group A *Streptococcus*. *Diagnostic Microbiology and Infectious Disease*, *71*(3), 217–223.

Diz, A. P., Carvajal-Rodríguez, A., & Skibinski, D. O. (2011). Multiple hypothesis testing in proteomics: a strategy for experimental work. *Molecular and Cellular Proteomics*, *10*(3), M110–004374.

Dogan, B., Schukken, Y., Santisteban, C., & Boor, K. J. (2005). Distribution of serotypes and antimicrobial resistance genes among *Streptococcus agalactiae* isolates from bovine and human hosts. *Journal of Clinical Microbiology*, *43*(12), 5899–5906.

Domelier, A. S., van der Mee-Marquet, N., Grandet, A., Mereghetti, L., Rosenau, A., & Quentin, R. (2006). Loss of catabolic function in *Streptococcus agalactiae* strains and its association with neonatal meningitis. *Journal of Clinical Microbiology*, *44*(9), 3245–3250.

Domelier, A. S., van der Mee-Marquet, N., Sizaret, P.-Y., Héry-Arnaud, G., Lartigue, M.-F., Mereghetti, L., & Quentin, R. (2009). Molecular characterization and lytic activities of *Streptococcus agalactiae* bacteriophages and determination of lysogenic-strain features. *Journal of Bacteriology*, *191*(15), 4776–4785.

Duarte, R. S., Bellei, B. C., Miranda, O. P., Brito, M. A., & Teixeira, L. M. (2005). Distribution of antimicrobial resistance and virulence-related genes among Brazilian group B streptococci recovered from bovine and human sources. *Antimicrobial Agents and Chemotherapy*, *49*(1), 97–103.

Duarte, R. S., Miranda, O. P., Bellei, B. C., Brito, M. A. V., & Teixeira, L. M. (2004). Phenotypic and molecular characteristics of *Streptococcus agalactiae* isolates recovered from milk of dairy cows in Brazil. *Journal of Clinical Microbiology*, *42*(9), 4214–4222.

Dykhuizen, D. E., & Green, L. (1991). Recombination in *Escherichia coli* and the definition of biological species. *Journal of Bacteriology*, *173*(22), 7257–7268.

Ebrahimi, A., Moatamedi, A., Lotfalian, S., & Mirshokraei, P. (2013). Biofilm formation,

hemolysin production and antimicrobial susceptibilities of *Streptococcus agalactiae* isolated from the mastitis milk of dairy cows in Shahrekord district, Iran. In *Veterinary Research Forum: an International Quarterly Journal* (Vol. 4, p. 269).

Edwards, M. S., Rench, M. A., Rinaudo, C. D., Fabbrini, M., Tuscano, G., Buffi, G., … Margarit, I. (2016). Immune response to invasive group B *Streptococcus* disease in adults. *Emerging Infectious Diseases*, *22*(11), 1877.

Elhadi, Y. A., Nyariki, D. M., & Wasonga, O. V. (2015). Role of camel milk in pastoral livelihoods in Kenya: contribution to household diet and income. *Pastoralism*, *5*(1), 1–8.

Elliott, J. A., Facklam, R. R., & Richter, C. B. (1990). Whole-cell protein patterns of nonhemolytic group B, type Ib, streptococci isolated from humans, mice, cattle, frogs, and fish. *Journal of Clinical Microbiology*, *28*(3), 628–630.

El Tigani-Asil, E. T. A., Abdelwahab, G. E., Veedu, J. T. V. P., Khalafalla, A. I., Mohamed, Z. S. A., Ishag, H. Z. A., … Al Muhairi, S. S. M. (2020). Gangrenous mastitis in dromedary camels in UAE caused by *Streptococcus agalactiae*. *BMC Veterinary Research*, *16*, 1–8.

Ershova, A. S., Rusinov, I. S., Spirin, S. A., Karyagina, A. S., & Alexeevski, A. V. (2015). Role of restriction-modification systems in prokaryotic evolution and ecology. *Biochemistry (Moscow)*, *80*(10), 1373–1386.

Evans, J. J., Pasnik, D. J., Klesius, P. H., & Al-Ablani, S. (2006). First report of *Streptococcus agalactiae* and *Lactococcus garvieae* from a wild bottlenose dolphin (*Tursiops truncatus*). *Journal of Wildlife Diseases*, *42*(3), 561–569.

Everitt, R. G., Didelot, X., Batty, E. M., Miller, R. R., Knox, K., Young, B. C., … Wilson, D. J. (2014). Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*. *Nature Communications*, *5*(1), 1–9.

Fabre, L., Zhang, J., Guigon, G., Le Hello, S., Guibert, V., Accou-Demartin, M., … Weill, F.-X. (2012). CRISPR typing and subtyping for improved laboratory surveillance of *Salmonella* infections. *PLoS One*, *7*(5), e36995.

FAOSTAT. (2018). *Trends in the livestock sector.* Paris.

Farhat, M. R., Freschi, L., Calderon, R., Ioerger, T., Snyder, M., Meehan, C. J., … Murray, M. (2019). GWAS for quantitative resistance phenotypes in *Mycobacterium tubercu-*

losis reveals resistance genes and regulatory regions. *Nature Communications*, *10*(1), 1–11.

Farley, M. M., Harvey, C., Stull, T., Smith, J. D., Schuchat, A., Wenger, J. D., & Stephens, D. S. (1993). A population-based assessment of invasive disease due to group B *Streptococcus* in nonpregnant adults. *New England Journal of Medicine*, *328*(25), 1807–1811.

Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P., & Spratt, B. G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *Journal of Bacteriology*, *186*(5), 1518–1530.

Ferguson, H., Morales, J., & Ostland, V. (1994). Streptococcosis in aquarium fish. *Diseases of Aquatic Organisms*, *19*(1), 1–6.

Ferretti, J. J., McShan, W. M., Ajdic, D., Savic, D. J., Savic, G., Lyon, K., … Lai, H. S. (2001). Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proceedings of the National Academy of Sciences*, *98*(8), 4658–4663.

Fillol-Salom, A., Martínez-Rubio, R., Abdulrahman, R. F., Chen, J., Davies, R., & Penadés, J. R. (2018). Phage-inducible chromosomal islands are ubiquitous within the bacterial universe. *The ISME Journal*, *12*(9), 2114.

Fischer, A., Liljander, A., Kaspar, H., Muriuki, C., Fuxelius, H.-H., Bongcam-Rudloff, E., … Jores, J. (2013). Camel *Streptococcus agalactiae* populations are associated with specific disease complexes and acquired the tetracycline resistance gene *tetM* via a Tn*916*-like element. *Veterinary Research*, *44*(1), 86.

Flores, A. R., Galloway-Peña, J., Sahasrabhojane, P., Saldaña, M., Yao, H., Su, X., … Shelburne, S. A. (2015). Sequence type 1 group B *Streptococcus*, an emerging cause of invasive disease in adults, evolves by small genetic changes. *Proceedings of the National Academy of Sciences*, *112*(20), 6431–6436.

Florindo, C., Damiao, V., Silvestre, I., Farinha, C., Rodrigues, F., Nogueira, F., … Santos-Sanches, I. (2014). Epidemiological surveillance of colonising group B *Streptococcus* epidemiology in the Lisbon and Tagus Valley regions, Portugal (2005 to 2012): emergence of a new epidemic type IV/clonal complex 17 clone. *Eurosurveillance*, *19*(23), 20825.

Foley, S. L., Johnson, T. J., Ricke, S. C., Nayak, R., & Danzeisen, J. (2013). *Salmonella*

pathogenicity and host adaptation in chicken-associated serovars. *Microbiology and Molecular Biology Reviews*, *77*(4), 582–607.

Forsbäck, L., Lindmark-Månsson, H., Andrén, A., Åkerstedt, M., & Svennersten-Sjaunja, K. (2009). Udder quarter milk composition at different levels of somatic cell count in cow composite milk. *Animal*, *3*(5), 710–717.

Fouts, D. E. (2006). Phage_finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Research*, *34*(20), 5839–5851.

Franken, C., Brandt, C., Bröker, G., & Spellerberg, B. (2004). ISSag1 in streptococcal strains of human and animal origin. *International Journal of Medical Microbiology*, *294*(4), 247–254.

Franken, C., Haase, G., Brandt, C., Weber-Heynemann, J., Martin, S., Lämmler, C., … Spellerberg, B. (2001). Horizontal gene transfer and host specificity of beta-haemolytic streptococci: the role of a putative composite transposon containing *scpB* and *lmb*. *Molecular Microbiology*, *41*(4), 925–935.

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, *3*(9), 722.

Galata, V., Fehlmann, T., Backes, C., & Keller, A. (2018). PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Research*, *47*(D1), D195–D202.

Garcia, J. C., Klesius, P. H., Evans, J. J., & Shoemaker, C. A. (2008). Non-infectivity of cattle *Streptococcus agalactiae* in Nile tilapia, *Oreochromis niloticus* and channel catfish, *Ictalurus punctatus*. *Aquaculture*, *281*(1-4), 151–154.

Gendrin, C., Vornhagen, J., Armistead, B., Singh, P., Whidbey, C., Merillat, S., … Iyer, L. M. (2017). A nonhemolytic group B *Streptococcus strain* exhibits hypervirulence. *The Journal of Infectious Diseases*, *217*(6), 983–987.

Geng, Y., Wang, K., Huang, X., Chen, D., Li, C., Ren, S., … Lai, W. M. (2012). *Streptococcus agalactiae*, an emerging pathogen for cultured ya-fish, *Schizothorax prenanti*, in China. *Transboundary and Emerging Diseases*, *59*(4), 369–375.

Gerdes, K., Christensen, S. K., & Løbner-Olesen, A. (2005). Prokaryotic toxin–antitoxin stress response loci. *Nature Reviews Microbiology*, *3*(5), 371–382.

Gianneechini, R., Concha, C., Rivero, R., Delucci, I., & López, J. M. (2002). Occurrence of

clinical and sub-clinical mastitis in dairy herds in the West Littoral Region in Uruguay. *Acta Veterinaria Scandinavica*, *43*(4), 1–10.

Glaser, P., Rusniok, C., Buchrieser, C., Chevalier, F., Frangeul, L., Msadek, T., ... Trieu-Cuot, P. (2002). Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. *Molecular Microbiology*, *45*(6), 1499–1513.

Gleich-Theurer, U., Aymanns, S., Haas, G., Mauerer, S., Vogt, J., & Spellerberg, B. (2009). Human serum induces streptococcal c5a peptidase expression. *Infection and Immunity*, *77*(9), 3817–3825.

Glibert, P. M., Landsberg, J. H., Evans, J. J., Al-Sarawi, M. A., Faraj, M., Al-Jarallah, M. A., ... Shoemaker, C. (2002). A fish kill of massive proportion in Kuwait Bay, Arabian Gulf, 2001: the roles of bacterial disease, harmful algae, and eutrophication. *Harmful Algae*, *1*(2), 215–231.

Godoy, D., Carvalho-Castro, G., Leal, C., Pereira, U., Leite, R., & Figueiredo, H. (2013). Genetic diversity and new genotyping scheme for fish pathogenic *Streptococcus agalactiae*. *Letters in Applied Microbiology*, *57*(6), 476–483.

Goerke, C., Pantucek, R., Holtfreter, S., Schulte, B., Zink, M., Grumann, D., ... Wolz, C. (2009). Diversity of prophages in dominant *Staphylococcus aureus* clonal lineages. *Journal of Bacteriology*, *191*(11), 3462–3468.

Gonçalves, J., Kamphuis, C., Martins, C., Barreiro, J., Tomazi, T., Gameiro, A., ... Dos Santos, M. (2018). Bovine subclinical mastitis reduces milk yield and economic return. *Livestock Science*, *210*, 25–32.

Gori, A., Harrison, O. B., Mlia, E., Nishihara, Y., Chan, J. M., Msefula, J., ... Heyderman, R. S. (2020). Pan-GWAS of *Streptococcus agalactiae* highlights lineage-specific genes associated with virulence and niche adaptation. *MBio*, *11*(3), e00728–20.

Granlund, M., Michel, F., & Norgren, M. (2001). Mutually exclusive distribution of IS*1548* and GBSi1, an active group II intron identified in human isolates of group B streptococci. *Journal of Bacteriology*, *183*(8), 2560–2569.

Guérillot, R., Da Cunha, V., Sauvage, E., Bouchier, C., & Glaser, P. (2013). Modular evolution of Tn*GBS*s, a new family of integrative and conjugative elements associating insertion sequence transposition, plasmid replication, and conjugation for their spreading. *Journal of Bacteriology*, *195*(9), 1979–1990.

Guérout, A.-M., Iqbal, N., Mine, N., Ducos-Galand, M., Van Melderen, L., & Mazel, D. (2013). Characterization of the phd-doc and ccd toxin-antitoxin cassettes from *Vibrio* superintegrons. *Journal of Bacteriology*, *195*(10), 2270–2283.

Guglielmini, J., Quintais, L., Garcillán-Barcia, M. P., De La Cruz, F., & Rocha, E. P. (2011). The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genetics*, *7*(8), e1002222.

Guinane, C. M., Ben Zakour, N. L., Tormo-Mas, M. A., Weinert, L. A., Lowder, B. V., Cartwright, R. A., … Fitzgerald, J. R. (2010). Evolutionary genomics of *Staphylococcus aureus* reveals insights into the origin and molecular basis of ruminant host adaptation. *Genome Biology and Evolution*, *2*, 454–466.

Guo, Y., Zhang, D., Fan, H., CHEN, X.-n., LI, T.-t., & LI, A.-h. (2012). Molecular epidemiology of *Streptococcus agalactiae* isolated from tilapia in Southern China. *Journal of Fisheries of China*, *3*, 012.

Gupta, S. K., Padmanabhan, B. R., Diene, S. M., Lopez-Rojas, R., Kempf, M., Landraud, L., & Rolain, J.-M. (2014). ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial Agents and Chemotherapy*, *58*(1), 212–220.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075.

Gutekunst, H., Eikmanns, B. J., & Reinscheid, D. J. (2004). The novel fibrinogen-binding protein FbsB promotes *Streptococcus agalactiae* invasion into epithelial cells. *Infection and Immunity*, *72*(6), 3495–3504.

Guttman, D. S., & Dykhuizen, D. E. (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science*, *266*(5189), 1380–1383.

Hall, J., Adams, N. H., Bartlett, L., Seale, A. C., Lamagni, T., Bianchi-Jassir, F., … Ip, M. (2017). Maternal disease with group B *Streptococcus* and serotype distribution worldwide: systematic review and meta-analyses. *Clinical Infectious Diseases*, *65*(suppl_2), S112–S124.

Hanage, W. P., Fraser, C., & Spratt, B. G. (2005). Fuzzy species among recombinogenic bacteria. *BMC Biology*, *3*(1), 1–7.

Handa, N., Nakayama, Y., Sadykov, M., & Kobayashi, I. (2001). Experimental genome evo-

lution: large-scale genome rearrangements associated with resistance to replacement of a chromosomal restriction–modification gene complex. *Molecular Microbiology*, *40*(4), 932–940.

Harris, J. R., Burton, P., Knoppers, B. M., Lindpaintner, K., Bledsoe, M., Brookes, A. J., … others (2012). Toward a roadmap in global biobanking for health. *European Journal of Human Genetics*, *20*(11), 1105–1111.

Hasman, H., Saputra, D., Sicheritz-Ponten, T., Lund, O., Svendsen, C. A., Frimodt-Møller, N., & Aarestrup, F. M. (2014). Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *Journal of Clinical Microbiology*, *52*(1), 139–146.

Hassan, A. A., Abdulmawjood, A., Yildirim, A. Ö., Fink, K., Lämmler, C., & Schlenstedt, R. (2000). Identification of streptococci isolated from various sources by determination of *cfb* gene and other camp-factor genes. *Canadian Journal of Microbiology*, *46*(10), 946–951.

Haudiquet, M., Buffet, A., Rendueles, O., & Rocha, E. P. (2021). Interplay between the cell envelope and mobile genetic elements shapes gene flow in populations of the nosocomial pathogen *Klebsiella pneumoniae*. *PLoS Biology*, *19*(7), e3001276.

Hayes, K., O'Halloran, F., & Cotter, L. (2020). A review of antibiotic resistance in Group B *Streptococcus*: the story so far. *Critical Reviews in Microbiology*, *46*(3), 253–269.

Heath, P. T. (2016). Status of vaccine research and development of vaccines for GBS. *Vaccine*, *34*(26), 2876–2879.

Heffernan, C., & Misturelli, F. (2000). The delivery of veterinary services to the poor: Findings from Kenya. *Agris FAO*.

Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E., & Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proceedings of the National Academy of Sciences*, *96*(5), 2192–2197.

Heng, N. C., Ragland, N. L., Swe, P. M., Baird, H. J., Inglis, M. A., Tagg, J. R., & Jack, R. W. (2006). Dysgalacticin: a novel, plasmid-encoded antimicrobial protein (bacteriocin) produced by *Streptococcus dysgalactiae* subsp. *equisimilis*. *Microbiology*, *152*(7), 1991–2001.

Herbert, M. A., Beveridge, C. J., McCormick, D., Aten, E., Jones, N., Snyder, L. A., &

Saunders, N. J. (2005). Genetic islands of *Streptococcus agalactiae* strains NEM316 and 2603VR and their presence in other Group B Streptococcal strains. *BMC Microbiology*, *5*(1), 31.

Herbert, M. A., Beveridge, C. J., & Saunders, N. J. (2004a). Bacterial virulence factors in neonatal sepsis: group B *Streptococcus*. *Current Opinion in Infectious Diseases*, *17*(3), 225–229.

Herbert, M. A., Beveridge, C. J., & Saunders, N. J. (2004b). Bacterial virulence factors in neonatal sepsis: group B *Streptococcus*. *Current Opinion in Infectious Diseases*, *17*(3), 225–229.

Heringstad, B., Klemetsdal, G., & Ruane, J. (2000). Selection for mastitis resistance in dairy cattle: a review with focus on the situation in the nordic countries. *Livestock Production Science*, *64*(2-3), 95–106.

Hernández, E., Figueroa, J., & Iregui, C. (2009). Streptococcosis on a red tilapia, *Oreochromis* sp., farm: a case study. *Journal of Fish Diseases*, *32*(3), 247–252.

Hernandez, L., Bottini, E., Cadona, J., Cacciato, C., Monteavaro, C., Bustamante, A., & Sanso, A. M. (2021). Multidrug resistance and molecular characterization of *Streptococcus agalactiae* isolates from dairy cattle with mastitis. *Frontiers in Cellular and Infection Microbiology*, *11*, 647324.

Héry-Arnaud, G., Bruant, G., Lanotte, P., Brun, S., Picard, B., Rosenau, A., ... Mereghetti, L. (2007). Mobile genetic elements provide evidence for a bovine origin of clonal complex 17 of *Streptococcus agalactiae*. *Applied and Environmental Microbiology*, *73*(14), 4668–4672.

Héry-Arnaud, G., Bruant, G., Lanotte, P., Brun, S., Rosenau, A., van der Mee-Marquet, N., ... Mereghetti, L. (2005). Acquisition of insertion sequences and the GBSi1 intron by *Streptococcus agalactiae* isolates correlates with the evolution of the species. *Journal of Bacteriology*, *187*(17), 6248–6252.

Hetzel, U., König, A., Yildirim, A. Ö., Lämmler, C., & Kipar, A. (2003). Septicaemia in emerald monitors (*Varanus prasinus* Schlegel 1839) caused by *Streptococcus agalactiae* acquired from mice. *Veterinary Microbiology*, *95*(4), 283–293.

Heymann, D. L. (2006). Control, elimination, eradication and re-emergence of infectious diseases: getting the message right. *Bulletin of the World Health Organization: the*

International Journal of Public Health 2006, *84*(2), 82.

High, K. P., Edwards, M. S., & Baker, C. J. (2005). Group B streptococcal infections in elderly adults. *Clinical Infectious Diseases*, *41*(6), 839–847.

Hilario, E., & Gogarten, J. P. (1993). Horizontal transfer of ATPase genes — the tree of life becomes a net of life. *Biosystems*, *31*(2-3), 111–119.

Hohwy, J., Reinholdt, J., & Kilian, M. (2001). Population dynamics of *Streptococcus mitis* in its natural habitat. *Infection and Immunity*, *69*(10), 6055–6063.

Holden, M. T., Hauser, H., Sanders, M., Ngo, T. H., Cherevach, I., Cronin, A., … Julian, P. (2009). Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen *Streptococcus suis*. *PLoS One*, *4*(7), e6072.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.

Holt, K. E., Wertheim, H., Zadoks, R. N., Baker, S., Whitehouse, C. A., Dance, D., … Thomson, N. R. (2015). Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proceedings of the National Academy of Sciences*, *112*(27), E3574–E3581.

Horodniceanu, T., Bouanchaud, D., Bieth, G., & Chabbert, Y. (1976). R plasmids in *Streptococcus agalactiae* (group B). *Antimicrobial Agents and Chemotherapy*, *10*(5), 795–801.

Hsu, J. F., Chen, C. L., Lee, C. C., Lien, R., Chu, S. M., Fu, R. H., … Chiu, C. H. (2019). Characterization of group B *Streptococcus* colonization in full-term and Late-Preterm neonates in Taiwan. *Pediatrics and Neonatology*, *60*(3), 311–317.

Huang, J., Liang, Y., Guo, D., Shang, K., Ge, L., Kashif, J., & Wang, L. (2016). Comparative genomic analysis of the ICESa2603 family ICEs and spread of *erm*(B)- and *tet*(O)-carrying transferable 89K-subtype ICEs in swine and bovine isolates in China. *Frontiers in Microbiology*, *7*, 55.

Huijps, K., Lam, T. J., & Hogeveen, H. (2008). Costs of mastitis: facts and perception. *Journal of Dairy Research*, *75*(1), 113.

Hunt, M., Mather, A. E., Sánchez-Busó, L., Page, A. J., Parkhill, J., Keane, J. A., & Harris, S. R. (2017). ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial Genomics*, *3*(10).

Husein, A., Haftu, B., Hunde, A., & Tesfaye, A. (2013). Prevalence of camel (*Camelus dromedaries*) mastitis in Jijiga Town, Ethiopia. *African Journal of Agricultural Research*, *8*(24), 3113–3120.

Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, *14*(1), 68–73.

Iandolo, J. J., Worrell, V., Groicher, K. H., Qian, Y., Tian, R., Kenton, S., ... Qi, S. (2002). Comparative analysis of the genomes of the temperate bacteriophages φ11, φ12 and φ13 of *Staphylococcus aureus* 8325. *Gene*, *289*(1), 109–118.

Iannelli, F., Santagati, M., Santoro, F., Oggioni, M. R., Stefani, S., & Pozzi, G. (2014). Nucleotide sequence of conjugative prophage Φ1207. 3 (formerly Tn*1207.3*) carrying the *mef*(A)/*msr*(D) genes for efflux resistance to macrolides in *Streptococcus pyogenes*. *Frontiers in Microbiology*, *5*, 687.

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M. B., Pope, B. J., Tomita, T., ... Holt, K. E. (2014). SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Medicine*, *6*(11), 90.

Jafar, Q., Sameer, A. Z., Salwa, A. M., Samee, A. A., Ahmed, A. M., & Al-Sharifi, F. (2008). Molecular investigation of *Streptococcus agalactiae* isolates from environmental samples and fish specimens during a massive fish kill in Kuwait Bay. *Pakistan Journal of Biological Sciences*, *11*(21), 2500–2504.

Jaillard, M., Lima, L., Tournoud, M., Mahé, P., Van Belkum, A., Lacroix, V., & Jacob, L. (2018). A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genetics*, *14*(11), e1007758.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 239.

Jamrozy, D., Coll, F., Mather, A. E., Harris, S. R., Harrison, E. M., MacGowan, A., ... Peacock, S. J. (2017). Evolution of mobile genetic element composition in an epidemic methicillin-resistant *Staphylococcus aureus*: temporal changes correlated with frequent loss and gain events. *BMC Genomics*, *18*(1), 684.

Jamrozy, D., De Goffau, M. C., Bijlsma, M. W., Van De Beek, D., Kuijpers, T. W., Parkhill,

J., ... Bentley, S. D. (2018). Temporal population structure of invasive Group B *Streptococcus* during a period of rising disease incidence shows expansion of a CC17 clone. *bioRxiv*, 447037.

Jantrakajorn, S., Maisak, H., & Wongtavatchai, J. (2014). Comprehensive investigation of streptococcosis outbreaks in cultured Nile tilapia, *Oreochromis niloticus*, and red tilapia, *Oreochromis* sp., of Thailand. *Journal of the World Aquaculture Society*, *45*(4), 392–402.

Javan, R. R., Ramos-Sevillano, E., Akter, A., Brown, J., & Brueggemann, A. B. (2019). Prophages and satellite prophages are widespread in *Streptococcus* and may play a role in pneumococcal pathogenesis. *Nature Communications*, *10*(1), 1–14.

Ji, X., Sun, Y., Liu, J., Zhu, L., Guo, X., Lang, X., & Feng, S. (2016). A novel virulence-associated protein, *vapE*, in *Streptococcus suis* serotype 2. *Molecular Medicine Reports*, *13*(3), 2871–2877.

Jiang, S. M., Cieslewicz, M. J., Kasper, D. L., & Wessels, M. R. (2005). Regulation of virulence by a two-component system in group B *Streptococcus*. *Journal of Bacteriology*, *187*(3), 1105–1113.

Jiang, X., Yang, Y., Zhou, J., Zhu, L., Gu, Y., Zhang, X., ... Fang, W. (2016). Roles of the putative type IV-like secretion system key component VirD4 and PrsA in pathogenesis of *Streptococcus suis* type 2. *Frontiers in Cellular and Infection Microbiology*, *6*, 172.

John, H., Nijkamp, J., & Veltkamp, E. (1981). Genetic organization and expression of non-conjugative plasmids. In *Molecular Biology, Pathogenicity, and Ecology of Bacterial Plasmids* (pp. 247–258). Springer.

Johnson, C. M., & Grossman, A. D. (2015). Integrative and conjugative elements (ICEs): what they do and how they work. *Annual Review of Genetics*, *49*, 577–601.

Johnson, D. R., & Ferrieri, P. (1984). Group B streptococcal Ibc protein antigen: distribution of two determinants in wild-type strains of common serotypes. *Journal of Clinical Microbiology*, *19*(4), 506–510.

Jørgensen, H., Nordstoga, A., Sviland, S., Zadoks, R., Sølverød, L., Kvitle, B., & Mørk, T. (2016). *Streptococcus agalactiae* in the environment of bovine dairy herds - rewriting the textbooks? *Veterinary Microbiology*, *184*, 64–72.

Kalimuddin, S., Chen, S. L., Lim, C. T., Koh, T. H., Tan, T. Y., Kam, M., ... Tang, W. Y.

(2017). 2015 epidemic of severe *Streptococcus agalactiae* sequence type 283 infections in Singapore associated with the consumption of raw freshwater fish: a detailed analysis of clinical, epidemiological, and bacterial sequencing data. *Clinical Infectious Diseases*, *64*(suppl_2), S145–S152.

Karimi, Z., Ahmadi, A., Najafi, A., & Ranjbar, R. (2018). Bacterial CRISPR regions: general features and their potential for epidemiological molecular typing studies. *The Open Microbiology Journal*, *12*, 59.

Katholm, J., Bennedsgaard, T., Koskinen, M., & Rattenborg, E. (2012). Quality of bulk tank milk samples from Danish dairy herds based on real-time polymerase chain reaction identification of mastitis pathogens. *Journal of Dairy Science*, *95*(10), 5702–5708.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Kawasaki, M., Delamare-Deboutteville, J., Bowater, R. O., Walker, M. J., Beatson, S., Zakour, N. L. B., & Barnes, A. C. (2018). Microevolution of *Streptococcus agalactiae* ST-261 from Australia indicates dissemination via imported tilapia and ongoing adaptation to marine hosts or environment. *Applied and Environmental Microbiology*, *84*(16), e00859–18.

Kayansamruaj, P., Pirarat, N., Katagiri, T., Hirono, I., & Rodkhum, C. (2014). Molecular characterization and virulence gene profiling of pathogenic *Streptococcus agalactiae* populations from tilapia (*Oreochromis* sp.) farms in Thailand. *Journal of Veterinary Diagnostic Investigation*, *26*(4), 488–495.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649.

Keefe, G. P. (1997). *Streptococcus agalactiae* mastitis: a review. *The Canadian Veterinary Journal*, *38*(7), 429.

Keefe, G. P., Chaffer, M., Ceballos, A., Jaramillo, M., Londoño, M., Toro, M., & Montoya, M. (2010). Prevalence of *Streptococcus agalactiae* in cooling tanks of Colanta. *VI Seminario Internacional en Competitividad en Carne y Leche, Colombia*.

Khoo, S. K., Loll, B., Chan, W. T., Shoeman, R. L., Ngoo, L., Yeo, C. C., & Meinhart, A. (2007). Molecular and structural characterization of the PezAT chromosomal toxin-antitoxin system of the human pathogen *Streptococcus pneumoniae*. *Journal of Biological Chemistry*, *282*(27), 19606–19618.

Kingsley, R. A., Kay, S., Connor, T., Barquist, L., Sait, L., Holt, K. E., ... Dougan, G. (2013). Genome and transcriptome adaptation accompanying emergence of the definitive type 2 host-restricted *Salmonella enterica* serovar Typhimurium pathovar. *MBio*, *4*(5).

Klima, C. L., Zaheer, R., Cook, S. R., Booker, C. W., Hendrick, S., Alexander, T. W., & McAllister, T. A. (2013). Pathogens of bovine respiratory disease in North American feedlots conferring multi-drug resistance via integrative conjugative elements. *Journal of Clinical Microbiology*, JCM–02485.

Kobayashi, M., Schrag, S. J., Alderson, M. R., Madhi, S. A., Baker, C. J., Sobanjo-ter Meulen, A., ... Vekemans, J. (2019). WHO consultation on group B *Streptococcus* vaccine development: report from a meeting held on 27-28 April 2016. *Vaccine*, *37*(50), 7307–7314.

Kong, F., Gowan, S., Martin, D., James, G., & Gilbert, G. L. (2002). Serotype identification of group B streptococci by PCR and sequencing. *Journal of Clinical Microbiology*, *40*(1), 216–226.

Koonin, E. V., & Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, *36*(21), 6688–6719.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453–4455.

Kunin, V., Goldovsky, L., Darzentas, N., & Ouzounis, C. A. (2005). The net of life: reconstructing the microbial phylogenetic network. *Genome Research*, *15*(7), 954–959.

Kwatra, G., Cunnington, M. C., Merrall, E., Adrian, P. V., Ip, M., Klugman, K. P., ... Madhi, S. A. (2016). Prevalence of maternal colonisation with group B *Streptococcus*: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, *16*(9), 1076–1084.

Laabei, M., Recker, M., Rudkin, J. K., Aldeljawi, M., Gulay, Z., Sloan, T. J., ... Massey,

R. C. (2014). Predicting the virulence of MRSA from its genome sequence. *Genome Research*, *24*(5), 839–849.

Lachenauer, C., Creti, R., Michel, J., & Madoff, L. (2000). Mosaicism in the alpha-like protein genes of group B streptococci. *Proceedings of the National Academy of Sciences*, *97*(17), 9630–9635.

Lambertsen, L., Ekelund, K., Skovsted, I., Liboriussen, A., & Slotved, H.-C. (2010). Characterisation of invasive group B streptococci from adults in Denmark 1999 to 2004. *European Journal of Clinical Microbiology and Infectious Diseases*, *29*(9), 1071–1077.

Lambowitz, A. M., & Zimmerly, S. (2004). Mobile group II introns. *Annual Review of Genetics*, *38*, 1–35.

Lämmler, C., Abdulmawjood, A., & Weiß, R. (1998). Properties of serological group B streptococci of dog, cat and monkey origin. *Journal of Veterinary Medicine, Series B*, *45*(1-10), 561–566.

Lamuka, P. O., Njeruh, F. M., Gitao, G. C., & Abey, K. A. (2017). Camel health management and pastoralists' knowledge and information on zoonoses and food safety risks in Isiolo County, Kenya. *Pastoralism*, *7*(1), 1–10.

Lamy, M. C., Zouine, M., Fert, J., Vergassola, M., Couve, E., Pellegrini, E., … Poyart, C. (2004). CovS/CovR of group B *Streptococcus*: a two-component global regulatory system involved in virulence. *Molecular Microbiology*, *54*(5), 1250–1268.

Lancefield, R. C. (1933). A serological differentiation of human and other groups of hemolytic streptococci. *Journal of Experimental Medicine*, *57*(4), 571–595.

Langille, M. G., & Brinkman, F. S. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics*, *25*(5), 664–665.

Langridge, G. C., Fookes, M., Connor, T. R., Feltwell, T., Feasey, N., Parsons, B. N., … Thomson, N. R. (2015). Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. *Proceedings of the National Academy of Sciences*, *112*(3), 863–868.

Larsen, M. V., Cosentino, S., Lukjancenko, O., Saputra, D., Rasmussen, S., Hasman, H., … Lund, O. (2014). Benchmarking of methods for genomic taxonomy. *Journal of Clinical Microbiology*, *52*(5), 1529–1539.

Las Heras, A., Domınguez, L., & Fernandez-Garayzabal, J. (1999). Prevalence and aetiology of subclinical mastitis in dairy ewes of the Madrid region. *Small Ruminant Research*, *32*(1), 21–29.

Lauer, P., Rinaudo, C. D., Soriani, M., Margarit, I., Maione, D., Rosini, R., … Telford, J. L. (2005). Genome analysis reveals pili in group B *Streptococcus*. *Science*, *309*(5731), 105–105.

Leal, C. A., Queiroz, G. A., Pereira, F. L., Tavares, G. C., & Figueiredo, H. C. (2019). *Streptococcus agalactiae* Sequence Type 283 in farmed fish, Brazil. *Emerging Infectious Diseases*, *25*(4), 776.

Le Doare, K., & Heath, P. T. (2013). An overview of global GBS epidemiology. *Vaccine*, *31*, D7–D12.

Le Doare, K., O'driscoll, M., Turner, K., Seedat, F., Russell, N. J., Seale, A. C., … Cutland, C. (2017). Intrapartum antibiotic chemoprophylaxis policies for the prevention of group B streptococcal disease worldwide: systematic review. *Clinical Infectious Diseases*, *65*(suppl_2), S143–S151.

Lees, J. A., Croucher, N. J., Goldblatt, D., Nosten, F., Parkhill, J., Turner, C., … Bentley, S. D. (2017). Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *eLife*, *6*, e26255.

Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N., & Corander, J. (2018). Pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, *34*(24), 4310–4312.

Lees, J. A., Harris, S. R., Tonkin-Hill, G., Gladstone, R. A., Lo, S. W., Weiser, J. N., … Croucher, N. J. (2019). Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Research*, *29*(2), 304–316.

Lees, J. A., Vehkala, M., Välimäki, N., Harris, S. R., Chewapreecha, C., Croucher, N. J., … Corander, J. (2016). Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nature Communications*, *7*(1), 1–8.

Lefébure, T., & Stanhope, M. J. (2007). Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biology*, *8*(5), R71.

Lehnherr, H., Maguin, E., Jafri, S., & Yarmolinsky, M. B. (1993). Plasmid addiction genes

of bacteriophage P1: doc, which causes cell death on curing of prophage, and phd, which prevents host death when prophage is retained. *Journal of Molecular Biology*, *233*(3), 414–428.

Lembo, A., Gurney, M. A., Burnside, K., Banerjee, A., De Los Reyes, M., Connelly, J. E., … Doran, K. S. (2010). Regulation of CovR expression in group B *Streptococcus* impacts blood-brain barrier penetration. *Molecular Microbiology*, *77*(2), 431–443.

León-Sampedro, R., Novais, C., Peixe, L., Baquero, F., & Coque, T. (2016). Diversity and evolution of the Tn*5801-tet*(M)-like integrative and conjugative elements among *Enterococcus, Streptococcus*, and *Staphylococcus*. *Antimicrobial Agents and Chemotherapy*, *60*, 1736–1746.

Letunic, I., & Bork, P. (2006). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, *23*(1), 127–128.

Li, C., Sapugahawatte, D. N., Yang, Y., Wong, K. T., Lo, N. W. S., & Ip, M. (2020). Multidrug-resistant *Streptococcus agalactiae* strains found in human and fish with high penicillin and cefotaxime non-susceptibilities. *Microorganisms*, *8*(7), 1055.

Li, J., Tai, C., Deng, Z., Zhong, W., He, Y., & Ou, H. Y. (2017). VRprofile: gene-cluster-detection-based profiling of virulence and antibiotic resistance traits encoded within genome sequences of pathogenic bacteria. *Briefings in Bioinformatics*, *19*(4), 566–574.

Li, L., Wang, R., Huang, Y., Huang, T., Luo, F., Huang, W., … Gan, X. (2018). High incidence of pathogenic *Streptococcus agalactiae* ST485 strain in pregnant/puerperal women and isolation of hyper-virulent human CC67 strain. *Frontiers in Microbiology*, *9*, 50.

Li, L. P., Wang, R., Liang, W. W., Huang, T., Huang, Y., Luo, F. G., … Gan, X. (2015). Development of live attenuated *Streptococcus agalactiae* vaccine for tilapia via continuous passage *in vitro*. *Fish & Shellfish Immunology*, *45*(2), 955–963.

Lima-Mendez, G., Van Helden, J., Toussaint, A., & Leplae, R. (2008). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*, *24*(6), 863–865.

Lin, S. M., Zhi, Y., Ahn, K. B., Lim, S., & Seo, H. S. (2018). Status of group B streptococcal vaccine development. *Clinical and Experimental Vaccine Research*, *7*(1), 76–81.

Lindsay, J. A. (2010). Genomic variation and evolution of *Staphylococcus aureus*. *International Journal of Medical Microbiology*, *300*(2-3), 98–103.

Lindsay, J. A., Ruzin, A., Ross, H. F., Kurepina, N., & Novick, R. P. (1998). The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in *Staphylococcus aureus*. *Molecular Microbiology*, *29*(2), 527–543.

Liu, F., Barrangou, R., Gerner-Smidt, P., Ribot, E. M., Knabel, S. J., & Dudley, E. G. (2011). Novel virulence gene and clustered regularly interspaced short palindromic repeat (CRISPR) multilocus sequence typing scheme for subtyping of the major serovars of *Salmonella enterica* subsp. *enterica*. *Applied and Environmental Microbiology*, *77*(6), 1946–1956.

Liu, G., Yin, J., Barkema, H. W., Chen, L., Shahid, M., Szenci, O., … Han, B. (2017). Development of a single-dose recombinant camp factor entrapping poly (lactide-co-glycolide) microspheres-based vaccine against *Streptococcus agalactiae*. *Vaccine*, *35*(9), 1246–1253.

Liu, L., Li, Y., He, R., Xiao, X., Zhang, X., Su, Y., … Li, A. (2014). Outbreak of *Streptococcus agalactiae* infection in barcoo grunter, *Scortum barcoo* (McCulloch and Waite), in an intensive fish farm in China. *Journal of Fish Diseases*, *37*(12), 1067–1072.

Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., … Ou, H.-Y. (2018). ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Research*, *47*(D1), D660–D665.

Loessner, M. J., Inman, R. B., Lauer, P., & Calendar, R. (2000). Complete nucleotide sequence, molecular analysis and genome structure of bacteriophage A118 of *Listeria monocytogenes*: implications for phage evolution. *Molecular Microbiology*, *35*(2), 324–340.

Lopez-Sanchez, M. J., Sauvage, E., Da Cunha, V., Clermont, D., Ratsima Hariniaina, E., Gonzalez-Zorn, B., … Glaser, P. (2012). The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. *Molecular Microbiology*, *85*(6), 1057–1071.

Lowder, B. V., Guinane, C. M., Zakour, N. L. B., Weinert, L. A., Conway-Morris, A., Cartwright, R. A., … Fitzgerald, J. R. (2009). Recent human-to-poultry host jump, adaptation, and pandemic spread of *Staphylococcus aureus*. *Proceedings of the Na-*

tional Academy of Sciences, *106*(46), 19545–19550.

Lu, M. (2010). Review of research on streptococcosis in tilapia. *South China Fisheries Science*, *6*(1), 75–79.

Luan, S. L., Granlund, M., Sellin, M., Lagergård, T., Spratt, B. G., & Norgren, M. (2005). Multilocus sequence typing of Swedish invasive group B *Streptococcus* isolates indicates a neonatally associated genetic lineage and capsule switching. *Journal of Clinical Microbiology*, *43*(8), 3727–3733.

Lusiastuti, A. M., Seeger, H., Indrawati, A., & Zschöck, M. (2013). The comparison of *Streptococcus agalactiae* isolated from fish and bovine using multilocus sequence typing. *Hayati Journal of Biosciences*, *20*(4), 157–162.

Lyhs, U., Kulkas, L., Katholm, J., Waller, K. P., Saha, K., Tomusk, R. J., & Zadoks, R. N. (2016). *Streptococcus agalactiae* serotype IV in humans and cattle, northern Europe. *Emerging Infectious Diseases*, *22*(12), 2097.

Magaš, V., Vakanjac, S., Pavlović, V., Velebit, B., Mirilović, M., Maletić, M., ... Nedić, S. (2013). Efficiency evaluation of a bivalent vaccine in the prophylaxis of mastitis in cows. *Acta Veterinaria-Beograd*, *63*(5-6), 525–536.

Maiden, M. C. J. (2006). Multilocus sequence typing of bacteria. *Annual Review Microbiology*, *60*, 561–588.

Maisey, H. C., Doran, K. S., & Nizet, V. (2008). Recent advances in understanding the molecular basis of group B *Streptococcus* virulence. *Expert Reviews in Molecular Medicine*, *10*.

Manning, S. D., Springman, A. C., Lehotzky, E., Lewis, M. A., Whittam, T. S., & Davies, H. D. (2009). Multilocus sequence types associated with neonatal group B streptococcal sepsis and meningitis in Canada. *Journal of Clinical Microbiology*, *47*(4), 1143–1148.

Manning, S. D., Springman, A. C., Million, A. D., Milton, N. R., McNamara, S. E., Somsel, P. A., ... Davies, H. D. (2010). Association of group B *Streptococcus* colonization and bovine exposure: a prospective cross-sectional cohort study. *PLoS One*, *5*(1), e8795.

Martinez, G., Harel, J., Higgins, R., Lacouture, S., Daignault, D., & Gottschalk, M. (2000). Characterization of *Streptococcus agalactiae* isolates of bovine and human origin by

randomly amplified polymorphic DNA analysis. *Journal of Clinical Microbiology*, *38*(1), 71–78.

Martínez-Rubio, R., Quiles-Puchalt, N., Martí, M., Humphrey, S., Ram, G., Smyth, D., … Penadés, J. R. (2017). Phage-inducible islands in the Gram-positive cocci. *The ISME Journal*, *11*(4), 1029.

Martins, E. R., Melo-Cristino, J., & Ramirez, M. (2010). Evidence for rare capsular switching in *Streptococcus agalactiae*. *Journal of Bacteriology*, *192*(5), 1361–1369.

Mata, C., Navarro, F., Miró, E., Walsh, T. R., Mirelis, B., & Toleman, M. (2011). Prevalence of SXT/R391-like integrative and conjugative elements carrying bla_{CMY-2} in *Proteus mirabilis*. *Journal of Antimicrobial Chemotherapy*, *66*(10), 2266–2270.

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, *14*(9), 1–9.

McNally, A., Oren, Y., Kelly, D., Pascoe, B., Dunn, S., Sreecharan, T., … Corander, J. (2016). Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genetics*, *12*(9), e1006280.

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics and Development*, *15*(6), 589–594.

Melin, P. (2011). Neonatal group B streptococcal disease: from pathogenesis to preventive strategies. *Clinical Microbiology and Infection*, *17*(9), 1294–1303.

Meltz Steinberg, K., & Levin, B. R. (2007). Grazing protozoa and the evolution of the *Escherichia coli* O157: H7 Shiga toxin-encoding prophage. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1621), 1921–1929.

Metcalf, B., Chochua, S., Gertz Jr, R., Hawkins, P., Ricaldi, J., Li, Z., … Beall, B. (2017). Short-read whole genome sequencing for determination of antimicrobial resistance mechanisms and capsular serotypes of current invasive *Streptococcus agalactiae* recovered in the USA. *Clinical Microbiology and Infection*, *23*(8), 574–e7.

Michael, G. B., Kadlec, K., Sweeney, M. T., Brzuszkiewicz, E., Liesegang, H., Daniel, R., … Schwarz, S. (2011). ICE*Pmu1*, an integrative conjugative element (ICE) of *Pasteurella multocida*: analysis of the regions that comprise 12 antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, *67*(1), 84–90.

Mikheyev, A. S., & Tin, M. M. Y. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources*, *14*(6), 1097–1102.

Morach, M., Stephan, R., Schmitt, S., Ewers, C., Zschöck, M., Reyes-Velez, J., ... Daubenberger, C. A. (2018). Population structure and virulence gene profiles of *Streptococcus agalactiae* collected from different hosts worldwide. *European Journal of Clinical Microbiology and Infectious Diseases*, *37*(3), 527–536.

Mortensen, B. L., & Skaar, E. P. (2013). The contribution of nutrient metal acquisition and metabolism to *Acinetobacter baumannii* survival within the host. *Frontiers in Cellular and Infection Microbiology*, *3*, 95.

Mukhopadhyay, S., & Puopolo, K. M. (2019). Preventing neonatal group B *Streptococcus* disease: the limits of success. *JAMA Pediatrics*, *173*(3), 219–220.

Munang'andu, H. M., Paul, J., & Evensen, Ø. (2016). An overview of vaccination strategies and antigen delivery systems for *Streptococcus agalactiae* vaccines in Nile tilapia (*Oreochromis niloticus*). *Vaccines*, *4*(4), 48.

Munch-Petersen, E., Christie, R., Simmons, R., & Beddome, H. (1945). Further notes on a lytic phenomenon shown by group B streptococci. *Australian Journal of Experimental Biology and Medical Science*, *23*(3), 193–195.

Murray, S., Pascoe, B., Meric, G., Mageiros, L., Yahara, K., Hitchings, M. D., ... Sheppard, S. K. (2017). Recombination-mediated host adaptation by avian *Staphylococcus aureus*. *Genome Biology and Evolution*, *9*(4), 830–842.

Mutua, J., Gitao, C., Bebora, L., & Mutua, F. (2017). Antimicrobial resistance profiles of bacteria isolated from the nasal cavity of camels in Samburu, Nakuru, and Isiolo Counties of Kenya. *Journal of Veterinary Medicine*, *2017*.

Mweu, M. M., Nielsen, S. S., Halasa, T., & Toft, N. (2012). Annual incidence, prevalence and transmission characteristics of *Streptococcus agalactiae* in Danish dairy herds. *Preventive Veterinary Medicine*, *106*(3-4), 244–250.

Mweu, M. M., Nielsen, S. S., Halasa, T., & Toft, N. (2014). Spatiotemporal patterns, annual baseline and movement-related incidence of *Streptococcus agalactiae* infection in Danish dairy herds: 2000-2009. *Preventive Veterinary Medicine*, *113*(2), 219–230.

Neave, F., Dodd, F., Kingwill, R., & Westgarth, D. (1969). Control of mastitis in the dairy herd by hygiene and management. *Journal of Dairy Science*, *52*(5), 696–707.

Neemuchwala, A., Teatero, S., Athey, T. B., McGeer, A., & Fittipaldi, N. (2016). Capsular switching and other large-scale recombination events in invasive sequence type 1 group B *Streptococcus*. *Emerging Infectious Diseases*, *22*(11), 1941.

Nesbø, C. L., Dlutek, M., & Doolittle, W. F. (2006). Recombination in *Thermotoga*: implications for species concepts and biogeography. *Genetics*, *172*(2), 759–769.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274.

Nguyen, S. V., & McShan, W. M. (2014). Chromosomal islands of *Streptococcus pyogenes* and related streptococci: molecular switches for survival and virulence. *Frontiers in Cellular and Infection Microbiology*, *4*, 109.

Nickmans, S., Verhoye, E., Boel, A., Van Vaerenbergh, K., & De Beenhouwer, H. (2012). Possible solution to the problem of nonhemolytic group B *Streptococcus* on Granada medium. *Journal of Clinical Microbiology*, *50*(3), 1132–1133.

Nielsen, C., & Emanuelson, U. (2013). Mastitis control in Swedish dairy herds. *Journal of Dairy Science*, *96*(11), 6883–6893.

Nocard, M., & Mollereau, R. (1887). Sur une mammite contagieuse des vaches laitières. *Institut Pasteur Annual*, *1*, 109.

Novick, R. P., Christie, G. E., & Penadés, J. R. (2010). The phage-related chromosomal islands of Gram-positive bacteria. *Nature Reviews Microbiology*, *8*(8), 541.

Nowrouzian, F. L., Wold, A. E., & Adlerberth, I. (2005). *Escherichia coli* strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants. *The Journal of Infectious Diseases*, *191*(7), 1078–1083.

Obied, A., Bagadi, H., & Mukhtar, M. (1996). Mastitis in *Camelus dromedarius* and the somatic cell content of camels' milk. *Research in Veterinary Science*, *61*(1), 55–58.

Ochoa-Díaz, M. M., Daza-Giovannetty, S., & Gómez-Camargo, D. (2018). Bacterial genotyping methods: from the basics to modern. In *Host-Pathogen Interactions* (pp. 13–20). Springer.

Ohlsson, A., & Shah, V. S. (2014). Intrapartum antibiotics for known maternal group B streptococcal colonization. *The Cochrane Database of Systematic Reviews*, *6*.

Olendzenski, L., & Gogarten, J. P. (2009). Gene transfer: who benefits? In *Horizontal Gene*

Transfer (pp. 3–9). Springer.

Oliveira, I. C. M., De Mattos, M. C., Pinto, T. A., Ferreira-Carvalho, B. T., Benchetrit, L. C., Whiting, A. A., … Figueiredo, A. M. S. (2006). Genetic relatedness between group B streptococci originating from bovine mastitis and a human group B *Streptococcus* type V cluster displaying an identical pulsed-field gel electrophoresis pattern. *Clinical Microbiology and Infection*, *12*(9), 887–893.

Oliveira, P. H., Touchon, M., Cury, J., & Rocha, E. P. (2017). The chromosomal organization of horizontal gene transfer in bacteria. *Nature Communications*, *8*(1), 841.

Oliveira, P. H., Touchon, M., & Rocha, E. P. (2016). Regulation of genetic flux between bacteria by restriction–modification systems. *Proceedings of the National Academy of Sciences*, *113*(20), 5658–5663.

Ong, S. W., Barkham, T., Kyaw, W. M., Ho, H. J., & Chan, M. (2018). Characterisation of bone and joint infections due to group B *Streptococcus* serotype III sequence type 283. *European Journal of Clinical Microbiology and Infectious Diseases*, *37*(7), 1313–1317.

Oppegaard, O., Skrede, S., Mylvaganam, H., & Kittang, B. R. (2020). Emerging threat of antimicrobial resistance in β-hemolytic htreptococci. *Frontiers in Microbiology*, *11*, 797.

Østerås, O., Sølverød, L., & Reksen, O. (2006). Milk culture results in a large Norwegian survey—effects of season, parity, days in milk, resistance, and clustering. *Journal of Dairy Science*, *89*(3), 1010–1023.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., … Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693.

Page, R., & Peti, W. (2016). Toxin-antitoxin systems in bacterial growth arrest and persistence. *Nature Chemical Biology*, *12*(4), 208–214.

Pang, M., Sun, L., He, T., Bao, H., Zhang, L., Zhou, Y., … Wang, R. (2017). Molecular and virulence characterization of highly prevalent *Streptococcus agalactiae* circulated in bovine dairy herds. *Veterinary Research*, *48*(1), 65.

Parkhill, J., Dougan, G., James, K., Thomson, N., Pickard, D., Wain, J., … Barrell, B. (2001). Complete genome sequence of a multiple drug resistant *Salmonella enterica*

serovar Typhi CT18. *Nature*, *413*(6858), 848–852.

Patras, K. A., Wang, N.-Y., Fletcher, E. M., Cavaco, C. K., Jimenez, A., Garg, M., … Doran, K. S. (2013). Group B *Streptococcus* CovR regulation modulates host immune signalling pathways to promote vaginal colonization. *Cellular Microbiology*, *15*(7), 1154–1167.

Pavón, A. B. I., & Maiden, M. C. J. (2009). Multilocus sequence typing. In *Molecular Epidemiology of Microorganisms* (pp. 129–140). Springer.

Penadés, J. R., & Christie, G. E. (2015). The phage-inducible chromosomal islands: a family of highly evolved molecular parasites. *Annual Review of Virology*, *2*, 181–201.

Pereira, U., Mian, G., Oliveira, I., Benchetrit, L., Costa, G., & Figueiredo, H. (2010). Genotyping of *Streptococcus agalactiae* strains isolated from fish, human and cattle and their virulence potential in Nile tilapia. *Veterinary Microbiology*, *140*(1-2), 186–192.

Persson, Y., Nyman, A.-K. J., & Grönlund-Andersson, U. (2011). Etiology and antimicrobial susceptibility of udder pathogens from cases of subclinical mastitis in dairy cows in Sweden. *Acta Veterinaria Scandinavica*, *53*(1), 1–8.

Phuoc, N. N., Linh, N. T. H., Crestani, C., & Zadoks, R. N. (2021). Effect of strain and enviromental conditions on the virulence of *Streptococcus agalactiae* (Group B *Streptococcus*; GBS) in red tilapia (*Oreochromis* sp.). *Aquaculture*, *534*, 736256.

Piepers, S., De Meulemeester, L., de Kruif, A., Opsomer, G., Barkema, H. W., & De Vliegher, S. (2007). Prevalence and distribution of mastitis pathogens in subclinically infected dairy cows in Flanders, Belgium. *Journal of Dairy Research*, *74*(4), 478–483.

Pitkälä, A., Haveri, M., Pyörälä, S., Myllys, V., & Honkanen-Buzalski, T. (2004). Bovine mastitis in Finland 2001—prevalence, distribution of bacteria, and antimicrobial resistance. *Journal of Dairy Science*, *87*(8), 2433–2441.

Podbielski, A., Blankenstein, O., & Lütticken, R. (1994). Molecular characterization of the *cfb* gene encoding group B streptococcal CAMP-factor. *Medical Microbiology and Immunology*, *183*(5), 239–256.

Pradeep, P. J., Suebsing, R., Sirthammajak, S., Kampeera, J., Jitrakorn, S., Saksmerprome, V., … Jeffs, A. (2016). Evidence of vertical transmission and tissue tropism of strepto-

coccosis from naturally infected red tilapia (*Oreochromis* spp.). *Aquaculture Reports*, *3*, 58–66.

Pretto-Giordano, L. G., Müller, E. E., Klesius, P., & Da Silva, V. G. (2010). Efficacy of an experimentally inactivated *Streptococcus agalactiae* vaccine in nile tilapia (*Oreochromis niloticus*) reared in brazil. *Aquaculture Research*, *41*(10), 1539–1544.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, *5*(3).

R Core Team. (2013). R: A language and environment for statistical computing.

Rabinovich, L., Sigal, N., Borovok, I., Nir-Paz, R., & Herskovits, A. A. (2012). Prophage excision activates *Listeria* competence genes that promote phagosomal escape and virulence. *Cell*, *150*(4), 792–802.

Rahman, M. M., Rahman, M. A., Monir, M. S., Haque, M. E., Siddique, M. P., Khasruzzaman, A., … Islam, M. A. (2021). Isolation and molecular detection of *Streptococcus agalactiae* from popped eye disease of cultured Tilapia and Vietnamese koi fishes in Bangladesh. *Journal of Advanced Veterinary and Animal Research*, *8*(1), 14.

Rajendram, P., Kyaw, W. M., Leo, Y. S., Ho, H., Chen, W. K., Lin, R., … Chow, A. (2016). Group B *Streptococcus sequence* type 283 disease linked to consumption of raw fish, Singapore. *Emerging Infectious Diseases*, *22*(11), 1974.

Rambaut, A., Lam, T. T., Max Carvalho, L., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, *2*(1), vew007.

Rato, M. G., Bexiga, R., Florindo, C., Cavaco, L. M., Vilela, C. L., & Santos-Sanches, I. (2013). Antimicrobial resistance and molecular epidemiology of streptococci from bovine mastitis. *Veterinary Microbiology*, *161*(3-4), 286–294.

Renard, A., Barbera, L., Courtier-Martinez, L., Dos Santos, S., Valentin, A.-S., Mereghetti, L., … van der Mee-Marquet, N. L. (2019). phiD12-like livestock-associated prophages are associated with novel subpopulations of *Streptococcus agalactiae* infecting neonates. *Frontiers in Cellular and Infection Microbiology*, *9*, 166.

Richards, V. P., Choi, S. C., Bitar, P. D. P., Gurjar, A. A., & Stanhope, M. J. (2013). Transcriptomic and genomic evidence for *Streptococcus agalactiae* adaptation to the bovine environment. *BMC Genomics*, *14*(1), 920.

Richards, V. P., Lang, P., Bitar, P. D. P., Lefébure, T., Schukken, Y. H., Zadoks, R. N., & 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*, *11*(6), 1263–1275.

Richards, V. P., Velsko, I. M., Alam, T., Zadoks, R. N., Manning, S. D., Pavinski Bitar, P. D., ... Stanhope, M. J. (2019). Population gene introgression and high genome plasticity for the zoonotic pathogen *Streptococcus agalactiae*. *Molecular Biology and Evolution*, *36*(11), 2572–2590.

Richardson, E. J., Bacigalupe, R., Harrison, E. M., Weinert, L. A., Lycett, S., Vrieling, M., ... Fitzgerald, J. R. (2018). Gene exchange drives the ecological success of a multi-host bacterial pathogen. *Nature Ecology and Evolution*, *2*(9), 1468.

Riekerink, R. G. O., Barkema, H. W., Scholl, D. T., Poole, D. E., & Kelton, D. F. (2010). Management practices associated with the bulk-milk prevalence of *Staphylococcus aureus* in Canadian dairy farms. *Preventive Veterinary Medicine*, *97*(1), 20–28.

Roberts, A. P., & Mullany, P. (2009). A modular master on the move: the Tn*916* family of mobile genetic elements. *Trends in Microbiology*, *17*(6), 251–258.

Roberts, R. J., Vincze, T., Posfai, J., & Macelis, D. (2005). REBASE—restriction enzymes and DNA methyltransferases. *Nucleic Acids Research*, *33*(suppl_1), D230–D232.

Rocha, E. P. C. (2004). Order and disorder in bacterial genomes. *Current Opinion in Microbiology*, *7*(5), 519–527.

Rodic, A., Blagojevic, B., Zdobnov, E., Djordjevic, M., & Djordjevic, M. (2017). Understanding key features of bacterial restriction-modification systems through quantitative modeling. *BMC Systems Biology*, *11*(1), 1–15.

Rodriguez-Granger, J., Spellerberg, B., Asam, D., & Rosa-Fraile, M. (2015). Non-haemolytic and non-pigmented group B *Streptococcus*, an infrequent cause of early onset neonatal sepsis. *FEMS Pathogens and Disease*, *73*(9), ftv089.

Rolland, K., Marois, C., Siquier, V., Cattier, B., & Quentin, R. (1999). Genetic features of *Streptococcus agalactiae* strains causing severe neonatal infections, as revealed by pulsed-field gel electrophoresis and *hylB* gene analysis. *Journal of Clinical Microbiology*, *37*(6), 1892–1898.

Rondini, S., Käser, M., Stinear, T., Tessier, M., Mangold, C., Dernick, G., ... Pluschke, G.

(2007). Ongoing genome reduction in *Mycobacterium ulcerans*. *Emerging Infectious Diseases*, *13*(7), 1008.

Rosa-Fraile, M., Dramsi, S., & Spellerberg, B. (2014). Group B streptococcal haemolysin and pigment, a tale of twins. *FEMS Microbiology Reviews*, *38*(5), 932–946.

Rosini, R., Rinaudo, C. D., Soriani, M., Lauer, P., Mora, M., Maione, D., … Buccato, S. (2006). Identification of novel genomic islands coding for antigenic pilus-like structures in *Streptococcus agalactiae*. *Molecular Microbiology*, *61*(1), 126–141.

Rosinski-Chupin, I., Sauvage, E., Mairey, B., Mangenot, S., Ma, L., Da Cunha, V., … Glaser, P. (2013). Reductive evolution in *Streptococcus agalactiae* and the emergence of a host adapted lineage. *BMC Genomics*, *14*(1), 252.

Rouli, L., Merhej, V., Fournier, P.-E., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, *7*, 72–85.

Russell, N. J., Seale, A. C., O'Driscoll, M., O'Sullivan, C., Bianchi-Jassir, F., Gonzalez-Guarin, J., … Ip, M. (2017). Maternal colonization with group B *Streptococcus* and serotype distribution worldwide: systematic review and meta-analyses. *Clinical Infectious Diseases*, *65*(suppl_2), S100–S111.

Saalfeld, W., & Edwards, G. (2010). Distribution and abundance of the feral camel (*Camelus dromedarius*) in Australia. *The Rangeland Journal*, *32*(1), 1–9.

Saleh, S. K., Al-Ramadhan, G., & Faye, B. (2013). Monitoring of monthly SCC in she-camel in relation to milking practice, udder status and microbiological contamination of milk. *Emirates Journal of Food and Agriculture*, 403–408.

Salloum, M., van der Mee-Marquet, N., Domelier, A. S., Arnault, L., & Quentin, R. (2010). Molecular characterization and prophage DNA contents of *Streptococcus agalactiae* strains isolated from adult skin and osteoarticular infections. *Journal of Clinical Microbiology*, *48*(4), 1261–1269.

Salloum, M., van der Mee-marquet, N., Valentin-Domelier, A.-S., & Quentin, R. (2011). Diversity of prophage DNA regions of *Streptococcus agalactiae* clonal lineages from adults and neonates with invasive infectious disease. *PLoS One*, *6*(5), e20256.

Sampimon, O., Barkema, H. W., Berends, I., Sol, J., & Lam, T. (2009). Prevalence of intramammary infection in Dutch dairy herds. *Journal of Dairy Research*, *76*(2), 129–

136.

San, J. E., Baichoo, S., Kanzi, A., Moosa, Y., Lessells, R., Fonseca, V., … de Oliveira, T. (2020). Current affairs of microbial genome-wide association studies: approaches, bottlenecks and analytical pitfalls. *Frontiers in Microbiology*, *10*, 3119.

Sánchez-Busó, L., Golparian, D., Parkhill, J., Unemo, M., & Harris, S. R. (2019). Genetic variation regulates the activation and specificity of Restriction-Modification systems in *Neisseria gonorrhoeae*. *Scientific Reports*, *9*(1), 1–11.

Santagati, M., Iannelli, F., Cascone, C., Campanile, F., Oggioni, M. R., Stefani, S., & Pozzi, G. (2003). The novel conjugative transposon Tn*1207.3* carries the macrolide efflux gene *mef*(A) in *Streptococcus pyogenes*. *Microbial Drug Resistance*, *9*(3), 243–247.

Saunders, N. J., & Snyder, L. A. S. (2002). The minimal mobile element. *Microbiology*, *148*(12), 3756–3760.

Schulein, R., Guye, P., Rhomberg, T. A., Schmid, M. C., Schröder, G., Vergunst, A. C., … Dehio, C. (2005). A bipartite signal mediates the transfer of type IV secretion substrates of *Bartonella henselae* into human cells. *Proceedings of the National Academy of Sciences*, *102*(3), 856–861.

Seale, A. C., Bianchi-Jassir, F., Russell, N. J., Kohli-Lynch, M., Tann, C. J., Hall, J., … Bartlett, L. (2017). Estimates of the burden of group B streptococcal disease worldwide for pregnant women, stillbirths, and children. *Clinical Infectious Diseases*, *65*, S200–S219.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069.

Seligsohn, D., Crestani, C., Gitahi, N., Flodin, E. L., Chenais, E., & Zadoks, R. N. (2021a). Genomic analysis of Group B *Streptococcus* from milk demonstrates the need for improved biosecurity: a cross-sectional study of pastoralist camels in Kenya. *BMC Microbiology*, *21*, 217.

Seligsohn, D., Crestani, C., Gitahi, N., Flodin, E. L., Chenais, E., & Zadoks, R. N. (2021b). Investigation of extramammary sources of Group B *Streptococcus* reveals its unusual ecology and epidemiology in camels. *bioRxiv*, 445946.

Seligsohn, D., Nyman, A., Younan, M., Sake, W., Persson, Y., Bornstein, S., … Chenais, E. (2020). Subclinical mastitis in pastoralist dairy camel herds in Isiolo, Kenya:

Prevalence, risk factors, and antimicrobial susceptibility. *Journal of Dairy Science*, *103*(5), 4717–4731.

Sena, D. S., Mal, G., Kumar, R., & Sahani, M. (2001). A preliminary study of prevalence of mastitis in camel. *Journal of Applied Animal Research*, *20*(1), 27–31.

Sendi, P., Furitsch, M., Mauerer, S., Florindo, C., Kahl, B. C., Shabayek, S., … Spellerberg, B. (2016). Chromosomally and extrachromosomally mediated high-level gentamicin resistance in *Streptococcus agalactiae*. *Antimicrobial Agents and Chemotherapy*, AAC–01933.

Shane, A. L., Sánchez, P. J., & Stoll, B. J. (2017). Neonatal sepsis. *The Lancet*, *390*(10104), 1770–1780.

Shaw, D. (2021). The association between satellite prophages and pneumococcal carriage and disease. *Abstract for the 31st European Congress of Clinical Microbiology and Infectious Diseases - online event*.

Shepheard, M. A., Fleming, V. M., Connor, T. R., Corander, J., Feil, E. J., Fraser, C., & Hanage, W. P. (2013). Historical zoonoses and other changes in host tropism of *Staphylococcus aureus*, identified by phylogenetic analysis of a population dataset. *PLoS One*, *8*(5), e62369.

Sheppard, A. E., Vaughan, A., Jones, N., Turner, P., Turner, C., Efstratiou, A., … Seale, A. C. (2016). Capsular typing method for *Streptococcus agalactiae* using whole-genome sequence data. *Journal of Clinical Microbiology*, *54*(5), 1388–1390.

Sheppard, S. K., Cheng, L., Méric, G., De Haan, C. P., Llarena, A.-K., Marttinen, P., … Jukka, C. (2014). Cryptic ecology among host generalist *Campylobacter jejuni* in domestic animals. *Molecular Ecology*, *23*(10), 2442–2451.

Sheppard, S. K., Didelot, X., Meric, G., Torralbo, A., Jolley, K. A., Kelly, D. J., … Falush, D. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proceedings of the National Academy of Sciences*, *110*(29), 11923–11927.

Sheppard, S. K., Guttman, D. S., & Fitzgerald, J. R. (2018). Population genomics of bacterial host adaptation. *Nature Reviews Genetics*, *19*(9), 549–565.

Siefert, J. L. (2009). Defining the mobilome. In *Horizontal Gene Transfer* (pp. 13–27). Springer.

Siguier, P., Gourbeyre, E., & Chandler, M. (2014). Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews*, *38*(5), 865–891.

Siguier, P., Pérochon, J., Lestrade, L., Mahillon, J., & Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research*, *34*(suppl_1), D32–D36.

Six, A., Firon, A., Plainvert, C., Caplain, C., Touak, G., Dmytruk, N., ... Poyart, C. (2015). Molecular characterization of non-haemolytic and non-pigmented group B streptococci responsible for human invasive infections. *Journal of Clinical Microbiology*, JCM–02177.

Six, A., Krajangwong, S., Crumlish, M., Zadoks, R. N., & Walker, D. (2019). *Galleria mellonella* as an infection model for the multi-host pathogen *Streptococcus agalactiae* reflects hypervirulence of strains associated with human invasive disease. *Virulence*, *10*(1), 600–609.

Skoff, T. H., Farley, M. M., Petit, S., Craig, A. S., Schaffner, W., Gershman, K., ... Albanese, B. A. (2009). Increasing burden of invasive group B streptococcal disease in nonpregnant adults, 1990-2007. *Clinical Infectious Diseases*, *49*(1), 85–92.

Slotved, H. C., Elliott, J., Thompson, T., & Konradsen, H. B. (2003). Latex assay for serotyping of group B *Streptococcus* isolates. *Journal of Clinical Microbiology*, *41*(9), 4445–4447.

Smeds, L., & Künstner, A. (2011). ConDeTri-a content dependent read trimmer for Illumina data. *PLoS One*, *6*(10), e26314.

Smith, Y. (Ed.). (1996). *Proceedings of the First National Conference on Porous Sieves: 27-30 June 1996; Baltimore.* Stoneham: Butterworth-Heinemann.

Snyder, L. A., McGowan, S., Rogers, M., Duro, E., O'farrell, E., & Saunders, N. J. (2007). The repertoire of minimal mobile elements in the *Neisseria* species and evidence that these are involved in horizontal gene transfer in other bacteria. *Molecular Biology and Evolution*, *24*(12), 2802–2815.

Solomon, J. M., & Grossman, A. D. (1996). Who's competent and when: regulation of natural genetic competence in bacteria. *Trends in Genetics*, *12*(4), 150–155.

Sørensen, U. B. S., Klaas, I. C., Boes, J., & Farre, M. (2019). The distribution of clones of *Streptococcus agalactiae* (group B streptococci) among herdspersons and dairy cows

demonstrates lack of host specificity for some lineages. *Veterinary Microbiology*, *235*, 71–79.

Sørensen, U. B. S., Poulsen, K., Ghezzo, C., Margarit, I., & Kilian, M. (2010). Emergence and global dissemination of host-specific *Streptococcus agalactiae* clones. *MBio*, *1*(3), e00178–10.

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, *16*(8), 472.

Spellerberg, B., Martin, S., Franken, C., Berner, R., & Lütticken, R. (2000). Identification of a novel insertion sequence element in *Streptococcus agalactiae*. *Gene*, *241*(1), 51–56.

Spencer, H. (1864). *The Principles of Biology*. Springer.

Spoor, L. E., Richardson, E., Richards, A. C., Wilson, G. J., Mendonca, C., Gupta, R. K., … Fitzgerald, J. R. (2015). Recombination-mediated remodelling of host–pathogen interactions during *Staphylococcus aureus* niche adaptation. *Microbial Genomics*, *1*(4), e000036.

Straume, D., Stamsås, G. A., & Håvarstein, L. S. (2015). Natural transformation and genome evolution in *Streptococcus pneumoniae*. *Infection, Genetics and Evolution*, *33*, 371–380.

Subbiah, M., Caudell, M. A., Mair, C., Davis, M. A., Matthews, L., Quinlan, R. J., … Call, D. R. (2020). Antimicrobial resistant enteric bacteria are widely distributed amongst people, animals and the environment in Tanzania. *Nature Communications*, *11*(1), 1–12.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*(1), vey016.

Sukhnanand, S., Dogan, B., Ayodele, M. O., Zadoks, R. N., Craver, M. P. J., Dumas, N. B., … Wiedmann, M. (2005). Molecular subtyping and characterization of bovine and human *Streptococcus agalactiae* isolates. *Journal of Clinical Microbiology*, *43*(3), 1177–1186.

Sullivan, M. J., Petty, N. K., & Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics*, *27*(7), 1009–1010.

Swe, P. M., Heng, N. C., Cook, G. M., Tagg, J. R., & Jack, R. W. (2010). Identification

of DysI, the immunity factor of the streptococcal bacteriocin dysgalacticin. *Applied Environmental Microbiology*, *76*(23), 7885–7889.

Takai, S., Hines, S. A., Sekizaki, T., Nicholson, V. M., Alperin, D. A., Osaki, M., … Prescott, J. F. (2000). DNA sequence and comparison of virulence plasmids from *Rhodococcus equi* ATCC 33701 and 103. *Infection and Immunity*, *68*(12), 6840–6847.

Tan, S., Lin, Y., Foo, K., Koh, H. F., Tow, C., Zhang, Y., … Lin, R. T. (2016). Group B *Streptococcus* serotype III sequence type 283 bacteremia associated with consumption of raw fish, Singapore. *Emerging Infectious Diseases*, *22*(11), 1970.

Tazi, A., Disson, O., Bellais, S., Bouaboud, A., Dmytruk, N., Dramsi, S., … Poyart, C. (2010). The surface protein HvgA mediates group B *Streptococcus* hypervirulence and meningeal tropism in neonates. *Journal of Experimental Medicine*, *207*(11), 2313–2322.

Teatero, S., Athey, T. B., Van Caeseele, P., Horsman, G., Alexander, D. C., Melano, R. G., … Fittipaldi, N. (2015). Emergence of serotype IV group B *Streptococcus* adult invasive disease in Manitoba and Saskatchewan, Canada, is driven by clonal sequence type 459 strains. *Journal of Clinical Microbiology*, *53*(9), 2919–2926.

Teatero, S., McGeer, A., Li, A., Gomes, J., Seah, C., Demczuk, W., … Fittipaldi, N. (2015). Population structure and antimicrobial resistance of invasive serotype IV group B *Streptococcus*, Toronto, Ontario, Canada. *Emerging Infectious Diseases*, *21*(4), 585.

Tenhagen, B. A., Köster, G., Wallmann, J., & Heuwieser, W. (2006). Prevalence of mastitis pathogens and their resistance against antimicrobial agents in dairy cows in Brandenburg, Germany. *Journal of Dairy Science*, *89*(7), 2542–2551.

Tettelin, H. (2009). The bacterial pan-genome and reverse vaccinology. In *Microbial Pathogenomics* (Vol. 6, pp. 35–47). Karger Publishers.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., … DeBoy, R. T. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, *102*(39), 13950–13955.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Eisen, J. A., Peterson, S., Wessels, M. R., … Madoff, L. C. (2002). Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proceedings*

of the National Academy of Sciences, *99*(19), 12391–12396.

Tettelin, H., Riley, D., Cattuto, C., & Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*, *11*(5), 472–477.

Thomas, C. M., & Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology*, *3*(9), 711.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680.

Tocher, D. R. (2003). Metabolism and functions of lipids and fatty acids in teleost fish. *Reviews in Fisheries Science*, *11*(2), 107–184.

Todd, J., Fishaut, M., Kapral, F., & Welch, T. (1978). Toxic-shock syndrome associated with phage-group-I Staphylococci. *The Lancet*, *312*(8100), 1116–1118.

Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D., & Corander, J. (2019). Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Research*, *47*(11), 5539–5549.

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., … Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, *21*(1), 1–21.

Toussaint, A., & Merlin, C. (2002). Mobile elements as a combination of functional modules. *Plasmid*, *47*(1), 26–35.

Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, *15*(11), 524.

Valentin-Domelier, A.-S., Girard, M., Bertrand, X., Violette, J., François, P., Donnio, P.-Y., … van der Mee-Marquet, N. (2011). Methicillin-susceptible ST398 *Staphylococcus aureus* responsible for bloodstream infections: an emerging human-adapted subclone? *PLoS One*, *6*(12).

van der Linden, M., Mamede, R., Levina, N., Helwig, P., Vila-Cerqueira, P., Carriço, J. A., … Martins, E. R. (2020). Heterogeneity of penicillin-non-susceptible group B streptococci isolated from a single patient in Germany. *Journal of Antimicrobial Chemother-*

apy, *75*(2), 296–299.

van der Mee-Marquet, N., Diene, S. M., Barbera, L., Courtier-Martinez, L., Lafont, L., Ouachée, A., … Francois, P. (2018). Analysis of the prophages carried by human infecting isolates provides new insight into the evolution of group B *Streptococcus* species. *Clinical Microbiology and Infection*, *24*(5), 514–521.

van der Mee-Marquet, N., Domelier, A. S., Mereghetti, L., Lanotte, P., Rosenau, A., van Leeuwen, W., & Quentin, R. (2006). Prophagic DNA fragments in *Streptococcus agalactiae* strains and association with neonatal meningitis. *Journal of Clinical Microbiology*, *44*(3), 1049–1058.

van der Mee-Marquet, N., Fourny, L., Arnault, L., Domelier, A. S., Salloum, M., Lartigue, M. F., & Quentin, R. (2008). Molecular characterization of human-colonizing *Streptococcus agalactiae* strains isolated from throat, skin, anal margin, and genital body sites. *Journal of Clinical Microbiology*, *46*(9), 2906–2911.

Verner-Jeffreys, D. W., Baker-Austin, C., Pond, M. J., Rimmer, G. S., Kerr, R., Stone, D., … Feist, S. W. (2012). Zoonotic disease pathogens in fish used for pedicure. *Emerging Infectious Diseases*, *18*(6), 1006.

Verner-Jeffreys, D. W., Wallis, T. J., Cano Cejas, I., Ryder, D., Haydon, D. J., Domazoro, J. F., … Bean, T. (2018). *Streptococcus agalactiae* multilocus sequence type 261 is associated with mortalities in the emerging Ghanaian tilapia industry. *Journal of Fish Diseases*, *41*(1), 175–179.

Vernikos, G., Medini, D., Riley, D. R., & Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology*, *23*, 148–154.

Viana, D., Blanco, J., Tormo-Más, M. Á., Selva, L., Guinane, C. M., Baselga, R., … Penadés, J. R. (2010). Adaptation of *Staphylococcus aureus* to ruminant and equine hosts involves SaPI-carried variants of von Willebrand factor-binding protein. *Molecular Microbiology*, *77*(6), 1583–1594.

Viana, D., Comos, M., McAdam, P. R., Ward, M. J., Selva, L., Guinane, C. M., … Penadés, J. R. (2015). A single natural nucleotide mutation alters bacterial pathogen host tropism. *Nature Genetics*, *47*(4), 361–366.

Wallden, K., Rivera-Calzada, A., & Waksman, G. (2010). Microreview: Type IV secretion systems: versatility and diversity in function. *Cellular Microbiology*, *12*(9), 1203–

1212.

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Weinert, L. A., Welch, J. J., Suchard, M. A., Lemey, P., Rambaut, A., & Fitzgerald, J. R. (2012). Molecular dating of human-to-bovid host jumps by *Staphylococcus aureus* reveals an association with the spread of domestication. *Biology Letters*, *8*(5), 829–832.

Wetzel, R. G. (2001). Structure and productivity of aquatic ecosystems. In *Limnology: Lake and River Ecosystems* (pp. 129–150). Elsevier.

Whidbey, C., Harrell, M. I., Burnside, K., Ngo, L., Becraft, A. K., Iyer, L. M., … Rajagopal, L. (2013). A hemolytic pigment of group B *Streptococcus* allows bacterial penetration of human placenta. *Journal of Experimental Medicine*, *210*(6), 1265–1281.

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, *13*(6), e1005595.

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, *20*(1), 129.

Wilkinson, H. W., Thacker, L., & Facklam, R. (1973). Nonhemolytic group B streptococci of human, bovine, and ichthyic origin. *Infection and Immunity*, *7*(3), 496.

Wilson, D. J., Gonzalez, R. N., & Das, H. H. (1997). Bovine mastitis pathogens in New York and Pennsylvania: prevalence and effects on somatic cell count and milk production. *Journal of Dairy Science*, *80*(10), 2592–2598.

Wozniak, R. A., Fouts, D. E., Spagnoletti, M., Colombo, M. M., Ceccarelli, D., Garriss, G., … Waldor, M. K. (2009). Comparative ICE genomics: insights into the evolution of the SXT/R391 family of ICEs. *PLoS Genetics*, *5*(12), e1000786.

Wozniak, R. A., & Waldor, M. K. (2009). A toxin–antitoxin system promotes the maintenance of an integrative conjugative element. *PLoS Genetics*, *5*(3), e1000439.

Wozniak, R. A., & Waldor, M. K. (2010). Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nature Reviews Microbiology*, *8*(8), 552.

Wu, B., Su, J., Li, L., Wu, W., Wu, J., Lu, Y., … Liang, X. (2019). Phenotypic and ge-

netic differences among group B *Streptococcus* recovered from neonates and pregnant women in Shenzhen, China: 8-year study. *BMC Microbiology*, *19*(1), 185.

Yang, Q. E., & Walsh, T. R. (2017). Toxin–antitoxin systems and their role in disseminating and maintaining antimicrobial resistance. *FEMS Microbiology Reviews*, *41*(3), 343–353.

Yang, Y., Liu, Y., Ding, Y., Yi, L., Ma, Z., Fan, H., & Lu, C. (2013). Molecular characterization of *Streptococcus agalactiae* isolated from bovine mastitis in Eastern China. *PLoS One*, *8*(7), e67755.

Yanong, R. P. E., & Francis-Floyd, R. (2010). Streptococcal infections of fish. *Fisheries and Aquatic Sciences Department, Univeristy of Florida/IFAS Extension*, *Circular 57*.

Yao, K., Poulsen, K., Maione, D., Rinaudo, C. D., Baldassarri, L., Telford, J. L., … Kilian, M. (2013). Capsular gene typing of *Streptococcus agalactiae* compared to serotyping by latex agglutination. *Journal of Clinical Microbiology*, *51*(2), 503–507.

Yildirim, A. Ö., Lämmler, C., & Weiss, R. (2002a). Identification and characterization of *Streptococcus agalactiae* isolated from horses. *Veterinary Microbiology*, *85*(1), 31–35.

Yildirim, A. Ö., Lämmler, C., Weiß, R., & Kopp, P. (2002b). Pheno- and genotypic properties of streptococci of serological group B of canine and feline origin. *FEMS Microbiology Letters*, *212*(2), 187–192.

Yoganandi, J., Mehta, B. M., Wadhwani, K., Darji, V., & Aparnathi, K. (2014). Evaluation and comparison of camel milk with cow milk and buffalo milk for gross composition. *Journal of Camel Practice and Research*, *21*(2), 259–265.

Younan, M. (2002). Traitement parentéral de mammites à *Streptococcus agalactiae* chez le dromadaire (*Camelus dromedarius*) au Kenya. *Revue d'élevage et de Médecine Vétérinaire des Pays Tropicaux*, *55*(3), 177–181.

Younan, M., & Bornstein, S. (2007). Lancefield group B and C streptococci in East African camels (*Camelus dromedarius*). *Veterinary Record*, *160*(10), 330–335.

Yue, M., Han, X., De Masi, L., Zhu, C., Ma, X., Zhang, J., … Schifferli, D. M. (2015). Allelic variation contributes to bacterial host specificity. *Nature Communications*, *6*(1), 1–11.

Zadoks, R. N., & Fitzpatrick, J. (2009). Changing trends in mastitis. *Irish Veterinary*

Journal, *62*(4), 1–12.

Zadoks, R. N., Middleton, J. R., McDougall, S., Katholm, J., & Schukken, Y. H. (2011). Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. *Journal of Mammary Gland Biology and Neoplasia*, *16*(4), 357–372.

Zadoks, R. N., & Schukken, Y. H. (2006). Use of molecular epidemiology in veterinary practice. *Veterinary Clinics: Food Animal Practice*, *22*(1), 229–261.

Zadoks, R. N., Van Leeuwen, W., Kreft, D., Fox, L., Barkema, H., Schukken, Y., & Van Belkum, A. (2002). Comparison of *Staphylococcus aureus* isolates from bovine and human skin, milking equipment, and bovine milk by phage typing, pulsed-field gel electrophoresis, and binary typing. *Journal Clinical Microbiology*, *40*(1), 3894–3902.

Zamri-Saad, M. (2018). GBS in fish and feed based vaccination. *Abstract for the 1st International Symposium on Streptococcus agalactiae Disease, Cape Town (South Africa)*.

Zamri-Saad, M., Amal, M. N. A., & Siti-Zahrah, A. (2010). Pathological changes in red tilapias (*Oreochromis* spp.) naturally infected by *Streptococcus agalactiae*. *Journal of Comparative Pathology*, *143*(2-3), 227–229.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., … Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, *67*(11), 2640–2644.

Zecconi, A., & Zanirato, G. (2013). Il controllo delle mastiti per un'allevamento sostenibile. *Litografia SAB-Budrio (BO)*, 1–69.

Zeng, L., & Burne, R. A. (2021a). Molecular mechanisms controlling fructose-specific memory and catabolite repression in lactose metabolism by *Streptococcus mutans*. *Molecular Microbiology*, *115*(1), 70–83.

Zeng, L., & Burne, R. A. (2021b). Subpopulation behaviors in lactose metabolism by *Streptococcus mutans*. *Molecular Microbiology*, *115*(1), 58–69.

Zeng, L., Das, S., & Burne, R. A. (2010). Utilization of lactose and galactose by *Streptococcus mutans*: transport, toxicity, and carbon catabolite repression. *Journal of Bacteriology*, *192*(9), 2434–2444.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829.

Zhang, D., Li, A., Guo, Y., Zhang, Q., Chen, X., & Gong, X. (2013). Molecular characterization of *Streptococcus agalactiae* in diseased farmed tilapia in China. *Aquaculture*, *412*, 64–69.

Zhang, X. Y., Fan, H. P., Zhong, Q. F., Zhuo, Y. C., Lin, Y., & Zeng, Z. Z. (2008). Isolation, identification and pathogenicity of *Streptococcus agalactiae* from tilapia. *Journal of Fisheries of China*, *5*, 772–779.

Zhang, Z., Yu, A., Lan, J., Zhang, H., Hu, M., Cheng, J., ... Wei, S. (2017). GapA, a potential vaccine candidate antigen against *Streptococcus agalactiae* in Nile tilapia (*Oreochromis niloticus*). *Fish and Shellfish Immunology*, *63*, 255–260.

Zhao, Z., Kong, F., Martinez, G., Zeng, X., Gottschalk, M., & Gilbert, G. L. (2006). Molecular serotype identification of *Streptococcus agalactiae* of bovine origin by multiplex PCR-based reverse line blot (mPCR/RLB) hybridization assay. *FEMS Microbiology Letters*, *263*(2), 236–239.

Zhou, K., Xie, L., Han, L., Guo, X., Wang, Y., & Sun, J. (2017). ICE*Sag37*, a novel integrative and conjugative element carrying antimicrobial resistance genes and potential virulence factors in *Streptococcus agalactiae*. *Frontiers in Microbiology*, *8*, 1921.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., & Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Research*, *39*(suppl_2), W347–W352.

Zubair, S., de Villiers, E., Younan, M., Andersson, G., Tettelin, H., Riley, D., ... RP, B. (2013). Genome sequences of two pathogenic *Streptococcus agalactiae* isolates from the one-humped camel, *Camelus dromedarius*. *Genome Announcements*, *1*, e00515-13.