Bioinformatics analysis of bacterial pathogens from East African camels

Comparative genomics of *Streptococcus agalactiae* and *Staphylococcus aureus*

Saima Zubair

Faculty of Veterinary Medicine and Animal Sciences
Department of Animal Breeding and Genetics
Uppsala

Doctoral Thesis
Swedish University of Agricultural Sciences
Uppsala 2015

Acta Universitatis agriculturae Sueciae
2015:22

Cover: Comparative genomics of *S. agalactiae* and *S. aureus* from camels
(photo: Saima Zubair)

ISSN 1652-6880
ISBN (print version) 978-91-576-8242-0
ISBN (electronic version) 978-91-576-8243-7

Print: SLU Service/Repro, Uppsala 2015

Bioinformatics analysis of bacterial pathogens from East African camels. Comparative genomics of *Streptococcus agalactiae* and *Staphylococcus aureus*.

Abstract

The camel is the most valuable livestock species in arid and semi-arid regions in the Greater Horn of Africa. *Streptococcus agalactiae* and *Staphylococcus aureus* are important pathogens for a wide range of hosts including camels, cattle and humans. *Streptococcus agalactiae* has been reported to cause infections of the skin, the respiratory tract, the mammary gland and the vaginal tract in camels. *Staphylococcus aureus* has been isolated from the nasal cavity, wound infections and mastitis from camels. Both pathogens account for decline in health and productivity of camels, hence causing economic losses to the inhabitants of arid and semi arid lands.

To define candidate virulence traits in these bacteria, we compared the genomes of *S. agalactiae* and *S. aureus*. We sequenced and completely assembled the genomes of two *S. agalactiae* isolates ILRI005 and ILRI112 from abscessed case camels and an *S. aureus* isolate ILRI_Eymole1/1 from the nasal swab of camel in Kenya. To perform comparative analysis, we also sequenced and assembled an *S. agalactiae* isolate 09mas018883 from subclinical mastitis case cattle in Sweden. Mapping assembly, *de novo* assembly and post-assembly genome finishing were performed to obtain completely assembled genomes.

Comparative genomics approach was applied to explore the genetic heterogeneity, core genome construction and protein repertoire comparison of these novel genomes, and to highlight potential virulence factors that could have contributed to the pathogenicity of these isolates in their hosts. Newly sequenced camel *S. agalactiae* genomes were compared with human and cattle *S. agalactiae* genomes. This comparison revealed that the two camel isolates were genetically close to each other but relatively distinct from other isolates, while cattle isolate 09mas018883 was genetically closer to the human isolates. Large proportion of the isolate-specific genes of the camel *S. agalactiae* isolates was clustered in putative phage insertions and genomic islands suggesting the lateral transfer of these putative phages. The two camel *S. agalactiae* isolates shared a novel potential virulent locus, the *CRISPR2* (Cluster Regularly Interspaced Palindromic Repeats) locus. The two cattle *S. agalactiae* isolates and three human *S. agalactiae* isolates contained similar putative phage insertions. Important potential pathogenic factors found in all *S. agalactiae* isolates were *CRISPR1* locus, *cyl* locus, capsular polysaccharide locus and pilus islands.

Phylogenetic analysis of novel camel *S. aureus* genome of strain type ST30 and previously sequenced human *S. aureus* genomes of type Clonal Complex 30 (CC30) revealed that camel *S. aureus* isolate is genetically distinct from human *S. aureus* isolates of the same sequence type. Important features were also identified such as genes encoding bacterial adhesins and secretory proteins.

The availability of genomic sequences of *S. agalactiae* and *S. aureus* from camels, their detailed bioinformatics analysis and identified potential virulence factors will foster the development of control measures such as molecular diagnostic assays and vaccines for control of *S. agalactiae* and *S. aureus* infections in camels. This will ensure improvement in health and productivity of camels.

Keywords: Streptococcus agalactiae, *Staphylococcus aureus*, genome assembly, comparative genomics, pathogenicity islands, virulence, abscesses, mastitis, camel, cattle

Author's address: Saima Zubair, SLU, Department of Animal Breeding and Genetics,
P.O. Box 7023, 750 07 Uppsala, Sweden
E-mail: Saima.Zubair@slu.se

Dedication

To Allah Almighty, Whose Knowledge and Power is above all.

> ***Oh Allah! Benefit me by that which You have taught me, and teach me that which will benefit me, and increase me in knowledge.***

Prayer by: Prophet Muhammad (PBUH)

Contents

List of Publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I **Zubair S**, de Villiers EP, Younan M, Andersson G, Tettelin H, Riley DR, Jores J, Bongcam-Rudloff E, Bishop RP (2013). Genome sequences of two pathogenic *Streptococcus agalactiae* isolates from the one-humped camel *Camelus dromedarius*. *Genome Announc* 1(4), e00515–13.

II **Zubair S**, de Villiers EP, Fuxelius HH, Andersson G, Johansson K-E, Bishop RP, Bongcam-Rudloff E (2013). Genome sequence of *Streptococcus agalactiae* strain 09mas018883, isolated from a Swedish cow. *Genome Announc* 1(4), e00456–13.

III **Zubair S**, de Villiers EP, Tettelin H, Mustafa MI, Andersson G, Bishop RP, Bongcam-Rudloff E (2015). Comparative Genomics of Mammalian *Streptococcus agalactiae* Isolates from Camels, Cattle and Humans. Manuscript.

IV **Zubair S**, Fischer A, Liljander A, Gourlé H, Bishop RP, Roebbelen I, Younan M, Mustafa MI, Mushtaq M, Bongcam-Rudloff E, Jores J (2014). Complete genome sequence of *Staphylococcus aureus*, strain ILRI_Eymole1/1, isolated from a Kenyan dromedary camel. Standards in Genomic Sciences. Submitted.

Papers I-II are reproduced with the permission of the publishers.

The contribution of Saima Zubair to the papers included in this thesis was as follows:

I Partly planned the study, performed genome assembly and annotation, and prepared the manuscript.

II Majorly planned the study, performed genome assembly and annotation, and prepared the manuscript.

III Planned the study, performed comparative analysis, and prepared the manuscript.

IV Partly planned the study, performed genome assembly, annotation and comparative analysis, and prepared the manuscript.

Additional publication

Mushtaq M, **Zubair S**, Råsbäck T, Bongcam-Rudloff E, Jansson DS (2015). *Brachyspira suanatina* sp. nov., an enteropathogenic intestinal spirochaete isolated from pigs and mallards: genomic and phenotypic characteristics. Manuscript.

Abbreviations

FAO	Food and Agriculture Organization
GBS	Group B *Streptococcus*
CMT	California mastitis test
SCC	Somatic cell count
IMI	Intramammary infection
BRCA1	Brease cancer 1 gene
SCS	Somaitic cell score
MHC	Major histocompatibility complex
BoLA-DRB3	Bovine leukocyte antigen, DR beta 3
NGS	Next generation sequencing
SBS	Sequencing by synthesis
CRT	Cyclic reverse termination
SBL	Sequencing by ligation
OLC	Overlap-Layout-Consensus
DBG	de Bruijn Graph
ANI	Average nucleotide identity
CC30	Clonal complex 30
MLST	Multi locus sequence type
CDS	Coding DNA sequence
COG	Clustering of orthologous groups
bp	Base pairs
Mbp	Million base pairs
Blastp	Protein blast
Blastn	Nucleotide blast
BLAST	Basic local alignment search tool
LGT	Lateral gene transfer
CRISPR	Clustered regularly interspaced short palindromic repeats
Cas	CRISPR-associated

1 Background

1.1 Significance of dromedary camels in arid lands

Although knowledge about precise origin and dispersion of camels is lacking, it is evident that these were domesticated in Arabian peninsular and Central Asia during second millennium BC, and were of great economic importance in these areas. The fossils of the genus *Camelus* were identified from north-eastern China, north-western Mongolia, Tadzhikistan, Kazakhstan, Harrapa and Mohenjo-daro, Pakistan and Kalibangan, north-western India. The specimens found from Harrapa (third millennium BC), Pakistan were recorded as *Camelus dromedarius* (Peters & von den Driesch, 1997). According to FAO Statistics in 2004, there is a total 18.9 million population of camels worldwide (Bornstein & Younan, 2013), about 95% of which are dromedary camels, of which 73% are located in Africa (Kaufmann, 1998; Bornstein & Younan, 2013). The camel pastoralists started migrating to northern Kenya between 10th and 13th centuries A.D. according to the traces found in Chalbi desert (Stiles, 1987). Surviving in the hot, harsh and arid climate is a big challenge for livestock, but camel is a special livestock species that efficiently survive and produce in such lands by tolerating lack of water and vegetation. The dromedary camels are excellent sources of food and food products for the pastoralists and inhabitant of these regions. In Kenya about 50 to 60% of the whole nutrient intake is fulfilled by camel milk among the pastoralists. Unlike other livestock animals, camels maintain their milk production during the entire dry seasons with longer lactation periods of between 12 to 18 months. According to a study conducted by Field and Simpkin in 1985, a lactating camel's milk production in dry season is equal to the milk production of five zebu cows in the wet season. The physical health status and growth of camels is extremely valuable for sufficient meat and milk production to meet the demands of a steadily increasing population of these countries (Bornstein &

Younan, 2013). Camels not only meet the economic demands of the pastoralists but are also used for transport, ecotourism as well as social and religious services, hence it is the most preferred livestock species in the region (Kagunyu & Wanjohi, 2014). In spite of its great economic significance for arid lands, only limited research has been carried out on camels. Infections arising from bacterial pathogens are greatly affecting camel health, production, and calf growth so it is essential to conduct research to explore the molecular biology of bacteria hazardous for camels. *Streptococcus agalactiae* and *Staphylococcus aureus* are the two most common pathogens found isolated from intramammary infections in Kenyan camels, and *S. agalactiae* from skin, joint, respiratory and vaginal infections (Bornstein & Younan, 2013).

1.2 *Streptococcus agalactiae*

Streptococcus agalactiae or Group B *Streptococcus* (GBS) are spherical cell shaped, non-motile, chain-forming and nonspore-forming, Gram-positive bacteria, shown in Figure 1 (left). In Gram-positive bacteria the cell wall is composed predominantly of peptidoglycan on which various carbohydrates, bacterial polysaccharides (teichoic acid) and surface antigens are attached. The cell wall polysaccharides of streptococcal species are critically important in determining the Lancefield serological grouping of strains on the basis of surface protein antigen (Lancefield, 1933). Capsular polysaccharide antigen and surface protein antigen determined ten serotypes Ia, Ib and II to IX in Group B *Streptococcus*. Majority of the neonatal infections in humans are caused by types I, II, III, and V (Whiley & Hardie, 2009; Imperi *et al.*, 2010).

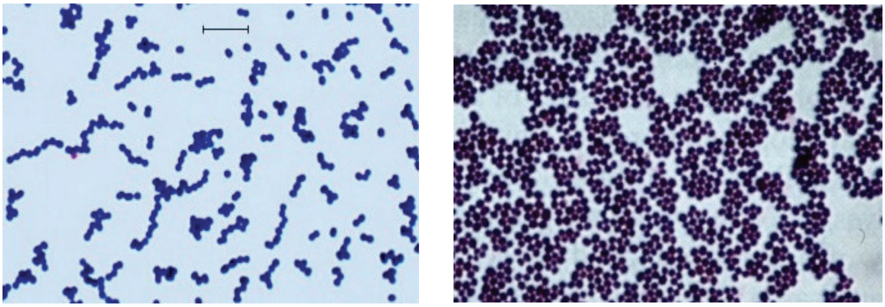


Figure 1. Gram stain view of *Streptococcus agalactiae* 09mas018883 (left): showing chain-forming clusters. Image adapted from VetBact, Karl-Erik Johansson, SVA Uppsala. *Staphylococcus aureus* (right): showing grape-like clusters. Image adapted from pixgood.com.

1.3 *Streptococcus agalactiae* infections in dromedary camels

Streptococcus agalactiae is a commensal and common opportunistic pathogen in East African camels. In a healthy carrier state, *S. agalactiae* is found on the nasopharynx and non-abscessed lymph nodes, while during the clinical infectious state it is found on skin abscesses, abscessed lesions, abscessed subcutaneous, peri-arthricular abscesses, abscessed lymph nodes, tick bite lesions, respiratory infections, vaginal infections, mastitis: udder infection, arthritis and gingivitis: gum infection (Younan & Bornstein, 2007; Bornstein & Younan, 2013).

1.4 Abscesses in camels associated with *S. agalactiae*

A skin abscess is the inflammation, swelling and soreness of the dermis and subcutaneous tissue in which pus accumulates (Singer & Talan, 2014). In camels, *S. agalactiae* has been found in the abscesses of skin, lesions, subcutaneous tissue and lymph nodes (Younan & Bornstein, 2007). According to a study conducted on camel calves in North Kenya, *S. agalactiae* causes a condition named peri-arthricular abscesses, with inflammation and pus accumulation around joints in camel calves. Peri-arthricular abscesses were present around elbow (33.3%), tarsus (29.2%), carpus (25.0%), knee (8.3%) and fetlock (4.2%) joints. The calves locomotion and suckling-ability was affected due to the pain in case of large multiple peri-arthricular abscesses. These abscesses can penetrate deep into the joints and cause the destruction of ligaments and tendons that lead to necrotising arthritis in nearby joints. Stunted growth and mortality was also observed in case of the chronic peri-arthricular abscesses. This study reported that 82% of the cases with peri-arthricular abscesses were exclusively associated with *S. agalactiae*, 4% of the cases showed infection of *S. agalactiae* and *Streptococcus equi zooepidemicus* and only 2% of the cases showed the single infection of *mucoid Streptococcus equi zooepidemicus*, Figure 2 (Younan *et al.*, 2007).

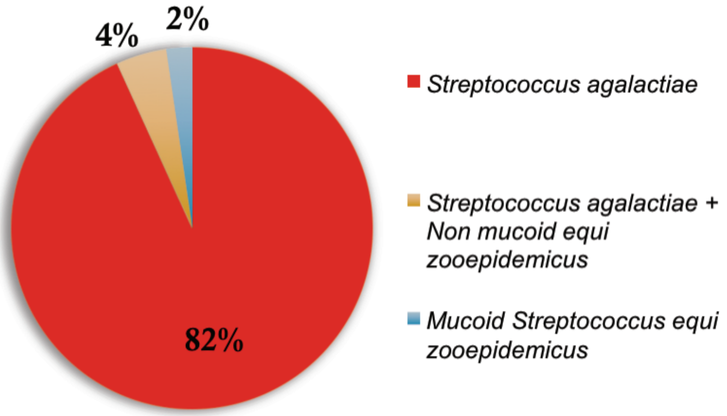


Figure 2. Percentage of abscesses cases in camels with their bacterial composition, in North Kenya.

1.5 *Staphylococcus aureus*

Staphylococcus aureus are Gram-positive, non-motile, nonspore-forming cocci occurring either singly or in pairs or in short chains of 3 to 4 cells characteristically arranging in grape like irregular clusters, shown in Figure 1 (right). Like other Gram-positive bacteria, the cell wall of *Staphylococci* is chemically composed of peptidoglycan, teichoic acid, and proteins (Schleifer & Bell, 2009).

Staphylococcus aureus not only colonizes and infects humans but also other mammalian animal species including cattle, camel, horse, goat, sheep, cat, dog, rabbit, pig as well as several bird species (Sung *et al.*, 2008; Smyth *et al.*, 2009). In humans, approximately 20% of individuals are found as persistent nasal carriers of *S. aureus*, about 30% as occasional carriers while about 50% as non-carriers. *Staphylococcus aureus* nasal carriers have high risk of getting infections associated with these bacteria but the underlying factors need to be elucidated. *Staphylococcus aureus* is the common cause of both community-acquired (CA) and hospital-acquired (HA) methicillin-susceptible *S. aureus* (MSSA) infections and methicillin-resistant *S. aureus* (MRSA) infections in humans, and it is becoming difficult to treat them due to the increased antibiotic resistance of *S. aureus* (Wertheim *et al.*, 2005; Aiken *et al.*, 2014). It also causes a wide range of other infections in humans like endocarditis, toxic shock syndrome and pneumonia (Fitzgerald, 2012). It is essential to elucidate underlying factors involved in conversion of *S. aureus* from nasal carrier state to the pathogenic state to prevent new infections.

1.6 *Staphylococcus aureus* infections in dromedary camels

Staphylococcus aureus is the major cause of ruminant mastitis infection and has strong economic impact on productivity losses of dairy industry worldwide. Some *S. aureus* strains pathogenic for animals appear to have zoonotic potential for humans through host adaptation (Guinane *et al.*, 2010; Fitzgerald, 2012). *Staphylococcus aureus* is the cause of mastitis in camels and according to a study performed on camels in Ethiopia was found as the most abundant pathogen in milk samples taken from camels affected by mastitis (Regassa *et al.*, 2013). In Kenya and Sudan, the intramammary infections in camels showed the prevalence of *S. aureus* as the second most frequent pathogen after *S. agalactiae* (Obied & Bagadi, 1996; Younan *et al.*, 2001). Moreover, infection of the joints in camel calves (Bani Ismail *et al.*, 2007), eye infections (Yeruh *et al.*, 2002), respiratory syndromes and subclinical pneumonia in dromedary camels have also been reported to be associated with *S. aureus* (Wareth *et al.*, 2014). A significant high percentage of 89.1% *S. aureus* has been observed in the nasal isolates of healthy dromedary camels in Saudi Arabia (Alhendi, 1999). Likewise in another study, samples from nasal swabs, tracheal swabs and pneumonic lung tissues were examined, and the predominant bacteria were *S. aureus* and *Corynebacterium* (Al-Doughaym *et al.*, 1999).

1.7 Mastitis in dairy camels and cattle

Mastitis, an inflammation of the mammary gland in dairy animals occurs either as non-infectious mastitis or infectious mastitis. Less often it occurs as non-infectious mastitis due to physical injury, improper milking and chilling, while most often it occurs as an infectious mastitis due to the bacterial pathogens (Sori *et al.*, 2005; Tamiru *et al.*, 2013). Clinical mastitis is characterized by clinical symptoms of swelling, hardening, redness, elevated temperature of the udder tissue, decreased and affected milk secretion, pain, depression, fever and loss of appetite. On the other hand, subclinical mastitis is not associated with apparent clinical signs and therefore this condition is usually undetectable and can cause the spread of bacteria among herd animals. California Mastitis Test (CMT) is used to detect increase in somatic cell count (SCC) in milk samples as a diagnostic measure for mastitis. The milk from an unaffected udder contains less than 200,000 somatic cells per ml while that from an affected one contains SCC of greater than 300,000 (Hillerton, 1999; Khan & Khan, 2006; Abdelgadir, 2014). A review study described that during last decades the cases of mastitis in dromedary camels have been reported from many camels-rearing counties of Africa and Asia such as Kenya, Somalia, Sudan, Egypt, Saudi

Arabia, Iraq and UAE (Abdelgadir, 2014). A variety of **factors** are involved in mastitis onset (Khan & Khan, 2006; Zhang *et al.*, 2009), a few are discussed here.

1.7.1 Bacterial pathogens

Many bacterial pathogens have been found associated with camel mastitis such as *S. aureus*, *S. agalactiae*, *Bacillus cereus*, *Actinomyces pyogenes*, *E. coli*, *Micrococcus spp.*, and *Corynebacterium bovis* (Abdelgadir, 2014). However, the two most common, mastitis causing contagious pathogens in camels and cattle are *S. aureus* and *S. agalactiae* (Younan *et al.*, 2001; Khan & Khan, 2006; Ahmad *et al.*, 2012). Bacterial pathogens cross the natural protective sphincter opening of the teat muscle, and proliferate inside the epithelium lining of the udder tissue. Various toxins, enzymes and cell wall components are released and cause fluid accumulation, as a result inflammatory mediators are produced to attract phagocytes. Large numbers of neutrophils or leukocytes are passed into the lumen and cause increases in SCC. The accumulation of these leukocytes and blood clotting factors may cause complete blockage of mammary ducts making it difficult for antibiotics to penetrate the affected udder tissue that may suffer permanent loss of function (Khan & Khan, 2006; Jones, 2009). Mastitis control through vaccine development can be a better solution.

Prevalence of mastitis associated with S. agalactiae and S. aureus

A study performed on lactating camels in Kenya from 1998 to 2000 has reported the prevalence of intramammary infections (IMIs) with *S. agalactiae* as 12% while IMIs with *S. aureus* as 11% of the sampled camels. CMT sensitivities for *S. agalactiae* and *S. aureus* in camels were 77% and 68%, respectively (Younan *et al.*, 2001). According to a cross-sectional study conducted in Jhang, Pakistan from November 2008 to October 2009 on 150 lactating camels, a total of 69 (46%) were positive for mastitis with 12 (8%) clinical and 57 (38%) subclinical. Sixty-four samples were culturally positive and contained 26.56% *S. aureus* and 15.63% *S. agalactiae* as the most predominant pathogens. Other pathogens were *E. coli*, *Bacillus spp.*, *Corynebacterium* and *Candida spp.* (Ahmad *et al.*, 2012). A study conducted in UAE showed the prevalence of clinical and subclinical mastitis in camels as 24.7% and 11.67%, respectively and the most abundant pathogens were *Staphylococcus* (41.67%) and *Streptococcus spp.* (21.67%). Other pathogens were *Enterobacter spp.*, *C. pyogenes*, *Micrococcus spp.*, *Pasteurells spp.* and *Pseudomonas aeruginosa* (Al-Juboori *et al.*, 2013).

A study conducted to estimate the prevalence of mastitis from October 2008 to May 2009 in and around Hawassa, southern Ethiopia, reported Staphylococci species and *S. agalactiae* as two of the most abundant pathogens found in infected samples with high SCC. The bacterial composition of mastitis samples is shown in Figure 3 (Abera *et al.*, 2012). According to a study performed in dairy farms of central Ethiopia to investigate the prevalence of mastitis from November 2008 to April 2009, a 71% prevalence of mastitis was observed in cattle, out of which the prevalence of subclinical and clinical mastitis were 48.6% and 22.4%, respectively; and *Staphylococci* and *Streptococci* were the two most abundant pathogens (Mekibib *et al.*, 2010).

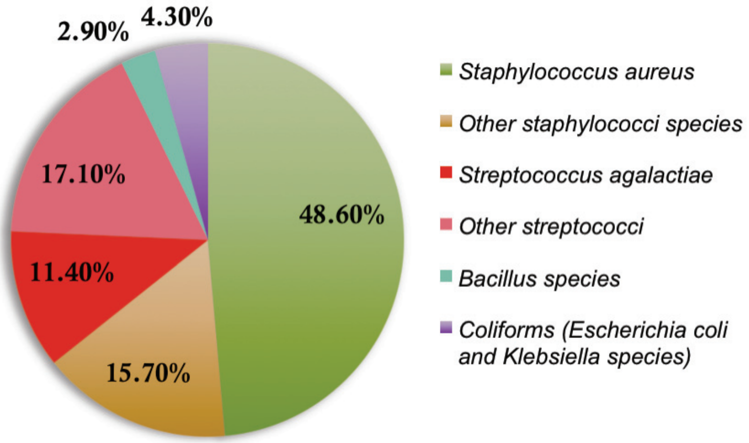


Figure 3. Relative abundance of bacterial isolates from cattle mastitis samples.

1.7.2 Risk factors

The susceptibility of camels and cows to develop mastitis infection depends on a variety of other factors like poor management practices, animal health, age, parity, lactation stage, cross-suckling by calves, milk machines, over-milking and cleanliness status of the area where animals are kept (Sori *et al.*, 2005; Khan & Khan, 2006; Zhang *et al.*, 2009; Abdelgadir, 2014). Moreover, use of anti-suckling devices to prevent suckling by camel calves, tick bites on udder, deformities of udder tissue due to thorny bushes in pastoral areas, and camel pox have been reported as risk factors for camel mastitis (Younan *et al.*, 2001; Abdelgadir, 2014).

1.7.3 Genetic risk factors of mastitis in cattle

Several genes have been reported to be associated with mastitis in cattle either by increasing susceptibility or resistance for mastitis. In an association study performed on Holstein, Sanhe and Simmental cows, a candidate gene for mastitis, called breast cancer 1 *BRCA1* gene is found to be associated with mastitis. Three genetic variants/SNPs G22231T, T25025A and C28300A were identified in *BRCA1* gene and the genetic effects of 24 combined genotypes on somatic cell score (SCS) were investigated. Genotype BBDDFF showed significant association with highest SCS while AACCEE had significant association with lowest SCS in milk samples (Yuan *et al.*, 2012), suggesting *BRCA1* gene as mastitis susceptibility or resistance gene based on the combination of alleles. Major histocompatibility complex (MHC), class II gene *BoLA-DRB3* known for its essential role in the immune response of dairy cattle against pathogens, is reported to be related to mastitis resistance as well as mastitis susceptibility under the influence of environmental factors like certain pathogens (Galal *et al.*, 2008; Sender *et al.*, 2013). No relation of allele *BoLA-DRB3.2*16* and *BoLA-DRB3.2*23* with SCC was observed in the presence of contagious pathogen *S. aureus*, however increased susceptibility of *BoLA-DRB3.2*23* to sub-clinical mastitis was observed in the presence of environmental pathogen *Streptococcus dysgalactiae* (Galal *et al.*, 2008). Likewise, *BoLA-DRB3.2*24* and *BoLA-DRB3.2**22 alleles showed association with mastitis susceptibility and *BoLA-DRB3.2*3* and *BoLA-DRB3.2*11* showed association with mastitis resistance, however many other *BoLA-DRB3.2* alleles had both responses (RUpp & BOichard, 2003). In another study, toll-like receptor 2 gene *TLR2* of essential role in the innate immune response to pathogens is reported to be important for mastitis resistance in cattle (Zhang *et al.*, 2009). Leptin gene *LEP* is found to be involved in reduction of SCC in Jersey cows (Kulig *et al.*, 2010). The chemokine gene interleukin 8 *IL8* and the chemokine receptor genes, interleukin 8 receptor, alpha *IL8RA* and *CCR2* are found to be associated with increased SCS and udder depth in Canadian Holsteins (Leyva-Baca *et al.*, 2007). Moreover a large number of other genes are also known to be related to mastitis, such as toll-like receptor 4 *TLR4*, lactoferrin gene, mannan-binding lectin *MBL*, ATPase subunit alpha-1 *ATP1A1*, complement component 5a receptor 1 *C5AR1*, *CD14* antigen, interferon gamma *IFNG*, interleukin 1 beta *IL1B*, interleukin 6 *IL6*, lipopolysaccharide binding protein *LBP*, serum amyloid A3 *SAA3* and tumor necrosis factor *TNF* (Detilleux, 2009; Ogorevc *et al.*, 2009; Sender *et al.*, 2013).

1.7.4 Economic loss of mastitis

Mastitis is an economic problem and is of great concern for the dairy industry worldwide due to associated economic losses, although the costs might vary for different regions. Parity, stage of lactation, bacterial pathogens and some other factors contribute to the economic loss. Under Dutch circumstances, the average cost per cow for clinical mastitis in dairy cattle is calculated as €277 during 1-3 months after calving and €168 during 4-9 months after calving. The cost for clinical mastitis are estimated as €293 for staphylococci, €270 for streptococci and €263 for E. coli (Hogeveen, 2005). The mastitis annual cost in USA is estimated as nearly $1.8 billion for about 9 million dairy cows, excluding additional costs such as costs related to antibiotic remnants in human diet, controlling milk quality and nutrition, and degradation of damaged milk (Schroeder, 2012). Both clinical and subclinical mastitis cause economic damages in the form of reduced milk production, discarded milk, reduced milk quality and unstable taste, decreased efficiency of milk processing, decreased shelf life, reduced yield of milk products such as cheese. Furthermore, the costs associated with drugs, management and treatment of disease-affected cattle, disease spread risk, culling, veterinarians and labour are substantial. Prevention of subclinical mastitis can be beneficial at many levels for mastitis management (Hogeveen, 2005).

1.8 Antibiotics and antibiotics resistance genes

Antibiotic therapy has been effectively used to treat infectious diseases, improve health, reduce disease incidence, morbidity and mortality of humans and animals, and increase the productivity of food-producing animals. However, the use of antibiotics is a key concern for veterinary and human health these days due to emergence and dissemination of antimicrobial resistance in pathogens (Oliver *et al.*, 2011). Many different antibiotics are used as control program for mastitis in dairy animals such as penicillin, ampicillin, erythromycin, tetracycline, oxacillin, ephalothin, ceftiofur, gentamicin, pirlimycin, cephalosporins, lincosamides, non-cephalosporin beta-lactams, aminoglycosides, kanamycin and chloramphenicol (Barlow, 2011; Oliver *et al.*, 2011; Abdelgadir, 2014). Antimicrobial resistance in *S. agalactiae* occurs due to many antimicrobial resistance genes such as *ermA*/*TR*, *ermB*, *ermC*, *mefA*, *tetK*, *tetL*, *tetM*, *tetO*, *aphA-3* and *aad-6*. These genes show resistance against erythromycin, tetracycline and aminoglycosides (Dogan *et al.*, 2005; Gao *et al.*, 2012). *In vitro* susceptibility testing performed on *S. agalactiae* isolates from Kenyan camels revealed the resistance to tetracycline through *tetM* gene in 34% isolates (Fischer et al., 2013). MRSA is a major cause of healthcare associated infections worldwide (Wertheim et al., 2005; Aiken et al., 2014). It has been reported to be spread among animals and have been shown to cause outbreaks in humans (Mishra et al., 2012). Some of the antimicrobial resistance genes in *S. aureus* are mecA (oxacillin), aac-6/aph-2 (gentamicin), ermA, ermB, *ermC*, *msrA* (erythromycin), *tetK*, *tetM* (tetracycline) and *blaZ* (penicillin) (Duran *et al.*, 2012). Bacterial resistance to antibiotics work in many ways, either by enzyme catalysed deactivation of the drug (Wright, 2011), pumping it out through efflux pump or transport proteins (Webber & Piddock, 2002), or inhibiting its binding to the target e.g RNA polymerase and DNA gyrase. The resistant genes disseminate to the susceptible bacterial strains through horizontal gene transfer e. g acquisition of the *mecA* gene encoding methicillin resistance in *S. aureus* (Lambert, 2005).

1.9 Next generation sequencing technologies

The demand for fast, inexpensive and reliable genomic information lead to the replacement of existing accurate but slow Sanger sequencing method with low cost and high throughput next generation sequencing (NGS) technologies. Genome assemblies, genome resequencing, transcriptomics by RNA-seq, metagenomics and ChIP-seq methods are common applications of these new technologies. These technologies work through template preparation, sequencing and imaging, and data analysis to generate data reads and then multiple sequence alignment of sequence reads for different purposes such as genome assembly, variants analysis. The template preparation and sequencing strategies are specific for each technology. Moreover, the quality scores are also NGS-platform dependent (Metzker, 2010). The commercially available NGS technologies are GS FLX Titanium/GS Junior from Roche/454, Genome Analyzer/HiSeq 2000/MiSeq from Illumina/Solexa, SOLiD/Ion Torrent PGM from Life Sciences, Helicos Biosciences and Pacific Biosciences (Metzker, 2010; Liu *et al.*, 2012). Some of the NGS technologies are described below.

Roche /454 uses emulsion PCR and generates sequence reads of length up to 700 bp with 99.9% accuracy and produces both fragment and paired end libraries. First of all, the genomic DNA sample is sheared into small fragments, whose ends are ligated with adapters. It is followed by the denaturation of the double stranded fragments to obtain single stranded DNA fragments that get annealed to particular beads. These fragment-bead complexes are mixed in emulsion oil and encapsulated in little oil droplets. These encapsulated fragment-bead complexes along with PCR reagents act as microreactors and clonal amplification of each fragment takes place separately inside a separate microreactor producing million of copies for each fragment on each bead.

DNA synthesis is initiated from primer sequence by addition of nucleotides using polymerase. These beads are loaded into the picotiter plate (PTP) that is designed to get one bead per well. Additional beads coupled with sulphurylase and luciferase, are also added into PTP wells. This PTP is loaded into sequencer and incorporation of each base complementary to the template base in detected by signal produced as a result of pyrophosphate release. Pyrosequencing reaction is repeated by the incorporation of another nucleotide and so on. The generated signal is recorded as a series of peaks called flowgram, in which the intensity of peak is consistent with the repetition of single base. Roche's high speed and longer reads length are the prominent advantage over other NGS technologies but major challenges are its high cost, low throughput and error rate for more than 6 polybase (Metzker, 2010; Liu *et al.*, 2012).

Illumina/Solexa uses solid-phase amplification and generates both fragment and mate pair libraries with read length of up to 150 bp. Illumina sequencing begins from the template/sample preparation in which the genomic DNA is extracted and purified. This DNA is fragmented into small molecules and the adaptor sequences are ligated at their ends. The double stranded DNA molecules are denatured to obtain single stranded DNA molecules. Several single stranded DNA molecules are simultaneously hybridized to one of the two types of oligonucleotides of the flow cell channels, which are complementary to the adapter constructs. The complement strand is created from the hybridized fragment using polymerase. The newly synthesized double stranded molecule is denatured and the original template is washed away. The newly synthesized strand clonally amplifies and folds over to bind to the second type of oligonucleotide attached to the flow cell surface, resulting in bridge amplification. A polymerase constructs a complementary strand making a double stranded bridge that is denatured and two copies of molecule are obtained, each of which repeats the same process. The process is repeated again and again simultaneously for million of clusters to produce clonal amplification of all the fragments on the flow cell surface. The reverse strands are cleaved and washed off and only forward strands are read for sequencing. Four fluorescently labelled nucleotides compete for addition to the extending chain of the primer sequence; only one complementary base is added based on the template sequence generating corresponding fluorescent signal in response to a light source. This process is called sequencing by synthesis (SBS). After the incorporation of single nucleotide the process of DNA synthesis is terminated by 3'-reversible terminators in process called cyclic reverse termination (CRT). The 'read length' is determined by the number of cycles for the addition of the nucleotides, the 'base call' is determined by the emission wavelength and the signal intensity. All identical copies of the strands are read simultaneously for a particular cluster. The sequencing of hundreds of millions of clusters occurs simultaneously in a parallel process producing billion of reads that are used in data analysis step such as genome assembly and variant identification. The major advantages of Illumina technology are high throughput and low cost but the shortcomings are short reads length and substitution errors particularly in case when the previous incorporated nucleotide is 'G' (Metzker, 2010; Liu *et al.*, 2012) (https://www.youtube.com/watch?v=womKfikWlxM).

SOLiD: Sequencing by/ Support oligonucleotide ligation detection also produces both the fragment paired and mate pair types of libraries, and could generate data with read lengths of 85 bp with the accuracy of 99.99%. It uses emulsion PCR followed by sequencing by ligation (SBL) and two-base encoding in which each target base is investigated twice. A fluorescently labelled probe sequence is hybridized to the complementary template sequence and ligated to the primer sequence using DNA ligase. After fluorescence scanning, the fluorescent dye is cleaved off using a cleaving agent, and ligation cycle is repeated. The advantage of this NGS technology is high accuracy, but the shortcoming is generation of very short sequence reads (Metzker, 2010; Liu *et al.*, 2012).

Ion Torrent can produce on average 200 bp long data reads with accuracy of 99%. Ion Personal Genome Machine (PGM) launched by Ion Torrent uses semiconductor sequencing technology in which a hydrogen ion or proton is released on incorporation of new nucleotide during DNA synthesis by polymerase. This technology also uses emulsion PCR (Quail *et al.*, 2012). Four nucleotides 'A', 'G', 'C' and 'T' compete on semiconductor chip to incorporate into newly synthesizing DNA strand based on reference strand. PH change or voltage is detected if it is the correct nucleotide; no voltage is detected if it is wrong nucleotide; and double voltage is found if two copies of same nucleotide are added. Unlike other NGS tehcnologies, Ion Torrent does not require fluorescence and camera scanning, therefore is fast, small in size and easily affordable by small labs (Metzker, 2010; Liu *et al.*, 2012).

All above NGS technologies are based on the clonal amplification methods that use large amount of genomic DNA in 3 to 20 μg, however few NGS platforms such as **Helicos Biosciences** and **Pacific Biosciences** use non-amplified single molecule template and require less than 1 μg starting DNA material. These platforms do not require PCR therefore the sequencing errors due to mutations and amplification bias are avoided. In these technologies, the single molecule templates are immobilized on solid support before initiating NGS reaction. In Helicos Biosciences either spatially distributed primer sequences or the adaptors-ligated template fragments are immobilized, followed by NGS reaction by DNA polymerase. In Pacific Biosciences spatially distributed DNA polymerase molecules are immobilized by attaching them to the solid surface, and the primed DNA molecule of tens of thousands bp long is bound to the polymerase, generating longer sequence reads (Metzker, 2010). Although PacBio produces relatively lower throughput than second-generation sequencers, is quite fast and produces nearly 1300 bp long sequence reads (Liu *et al.*, 2012).

1.10 Genome assembly

The short sequence reads generated by NGS platforms are assembled through genome assembly process. A genome assembly produces a set of contigs, each one of that is the multiple sequence alignment of reads (Dear *et al.*, 1998), these set of contigs are then ordered, oriented and joined to make scaffolds (Huson *et al.*, 2002). There are two common methods for genome assembly, *de novo* assembly and mapping assembly. In *de novo* assembly, the sequence reads are assembled on the basis of overlapping reads generating new unknown sequence in the form of contigs or short scaffolds. Whereas in mapping assembly these sequence reads are assembled using a backbone reference sequence generating a consensus sequence similar to the reference sequence but not principally identical (Nishito *et al.*, 2010).

Genome assembly algorithms follow three different strategies for assembly process; the Overlap-Layout-Consensus (OLC) strategy, the de Bruijn Graph (DBG) strategy, and the greedy graph strategy. These are based on graphs that are set of nodes/vertices and the set of edges/arcs connecting these nodes. The nodes represent reads, the edges represent the overlaps between reads and the set of directed edges represent paths. The OLC method uses an overlap graph that is based on reads and their overlaps, the DBG method uses K-mer graph that is based on overlaps of fixed-length, and the greedy graph method may use OLC or DBG, and is based on the greedy extension process of adding more reads or more contigs to any given read or contig taking into account the highest scoring overlap (Miller *et al.*, 2010). The OLC method uses three steps, the *Overlap* in which potential overlap regions are identified among reads, the *Layout* in which the multiple selected reads are aligned based on their overlaps, and the *Consensus* in which aligned reads generate a final sequence estimate. In mapping assembly, the *Overlap* step is replaced by an *Align* step in which reads are aligned relative to the reference genome (Peltola *et al.*, 1984; Huang, 1992; Pop *et al.*, 2004).

1.11 Hypothesis

The hypothesis of this thesis is that *S. agalactiae* play a role in pathogenesis and cause infections of various tissues in camels such as skin abscesses, infection of joints or peri-arthricular abscesses, and infection of udder or mastitis. This pathogen is also the cause of pathogenesis in other animals like cattle and humans, such as mastitis in cattle and neonatal infections in humans. The pathogenicity of *S. agalactiae* in different infections of different hosts is due to certain virulence factors, and the differences in host specificity/adaptation between various *S. agalactiae* isolates are due to the acquisition of new genes or the loss of existing genes. Our second hypothesis is that the *S. aureus* isolate of Strain type 30 from camels is relatively distinct from Clonal complex 30 *S. aureus* isolates from humans.

2 Aims of this thesis

The **basic aim** of this research thesis was to understand the molecular biology of zoonotic pathogens in camels, by screening *S. agalactiae*'s underlying potential pathogenicity factors that could be involved in introducing mastitis and skin infections in camels, and to understand the mechanism of *S. agalactiae*'s adaptation and pathogenicity from one host to the other. Moreover, we aimed to analyse the genetic heterogeneity of ST30 *S. aureus* isolate from camel compared with CC30 *S. aureus* isolates of human origin, and identify the candidate factors that could be responsible for *S. aureus* host tropism in camels.

The **specific aims** of this study were as below;

- To assemble new genome sequences of *S. agalactiae* from camels and cattle, using NGS data; and annotate them.
- To compare newly sequenced *S. agalactiae* genomes with previously sequenced *S. agalactiae* genomes from cattle and humans to investigate the genetic heterogeneity and diversity of *S. agalactiae* across the strains in multiple hosts.
- To identify potential virulence genes that could be used as specific markers for *S. agalactiae*.
- To assemble and annotate new *S. aureus* genome sequence of type ST30 from camel, and perform comparative and phylogenetic analysis with all CC30 *S. aureus* genome sequences from humans to investigate its genetic heterogeneity.

3 Introduction (Paper I-IV)

Dromedary camels, being a valuable livestock species in providing a good source of food such as milk and meat, and transport for the pastoralists of semiarid and arid regions of the Greater Horn of Africa, are of great economic importance for the livelihood of these inhabitants (Kagunyu & Wanjohi, 2014). However, bacterial pathogens like *S. agalactiae* and *S. aureus* are enormously deteriorating the health of these camels by introducing different kinds of infections; *S. agalactiae* causes mastitis and skin infections, and *S. aureus* causes mastitis, bacteraemia, respiratory infections and wound infections (Ladhani, 2004; Guinane *et al.*, 2010; Fitzgerald, 2012; Maina *et al.*, 2013). These infections not only lead to economic losses by declining milk and meat productivity in camels, but the zoonotic transmission of these pathogens also affects the health of human themselves (Christou, 2011; Petersen *et al.*, 2013). The consumption of raw milk increases the risk of acquiring infections with zoonotic pathogens (Sprague *et al.*, 2012; Gautret *et al.*, 2013). Although camels are of great economic significance for the Horn of Africa, research on camels and their pathogens is lacking.

Until now, no genome sequences of bacterial pathogens affecting camels were available. Detailed sequence analysis of *S. agalactiae* and *S. aureus* from camels was an essential first step in exploring their molecular basis for host-specificity and pathogenesis in camels. We reported the assembly and annotation of the two first published genomic sequences of *S. agalactiae* isolates from abscesses in dromedary camels (Paper **I**) and the first published genome sequence of *S. aureus* isolate from the nasal swab of dromedary camel (Paper **IV**), from Kenya. *S. agalactiae* not only affects camels but causes mastitis in dairy cattle, and neonatal infections in humans. A total of eight *S. agalactiae* genome sequences from humans have already been sequenced, however only a single *S. agalactiae* genome from cattle has been sequenced and is in draft or un-finished status. We sequenced and annotated the first complete genome sequence of *S. agalactiae* isolate from cattle with subclinical mastitis, from Sweden (Paper **II**). The work performed in paper **I** and **II** allowed the detailed comparative analysis of *S. agalactiae* assembled genomes in paper **III**. In this paper, the comparative analysis of *S. agalactiae* genome sequences from three different hosts, camels, cattle and humans were compared to explore the genetic variability of these pathogens based on different hosts. A total of seven GBS isolates were used in comparison, two isolates ILRI005 and ILRI112 from infection in camels from Kenya described in paper I, one isolate 09mas018883 from mastitis in cattle from Sweden described in paper II and one published isolate FSL-S3-026 from mastitis in cattle from USA, two published isolates A909, 2603V/R from infection in neonates from USA, and another published isolate NEM316 from neonatal infection from France. We investigated many important virulence loci in these seven GBS isolates. Potential virulence loci were found to be present in GBS isolates potentially causing pathogenicity in hosts, however number of genes were variable in these loci. This variation or gain/loss of genes are probably of adaptive nature from one host to the other. Similarly the resistance gene *tetM* is also found to be present in some isolates. Paper **IV** provides first complete genome sequence and annotation of ST30 *S. aureus* isolate from camel, and its comparative analysis with CC30 *S. aureus* isolates from humans, to investigate the genetic diversity of camel isolate compared to human isolates.

The availability of these new genome sequences and their detailed comparative analysis aided us to identify virulence candidates e.g *CRISPR2* locus in *S. agalactiae* and putative phage insertions in *S. aureus*, potentially responsible for pathogenicity in their hosts. Our research provides novel insights on core genome, shared genome and isolate-specific genome content that could be relevant for developing control measures for *S. agalactiae* and *S. aureus* infections in camels. A deeper understanding of the identified virulence factors would ensure the growth, health and productivity of camels as well as human health and income in these developing countries. Moreover, it would be important to contemplate how the transfer of virulence and resistance genes has occurred in GBS isolates from different regions of the globe.

4 Materials and methods

4.1 Isolation of strains and DNA extraction

Streptococcus agalactiae isolate ILRI005 was isolated from an abscessed lesion of a *Camelus dromedarius* in Isiolo, Kenya, and ILRI112 was obtained from a periarthricular lesion of a *Camelus dromedarius*, in Laikipia Kenya. *Streptococcus agalactiae* isolate 09mas018883 was isolated from milk obtained from a single cow (*Bos taurus*) in Uppsala, Sweden, that was diagnosed as having subclinical mastitis case by SCC. DNA extraction was performed with standard phenol/chloroform extraction at the place of their isolation (Paper **I**, **II**). *Staphylococcus aureus* isolate ILRI_Eymole1/1 was isolated in Kenya from the nasal swab of a *Camelus dromedarius* that had rhinitis symptoms. DNA was isolated using the PureLink™ Genomic DNA Mini Kit (Invitrogen) according to manufacturer's instructions (Paper **IV**).

4.2 Genome Sequencing by NGS

Two *S. agalactiae* isolates ILRI005 and 09mas018883, and one *S. aureus* isolate ILRI_Eymole1/1 were sequenced using Illumina Genome Analyser GAIIx. Paired-end libraries were generated for these three isolates. Only one of the isolates, the *S. agalactiae* ILRI112 was sequenced with Ion Torrent, from a single end library. In a single end library, the genomic template is sequenced only from one end to generate single end sequence reads, while in paired end library the genomic template is sequenced from both ends, producing paired end sequence reads (Margulies *et al.*, 2005), (http://res.illumina.com/documents/products/datasheets/datasheet_genomic_sequence.pdf). The details of the NGS data used for four isolates are specified in Table 1 (Paper **I**, **II**, **IV**).

Table 1. *Data description of four isolates used in this study*

Bacteria	*Streptococcus agalactiae*			*Staphylococcus aureus*
Isolate	ILRI005	ILRI112	09mas018883	ILRI_Eymole1/1
Host	Camel	Camel	Cattle	Camel
NGS technology	Illumina	Ion Torrent	Illumina	Illumina
Avg. Read Length	100 bp	200 bp	75 bp	300 bp
Library	Paired-end	Single-end	Paired-end	Paired-end
Avg. Insert size	210 bp	N/A	545 bp	550 bp

4.3 Comparative genome assembly and genome finishing

The schematic representation of genome assembly process followed for all four bacterial genomes assembled in this study is shown in Figure 4. Shotgun sequence reads were assembled using two assembly methods.

1. A ***de novo*** assembly that was independent of a reference sequence.
2. A **mapping** or **reference-guided** genome assembly that mapped reads onto the chosen reference sequence (09mas018883 was mapped to A909 as reference, ILRI005 to 09mas018883 as reference, and ILRI112 to ILRI005 as reference).

Reference-guided assembly can identify variations among closely related prokaryotic and eukaryotic genomes but cannot expose species-specific sequences (Nishito *et al.*, 2010). It does not reveal divergent sequences like chromosomal rearrangements, large insertions and deletions due to their high levels of divergence from the reference sequence (Zubair, 2010), so combining both assemblies is a better approach to accurately identify regions similar to the reference genome as well as different from it.

Multiple appropriate tools were used to carry out the genome assembly of four different isolates. Both mapping and *de novo* assembly of the cattle *S. agalactiae* isolate 09mas018883 and a camel *S. agalactiae* isolate ILRI005 was performed using MIRA v 3.0 (Chevreux *et al.*, 1999). The mapping assembly of a camel *S. agalactiae* isolate ILRI112 was carried out using MIRA v 3.4.1.1 (Chevreux *et al.*, 1999), while its *de novo* assembly was done using Newbler v 2.8 (Margulies *et al.*, 2005). Reference genomes for the assembly process were selected on the basis of the alignment of the maximum percentage of the input data reads. The *S. agalactiae* genome A909 was used to perform the reference-guided assembly of 09mas018883 data reads, based on its maximum alignment of 92.2% reads compared to other previously sequenced GBS genomes. After getting complete sequence of 09mas018883, it was used as a reference genome for the mapping assembly of camel *S. agalactiae* ILRI005 data reads.

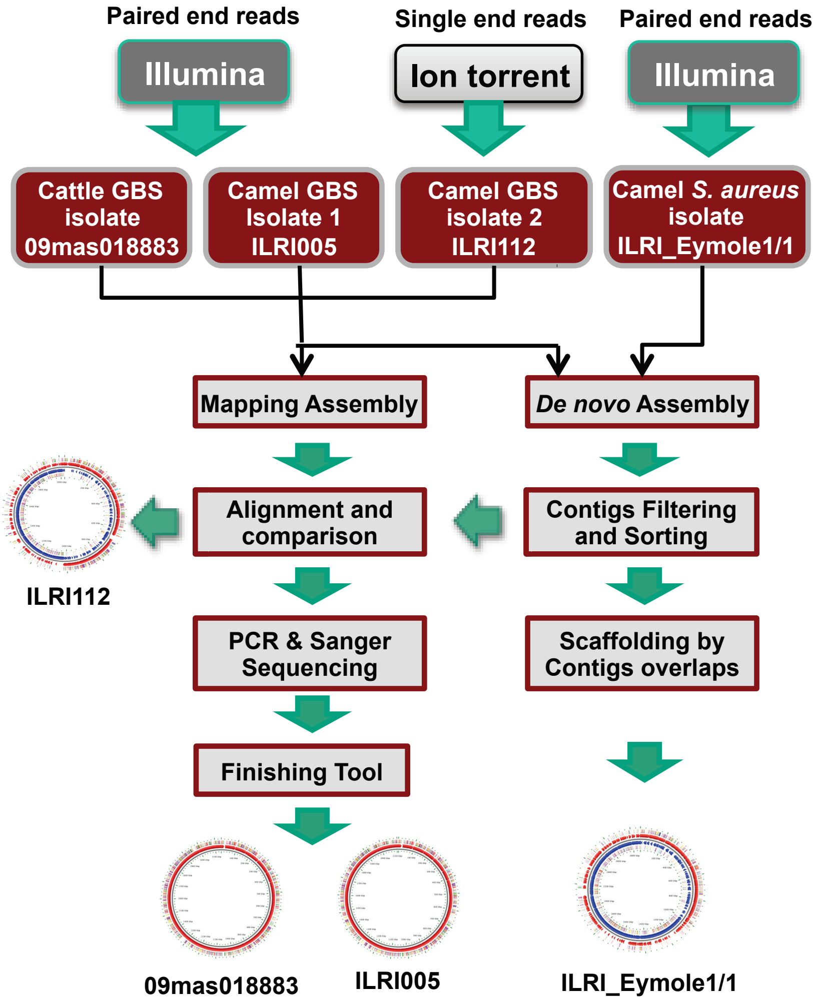
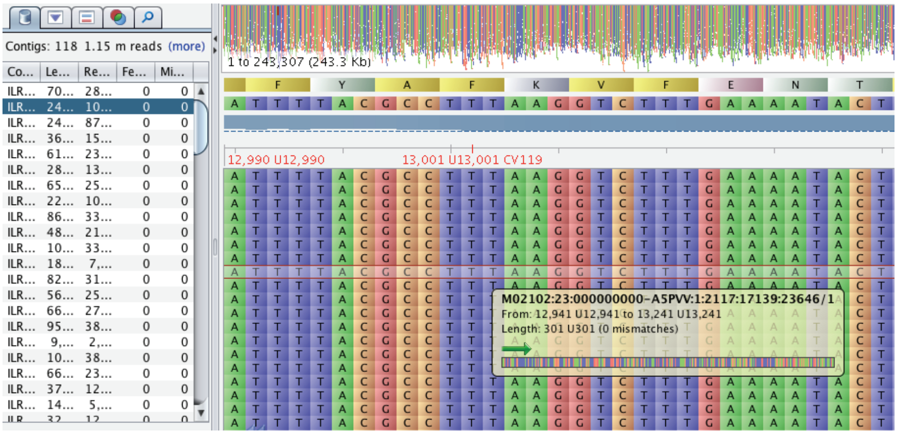


Figure 4. Comparative assembly process used to assemble four complete genomes.

ILRI005 complete genome sequence was further used as a reference sequence for the reads mapping of camel *S. agalactiae* ILRI112 isolate (Paper **I**, **II**). In addition, comparative assembly approach was included to combine both mapping and *de novo* assemblies (Nishito *et al.*, 2010). The *de novo* assembled contigs were filtered by discarding contigs with less than 10X coverage and 1000 bp length, and were then sorted against the reference genome sequence using ABACAS perl script (Assefa *et al.*, 2009), and alignment tool MUMmer v 3.2.2 (Kurtz *et al.*, 2004). The consensus sequence from the mapping assembly was aligned against the sorted *de novo* contigs using Mauve, a whole genome alignment tool (Darling *et al.*, 2004). The regions where mapping assembly showed different result than *de novo* assembly, and gapped regions were further analysed. The combined results of both assemblies, together with regular PCR, long range PCR, Sanger sequencing, a finishing tool GapFiller (Boetzer & Pirovano, 2012) and additional *de novo* assembly by Velvet assembler (Zerbino & Birney, 2008), finally produced two complete genome sequences for isolates 09mas018883 and ILRI005 (Zubair *et al.*, 2013a; b) (Paper **I**, **II**). The *de novo* assembly of a camel *S. aureus* isolate ILRI_Eymole1/1 was done using MIRA v 4.0 (Chevreux *et al.*, 1999), contigs were sorted according to a reference *S. aureus* genome MRSA252 (Holden *et al.*, 2004) using MUMmer v 3.2.2 (Kurtz *et al.*, 2004) and Mauve alignment (Darling *et al.*, 2004), and were concatenated on the basis of overlaps between contigs to reach a single scaffold (Paper **IV**).

4.4 Sequence Visualization methods

The assembly output (ACE) files produced by the assemblers were viewed in Tablet version 1.10.03.04 (Paper **I, II**) and 1.13.05.17 (Paper **IV**), a memory efficient assembly viewer tool for NGS technologies (Milne *et al.*, 2013). An example of genome assembly view for *S. aureus* isolate ILRI_Eymole1/1 is shown in Figure 5.



Contigs: 118 1.15 m reads (more)

Co...	Le...	Re...	Fe...	Mi...
ILR...	70...	28...	0	0
ILR...	24...	10...	0	0
ILR...	24...	87...	0	0
ILR...	36...	15...	0	0
ILR...	61...	23...	0	0
ILR...	28...	13...	0	0
ILR...	65...	25...	0	0
ILR...	22...	10...	0	0
ILR...	86...	33...	0	0
ILR...	48...	21...	0	0
ILR...	10...	33...	0	0
ILR...	18...	7,...	0	0
ILR...	82...	31...	0	0
ILR...	56...	25...	0	0
ILR...	66...	27...	0	0
ILR...	95...	38...	0	0
ILR...	9,...	2,...	0	0
ILR...	10...	38...	0	0
ILR...	66...	23...	0	0
ILR...	37...	12...	0	0
ILR...	14...	5,...	0	0
ILR	32	12	0	0

1 to 243,307 (243.3 Kb)

F Y A F K V F E N T

A T T T T A C G C C T T T A A G G T C T T T G A A A A T A C T

12,990 U12,990 13,001 U13,001 CV119

M02102:23:000000000-A5PVV:1:2117:17139:23646/1
From: 12,941 U12,941 to 13,241 U13,241
Length: 301 U301 (0 mismatches)

Figure 5. Visualization of genome assembly file (*.ace) loaded into Tablet viewer. Left pane is showing the list of contigs in *de novo* assembly. Right top pane is showing the coverage view of the selected contig, and right bottom pane is indicating the sequence reads aligned to each other with certain coverage. The label, length and the direction of the selected read are highlighted in yellow box. The consensus sequence is also indicated between coverage view (top), and reads view (bottom).

4.5 Genome annotation

Genome annotation is a multi-step process of interpreting a raw DNA sequence to understand its biological significance. It comprises of three different steps; a nucleotide-level annotation *Where*, a protein-level annotation *What?* and a process-level annotation *How?* (Stein, 2001). Specific tools used for the specific steps in the annotation process of the *S. agalactiae* and *S. aureus* genomes, are named in Table 2 (Paper **I-IV**).

Table 2. *Various tools used for different kinds of annotation.*

Annotation type	Servers/Tools used	Reference
Whole genome protein prediction	RAST, Basys, Mage, Sybil, PATRIC, NCBI FTP site	(Aziz *et al.*, 2008), (Van Domselaar *et al.*, 2005), (Vallenet *et al.*, 2006), (Riley *et al.*, 2012), (Wattam *et al.*, 2014), (https://www.ncbi.nlm.nih.gov/Ftp/)
Annotation analyser/viewer	Artemis	(Carver *et al.*, 2012)
rRNA prediction	RNAmmer v 1.2	(Lagesen *et al.*, 2007)
tRNA prediction	tRNAscan-SE v 1.21	(Lowe & Eddy, 1997)
CRISPR identification	CRISPRFinder	(Grissa *et al.*, 2007)
Phages identification	PHAST	(Zhou *et al.*, 2011)
Genomic Islands prediction	IslandViewer	(Langille & Brinkman, 2009)
Identification of Lateral/horizontal transfers (LGT)	GOHTAM	(Ménigaud *et al.*, 2012)
Signal peptides prediction	SignalP v 4.1	(Petersen *et al.*, 2011)
Transmembrane helices prediction	TopPred2	(Heijne, 1992)

(Paper **I-IV**).

4.6 Comparative analysis of *S. agalactiae* isolates (Paper III)

Seven *S. agalactiae* sequenced genomes were used in the comparative analysis to explore the genetic similarities and differences of these genome sequences. Among these GBS genomes were two newly sequenced *S. agalactiae* camel isolates ILRI005 and ILRI112, one newly sequenced *S. agalactiae* cattle isolate 09mas018883, one previously sequenced *S. agalactiae* cattle isolate FSL-S3-026, and three previously sequenced *S. agalactiae* human isolates A909, NEM316 and 2603V/R.

These seven *S. agalactiae* genomes were compared at two levels; *sequence-level comparison*, and *annotation-level comparison*. The ***sequence-level comparison*** was performed through pairwise alignment of the genomes using

MUMmer v 3.2.2 (Kurtz *et al.*, 2004), analysing genomic architecture and genomic rearrangements among genomes using Mauve tool (Darling *et al.*, 2004), visualizing genome synteny using Sybil server (Riley *et al.*, 2012), finding average nucleotide identity (ANI) using Jspecies v 1.2.1 (Richter & Rossello, 2009) and generating genome identity plots using BRIG (Alikhan *et al.*, 2011). The ***annotation-level comparison*** among seven *S. agalactiae* genomes was carried out by comparing general genomic features among them such as number of predicted CDS, rRNA genes, tRNA genes; doing pan proteome analysis using protein blast searches and custom scripts to find the common, variable and isolate-specific protein encoding genes among these *S. agalactiae* genomes; performing COG classification of core genes; and identifying and comparing potentially virulent features shared by either all *S. agalactiae* isolates or shared by some of the *S. agalactiae* isolates. Phylogenetic relationship among seven *S. agalactiae* genomes was inferred by phylogeny based on their core genome content identified through whole genome alignment as well as based on conserved core genes. Mugsy aligner (Angiuoli & Salzberg, 2011) was used for multiple sequence alignment, Phylomark tool (Sahl *et al.*, 2012) was used for concatenation of sequences and the phylogenetic trees were constructed using MEGA v 6.06 (Tamura *et al.*, 2013).

4.7 Comparative analysis of *S. aureus* isolates (Paper IV)

Newly sequenced *S. aureus* isolate ILRI_Eymole1/1 was compared with twenty previously sequenced CC30 type *S. aureus* isolates from humans. Their sequence types were analysed using MLST database (Enright *et al.*, 2000). The protein encoding genes common (core genes) in all CC30 *S. aureus* isolates were extracted using protein blast searches and custom Perl scripts (Supplementary data, paper **IV**). Functional classification of the core genes was carried out by protein blast search against a collection of genes in COG (Clustering of Orthologous Groups) database (Tatusov *et al.*, 2000). Genes shared between *S. aureus* ILRI_Eymole1/1 and several type CC30 *S. aureus* isolates (variable genes) were identified. ILRI_Eymole1/1's genes not found in other CC30 *S. aureus* isolates (isolate-specific genes) were also identified. A refined set of core genes were extracted and used to determine the phylogenetic relationships of novel *S. aureus* isolate with previously sequenced *S. aureus* isolates of type CC30. Two non-CC30 *S. aureus* isolates were used as an outgroup in phylogenetic tree construction. Multiple sequence alignment was carried out using Mugsy aligner (Angiuoli & Salzberg, 2011), and the phylogeny was performed using PhyML v 3.0 (Guindon & Gascuel, 2003).

5 Summary of Results with brief Discussion

5.1 Assembled genomes (Paper I, II and IV)

Comparing the results of *de novo* assembly with the mapping assembly for *S. agalactiae* ILRI005, 25 of 142 *de novo* contigs appeared as unaligned or orphan contigs. The orphan contigs of the ILRI005 *S. agalactiae* genome consisted of phage-related sequences and were the most difficult to assemble. All 25 orphan contigs were ultimately incorporated into the camel *S. agalactiae* ILRI005 genome, by combining the results of mapping assembly, *de novo* assembly, regular PCR, long range PCR, and Sanger sequencing. ILRI005 complete genome sequence acted as a good reference for the mapping assembly of a second GBS isolate from camel, the combined results of mapping and *de novo* assembly being sufficient to assemble a complete ILRI112 genome sequence. In case of cattle *S. agalactiae* 09mas018883 there was only one unaligned contig containing the tetracycline resistance gene *tetM*. This orphan contig was assembled through PCR and Sanger sequencing between the flanking ends of an orphan contig and the final gap; and an additional *de novo* assembly by Velvet assembler (Zerbino & Birney, 2008). The comparative assembly approach was useful in solving the problem of unaligned contigs, gap closure, and the identification of genomic regions for which the assembly differed between mapping and *de novo* approaches. Genomic regions where both assemblies were concordant and the coverage was good were incorporated in the final assembly. Mapping assembly bridged the gaps between two consecutive *de novo* contigs; likewise *de novo* contigs filled the gaps in mapping assembly. The regions where both assemblies had sequence but of different length were analysed further by PCR and Sanger sequencing to verify the results of *de novo* assembly (Paper **I, II**). In case of *S. aureus* ILRI_Eymole1/1, the mapping assembly was not successful due to the high number of chromosomal rearrangements in ILRI_Eymole1/1 compared to the reference genomes *S. aureus* MRSA252, TCH60 and 55/2053. *De novo* assembly followed by sorting of contigs based on overlaps according to a reference genome was an appropriate strategy in this case (Paper **IV**).

Finally we assembled each genome in the form of a single circular chromosome. The genome sizes of newly assembled *S. agalactiae* isolates were similar to those of the published *S. agalactiae* genomes from humans (Table 2, paper **III**). Likewise, the genome size of newly assembled *S. aureus* isolate was similar to that of published *S. aureus* isolates (Table 5, paper **IV**). Camel *S. agalactiae* isolates ILRI005 and ILRI112 had genome size of ~2.11 and ~2.03 Mbp respectively whereas the cattle *S. agalactiae* isolate 09mas018883 had genome size of ~2.14 Mbp. Camel *S. aureus* isolate ILRI_Eymole1/1 had genome size of ~2.87 Mbp. The assembly statistics of the four assembled genomes is given in Table 3.

Table 3. *Assembly statistics for four isolates*

Bacteria	*Streptococcus agalactiae*			*Staphylococcus aureus*
Isolate	ILRI005	ILRI112	09mas018883	ILRI_Eymole1/1
Host	Camel	Camel	Cattle	Camel
Filtered *de novo* contigs	142	43	43	69
Total filtered reads	20,687,942	3,123,413	10,079,600	1,176,591
Reads assembled	20,189,204 (97.6%)	2,994,027 (96%)	10,035,130 (99.6%)	1,154,246 (98.1%)
Average consensus coverage	936X	224X	351X	109X
Average consensus quality	79	75	87	83
Genome size (bp)	2,109,759	2,029,198	2,138,694	2,874,302
Reference genome used	09mas018883	ILRI005	A909	MRSA252

5.2 Comparative analysis of *S. agalactiae* isolates (Paper III)

In this study we determined genetic similarities, differences and phylogenetic relationship of *S. agalactiae* isolates from camels, cattle and humans using comparative genomics. Until now, genome sequences of only two *S. agalactiae* isolates from camels *i.e* ILRI005 and ILRI112 are available (**Paper I,** 2013) and only single completely sequenced *S. agalactiae* isolate from cattle 09mas018883 is available (**Paper II,** 2013). There is a previously published draft genome sequence of *S. agalactiae* isolate FSL-S3-026 from cattle (Richards *et al.*, 2011), therefore we included it in our analysis. A total of eight *S. agalactiae* genomes of human origin have been sequenced previously, however we focused our analysis on the three complete genome sequences A909, NEM316 and 2603V/R (Glaser *et al.*, 2002; Tettelin *et al.*, 2002, 2005), in order to use comparable number of genomes from each group. In total we used seven *S. agalactiae* isolates in this comparative study.

5.2.1 General genomic features of seven isolates

The comparison of general genomic features of *S. agalactiae* isolates used in study is shown in Figure 6. In general, the genomic features were similar in all GBS isolates, however GBS isolates from camels had relatively less genomic size and GC%. The deviation of an unfinished GBS genome sequence FSL-S3-026 is prominent in all features.

5.2.2 Taxonomic relationship

Average nucleotide identity (ANI) is the measurement of pairwise comparison of the genome sequences taking into account the tetra nucleotide signature frequencies, and is used for the taxonomic classification of prokaryotes. The optimal threshold ANI value of > 94% ANI between two genomes specifies them as of the same bacterial species. This value corresponds to DNA-DNA Hybridization (DDH) recommended cut-off value of 70% for species classification. DDH has been used to establish the relatedness of bacterial strains and species delineation, however can be substituted by ANI (Goris *et al.*, 2007; Richter & Rossello, 2009). We used these ANI values to estimate relative closeness of all seven *S. agalactiae* isolates among each other. The seven *S. agalactiae* isolates used in this study exhibited an ANI value of greater than 99%, strongly suggesting these isolates are closely related. Two camel *S. agalactiae* isolates ILRI005 and ILRI112 showed highest ANI between each other hence were closely related; three human *S. agalactiae* isolates showed highest ANI with cattle *S. agalactiae* 09mas018883 whereas the cattle *S. agalactiae* isolate FSL-S3-036 was less similar to each of the other isolate.

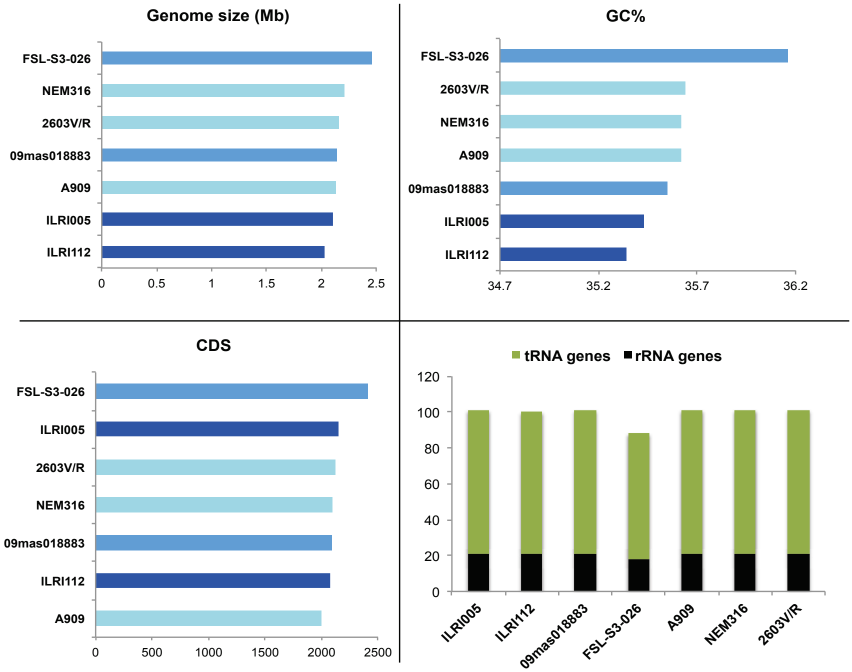


Figure 6. General genomic features of seven *S. agalactiae* isolates. Genome size in Mbp, GC% and number of protein encoding genes (CDS) are represented in ascending order from bottom to top. *S. agalactiae* from the same host type are depicted in same shade of blue color. Noncoding tRNA and rRNA genes of all *S. agalactiae* are shown in green and black color respectively.

5.2.3 Gene synteny

Gene synteny is the way to estimate the conservation of genomic placement of shared genes of a query genome relative to the reference genome. Gene synteny among seven *S. agalactiae* isolates was estimated using Sybil server (Riley *et al.*, 2012). Synteny gradient display or the visual representation of the arrangement of a query genomes with respect to the reference genome is shown in Figure 2C, paper **III**. The results revealed that the order of genes in *S. agalactiae* isolates is highly conserved, however FSL-S3-026 showed difference in synteny gradient probably due to the availability of this genomic sequence in the form of eight contigs or unfinished genome. All other genomes used in the study were complete. Two of the camel GBS isolates ILRI112 and ILRI005 were relatively far in synteny to reference genome '09mas018883' than human GBS isolates A909, 2603V/R and NEM316 were to the reference genome.

5.2.4 Phylogenetic relationship

We determined phylogenetic relationship among seven *S. agalactiae* isolates in two different ways. Firstly, we performed multiple sequence alignment of whole genomes using Mugsy tool specialized for whole genome alignment of closely related species (Angiuoli & Salzberg, 2011). From this alignment, we extracted and concatenated the conserved blocks with phylogenetic markers using Phylomark tool (Sahl *et al.*, 2012), and the phylogeny established. The phylogenetic tree is shown in Figure 3A, paper **III**. Secondly, we identified conserved genes on the basis of an all-against-all blastn comparison of the gene sequences of seven *S. agalactiae* isolates. We concatenated sequences of conserved genes and a phylogenetic tree was constructed, shown in Figure 3B, paper **III**. The phylogenetic trees based on both sequences (whole genome conserved content, and conserved genes) presented the same result. Both camel *S. agalactiae* isolates were relatively distant to human and cattle *S. agalactiae* isolates while being relatively close to each other.

5.2.5 Core, shared and isolate-specific genes

We identified 'core genes' as the genes common in all seven *S. agalactiae* isolates, 'shared or variable genes' the genes common in some of the *S. agalactiae* isolates, and 'isolate-specific genes' as the genes present in only one *S. agalactiae* isolate. The distribution of core, variable and isolate-specific genes in each *S. agalactiae* isolate is depicted in Figure 7.

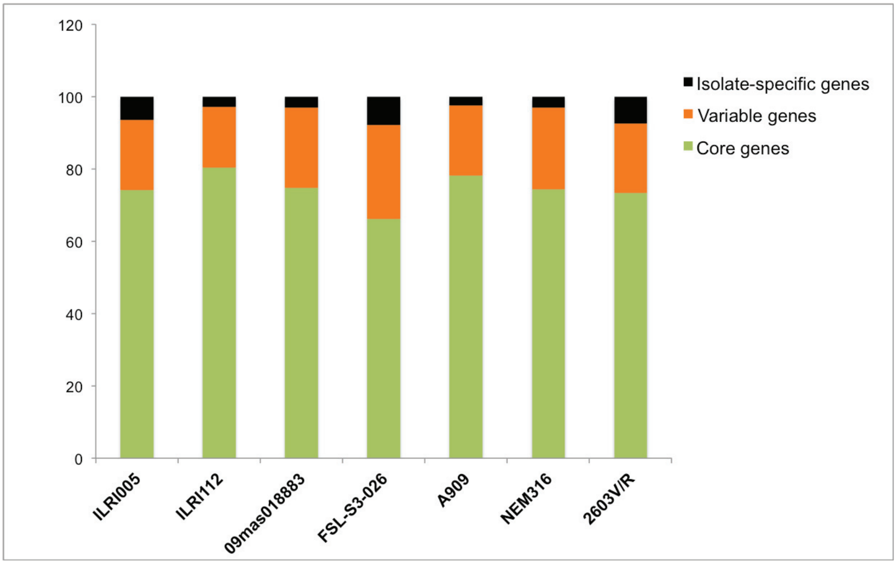


Figure 7. The distribution of core, variable and isolate-specific genes of seven *S. agalactiae* genomes (Paper III)

5.2.6 Isolate-specific genes in camel *S. agalactiae*

We identified isolate-specific genes in seven GBS isolates, but focused our analysis on isolate-specific genes of the newly sequenced genomes of camel GBS isolates, ILRI005 and ILRI112. We identified lateral gene transfer (LGT) signatures, genomic islands and putative phage insertions for these isolates. We mapped the positions of LGT signatures, genomic islands, putative phage insertions and isolate specific genes along the circular plot, in order to evaluate the genomic positions and relationship of isolate-specific genes relative to these features known to be under the influence of horizontal gene transfer. We observed that ~64% of isolate-specific genes in camel *S. agalactiae* ILRI005 were mapped at the genomic positions of putative phage insertion sequences. Approximately 74% of the isolate-specific genes in ILRI112 were clustered in genomic islands. The isolate-specific genes in both camel *S. agalactiae* isolates ILRI005 and ILRI112 were either clustered with in the putative phage insertions, genomic islands and LGT signatures separately or with two to three of them simultaneously, however a few exceptions were observed. The high proportions of isolate-specific genes in these areas suggest their acquisition via lateral transfer events in camel GBS isolates.

5.2.7 Tetracycline resistance gene *tetM* and associated transposon Tn*916*

Tetracycline resistance gene *tetM* and its associated transposon Tn*916*, is of significance in developing antibiotic resistance in microbial communities (Roberts & Mullany, 2011), and was found present in three of the *S. agalactiae* isolates; the human isolate 2603V/R, the cattle isolate 09mas018883 and the camel isolate ILRI112. It was found lacking in all other sequenced *S. agalactiae* isolates. A study conducted on Kenyan camels in 2013 reported the presence of transposon Tn*916* in all camel *S. agalactiae* isolates of resistance to tetracycline. And 34% of total GBS isolates were tetracycline resistant and possessed the *tetM* gene (Fischer *et al.*, 2013). It indicates the frequent use of tetracycline as antimicrobial treatment of GBS infections in Kenyan camels. Domestication of camels and other community ruminants might also have contributed to the increased rate of transfer of this resistance gene among GBS populations. An appropriate strategy would be to eliminate the use of tetracycline in camels, cattle and other animals in these regions, and to treat GBS infections with some alternate antibiotic.

S. agalactiae resistance for tetracycline and many other antibiotics has been reported. A study performed on antimicrobial susceptibility testing of GBS isolates from cases of cattle with subclinical mastitis indicated GBS resistance to streptomycin (85.1%), followed by tetracycline (55.5%), erythromycin (33.3%), cotrimoxazole (11.1%), ampicillin (11.1%), enrofloxacin (7.4%) and gentamicin (3.7%) (Jain *et al.*, 2012). Likewise more than 80% of GBS isolates showed resistance to tetracycline (Poyart *et al.*, 2003; Nakamura *et al.*, 2011). GBS were also found resistant to erythromycin, clindamycin, and levofloxacin (Borchardt *et al.*, 2006; Nakamura *et al.*, 2011). However, GBS were found susceptible to penicillin, vancomycin (Liddy & Holliman, 2002; Borchardt *et al.*, 2006; Nakamura *et al.*, 2011), ceftazidime (Nakamura *et al.*, 2011), cefotaxime, teicoplanin and rifampin (Poyart *et al.*, 2003). We should prefer one of these antibiotics or any other known susceptible antibiotics for treatment against infections in camels and other community animals. Overuse and misuse of antibiotics should also be avoided. Further the investment on management, nutrition and hygiene of camels and cattle is also important to reduce the need for antibiotics.

5.2.8 CRISPR/Cas system

Clustered regularly interspaced short palindromic repeats (CRISPRs) and CRISPR-associated Cas proteins in 40% bacteria and 90% archaea together make a well defined CRISPR/Cas system (Horvath & Barrangou, 2010) that generally undergoes two phases, the adaptation phase and the interference phase. In the adaptation phase, new spacer sequences are acquired from external DNA while in the interference phase these acquired spacers are used as antiviral defence mechanism to cleave the foreign invasive DNA (Deveau *et al.*, 2010).

5.2.9 *CRISPR1* locus in all *S. agalactiae* isolates

The *cas* genes in *CRISPR1* locus were identified in the core genome of all seven *S. agalactiae* isolates. At 5’ end of the *CRISPR* locus are four *cas* genes *Csn1*, *Cas1*, *Cas2* and *Csn2*, while at 3’ end there are the spacers and repeats. The spacers and repeats exist in non-coding sequence of the *S. agalactiae* genomes. The length and sequence of repeats was fixed in *CRISPR1* locus of all *S. agalactiae* isolates, however the sequence of the last repeat has a few base pairs variation. The repeats in this locus were 36 bp long with sequence ‘*GTTTTAGAGCTGTGCTGTTTCGAATGGTTCCAAAAC*’. The length of spacer was fixed at 30 bp while the sequence was variable. The number of repeats and spacers in *CRISPR1* locus were variable from one isolate to the other. The simplistic representation of *CRISPR1* locus is given below;

5’__[*Csn1* - *Cas1* - *Cas2* - *Csn2*] - [Repeats and Spacers]**__3’**

ILRI005	4 X [Repeat of 36 bp] 3 X [Spacer of 30 bp]

ILRI112	13 X [Repeat of 36 bp]
	12 X [Spacer of 30 bp]
09mas018883	11 X [Repeat of 36 bp]
	10 X [Spacer of 30 bp]
A909	15 X [Repeat of 36 bp]
	14 X [Spacer of 30 bp]
2603V/R	25 X [Repeat of 36 bp]
	24 X [Spacer of 30 bp]
NEM316	14 X [Repeat of 36 bp]
	13 X [Spacer of 30 bp]
FSL-S3-026	15 X [Repeat of 36 bp]
	14 X [Spacer of 30 bp]

The function of *cas* genes is not known, however *Cas1* and *Cas2* genes are considered as universal markers for the CRISPR system due to their high conservation (Deveau *et al.*, 2010). The variation in number of repeats and spacers in *CRISPR1* locus of all GBS isolates is as a result of an adaptive response to external phages. The relative loss of repeats and spacers might have occurred when foreign bacteriophages directly attacked these GBS isolates. Camel isolate ILRI005 had the least number of repeats and spacers in this locus. This suggest that camel *S. agalactiae* isolate ILRI005 was probably first to acquire *CRISPR1* locus, hence undergoing more loss of repeats and spacers during the course of time, however detailed mode of action needs to be elucidated in future. A review study on CRISPR system described that the loss of repeats and spacers can also occur due to homologous recombination event between repeats. The spacer content of *CRISPR* locus has also been found correlated with the susceptibility of phage that suggests spacers might provide a clue about past exposure of kind of phages (Deveau *et al.*, 2010). The BlastN search of *CRISPR1* locus of two camels isolates showed relative less %identity compared to human vs human, or bovine vs human isolates. Cattle *S. agalactiae* 09mas018883 and human *S. agalactiae* isolates' *CRISPR1* locus exhibited 99-100% (few had 96%) identity to all available human *S. agalactiae* whereas both camel *S. agalactiae* isolates showed <= 97% identity with each other and other GBS isolates. This relative low identity of camel *S. agalactiae* isolates suggests their earlier acquisition of this locus compared to human and cattle *S. agalactiae*. Future investigation on a larger set of GBS population from camels from the Horn of Africa would help us to elucidate the real mechanism of this adaptive activity of *CRISPR1* locus.

5.2.10 *CRISPR2* locus shared in two camel *S. agalactiae*

CRISPR2 locus identified was only shared among the camel *S. agalactiae* isolates ILRI005 and ILRI112. The repeats were of fixed length, 32 bp in ILRI005 and 33 bp in ILRI112. The non-repetitive spacers were of variable length ranging from 34-36 bp in ILRI005 while 31-36 bp in ILRI112.

Below are repeats (represented with black background) and spacers (represented with grey background) in *CRISPR2* locus of camel GBS ILRI005;

5’__[467351- *cas* genes -475383][475530- Repeats and Spacers -475895]__**3’**

```
GTCGCACCCTTTGCGGGTGCGTGGATTGAAAT
TATACAAACTTCTGCGTTATCTTCGTCATAATTA
GTCGCACCCTTTGCGGGTGCGTGGATTGAAAT
AAGTGGGTTAGTACAACTGAATGGGATGAAAAAC
GTCGCACCCTTTGCGGGTGCGTGGATTGAAAT
CTAAAGGTGTCTTATGGGATTCGAACCCATAGTGGC
GTCGCACCCTTTGCGGGTGCGTGGATTGAAAT
ATGCATTGATGTAACTTTCTATATTATTGACAACT
GTCGCACCCTTTGCGGGTGCGTGGATTGAAAT
TCCCAGTCCAATGTTTTATTAGCCATCTCAGCCTC
GTCGCACCCTTTGCGGGTGTGTAGTTTCAACT
```

Below are repeats and spacers in *CRISPR2* locus of camel GBS ILRI112;

5’__[468574- *cas* genes -475490][475637- Repeats and Spacers -476066]__**3’**

```
GTCGCACCCTTTGCGGGTGCGTGGATTGAAATA
AGAGATGCAAGTGTGGCAATGAAGAATTTTACA
GTCGCACCCTTTGCGGGTGCGTGGATTGAAATA
CCAACCTTGGGCGGTAGACTTTGACAAAAAGTCA
GTCGCACCCTTTGCGGGTGCGTGGATTGAAATA
GCTTGGTAGCCTCATTGATAGCTTGTATTGTT
GTCGCACCCTTTGCGGGTGCGTGGATTGAAATA
```

GTATTCCAAGTCAATGTTTTATGTAGCAATAAT
GTCGCACCCTTTGCGGGTGCGTGGATTGAAATT
GGAATGGTGCTGAATGGATTATCAATTCTTTTAGCG
GTCGCCCCCTTTGCGGGTGCATGGATTGAAATT
AATAATGATTATCTTTTTATTAATTCATTAT
CATCGCCCCTTTGCGGGTGTGTAGTTTTAACTA

Moreover, we found that the putative bacteriophage insertions in cattle and human *S. agalactiae* isolates were similar to each other, for example Phage-*Streptococcus*-PH10 was predicted in two cattle GBS isolates 09mas018883 and FSL-S3-026 as well as two human GBS isolates A909 and 2603V/R. In contrast, all putative phage insertions of camel *S. agalactiae* isolates were of a distinct kind, *i.e.* four predicted phage insertions in ILRI005, Phage-*Streptococcus-pyogenes*_315_1, Phage-*Streptococcus*-TP_J34, Phage-*Streptococcus*-Abc2 and Phage-*Bacillus*-virus_1; and Phage-OH2 in ILRI112 (Table 3, Paper **III**). The presence of similar kinds of putative phage insertions in cattle and human *S. agalactiae* isolates and the possession of single *CRISPR1* locus by these isolates strongly suggest that *CRISPR1* locus was sufficient for the defence activity of these isolates. The presence of *CRISPR2* locus with additional copies of *cas* genes, repeats and spacers in camel *S. agalactiae* isolates ILRI005 and ILRI112 is probably due to their exposure to different kinds of bacteriophages hence they acquired the *CRISPR2* locus as an adaptive mechanism against foreign phage DNA. Due to acquisition of these *CRISPR* loci, these *S. agalactiae* isolates possibly became virulent and pathogenic for their hosts. Detailed investigation of cas genes, repeats and spacers of *CRISPR1* locus and *CRISPR2* locus in camel GBS population, and the type of phages to which camel GBS are exposed to, would aid us to elucidate the mechanism of CRISPR system in *S. agalactiae* from camels in detail.

5.2.11 Other important features in *S. agalactiae*

In this study we discussed the possible organization and role of various operons and their possible role in *S. agalactiae* virulence, such as Lactose (*lac*) operon I, *lac* operon II, *cyl* operon, competence operon; capsular polysaccharide locus, pilus islands; and secretory proteins associated with type VII secretion system or Esx pathway. All these virulent operons and loci contained either core genes that were found in common in all GBS isolates, or shared genes found in common in some of the GBS isolates. The detail of each locus, its significance and possessed genes are discussed in detail in paper **III**.

5.2.12 General COG classification of core genes

A total of 11090 core genes were identified in seven *S. agalactiae* isolates, out of which 9041 (81.52%) core genes were found homologous with COG genes while 2049 (18.48%) had no matches to COG genes. The functional classification of homologus genes is shown in Table 4.

Table 4. *Functional classification of S. agalactiae core genes homologous with COG genes*

Functional Classification	Code	Number of genes	Percentage of core genome
Cellular processes and signalling			
Intracellular trafficking, secretion, and vesicular transport	U	98	0.88%
Cell cycle control, cell division, chromosome partitioning	D	112	1.01%
Signal transduction mechanisms	T	183	1.65%
Defence mechanisms	V	203	1.83%
Posttranslational modification, protein turnover, chaperones	O	301	2.71%
Cell wall/membrane biogenesis	M	453	4.08%
Information storage and processing			
Transcription	K	455	4.10%
Replication, recombination and repair	L	519	4.68%
Translation, ribosomal structure and biogenesis	J	972	8.76%
Metabolism			
Secondary metabolites biosynthesis, transport and catabolism	Q	38	0.34%
Lipid transport and metabolism	I	186	1.68%
Coenzyme transport and metabolism	H	290	2.62%
Energy production and conversion	C	322	2.90%
Nucleotide transport and metabolism	F	392	3.54%
Inorganic ion transport and metabolism	P	461	4.16%
Amino acid transport and metabolism	E	670	6.04%
Carbohydrate transport and metabolism	G	735	6.63%
Poorly characterized			
Function unknown	S	810	7.30%
General function prediction only	R	999	9.01%
Other categories			
Multi-functions	-	842	7.59%

5.3 Comparative analysis of *S. aureus* isolates (Paper IV)

Staphylococcus aureus ILRI_Eymole1/1 isolated from a dromedary camel in Kenya belongs to MLST 30 based on the analysis of its seven house-keeping genes using the MLST system (Enright *et al.*, 2000). The seven house-keeping genes were *arcc*, *aroe*, *glpf*, *gmk*, *pta*, *tpi* and *yqil*; encoding carbamate kinase, shikimate dehydrogenase, glycerol kinase, guanylate kinase, phosphate acetyltransferase, triosephosphate isomerase, and acetyl coenzyme A acetyltransferase, respectively. This newly sequenced camel *S. aureus* isolate was compared with nineteen previously sequenced human *S. aureus* isolates of type CC30, to investigate its genetic similarities and heterogeneity from previously sequenced CC30 *S. aureus* isolates.

5.3.1 General Genomic features of CC30 *S. aureus* isolates

Among 20 CC30 *S. aureus* isolates, only four are completely finished (ILRI_Eymole1/1, MRSA252, 55/2053 and TCH60) while 16 are in draft/unfinished status (Table 5, paper **IV**). The genome size of each draft genome was estimated by total sum of the size of all contigs/scaffolds in that genome (Stretches of Ns were also excluded). Camel *S. aureus* ILRI_Eymole1/1 has a genome size of 2.87 Mbp. The genomic size of all twenty CC30 *S. aureus* isolates ranged from 2.74 Mbp for *S. aureus* WW2703_97 to 2.90 Mbp for *S. aureus* MRSA252. GC% for these *S. aureus* isolates ranged from 32.66 for *S. aureus* MRSA-M2 to 32.88 for *S. aureus* ILRI_Eymole1/1. Ribosomal RNA genes were 12 in *S. aureus* 55/2053, 58_424 and EMRSA16, 13 in Btn1260, 14 in MN8, 16 in ILRI_Eymole1/1, MRSA252 and A017934_97, and 19 in *S. aureus* TCH60. All other CC30 *S. aureus* genomes had only four rRNA genes, and MRSA-M2 had five, possibly due to their draft genome status. CC30 *S. aureus* isolates contained tRNA genes ranging from 40 for *S. aureus* M809 to 60 for *S. aureus* ILRI_Eymole1/1 and MRSA252. The number of protein-encoding genes was 2532 for *S. aureus* 55/2053 to 2770 for *S. aureus* MRSA-M2; camel *S. aureus* ILRI_Eymole1/1 had 2755.

5.3.2 Core genome of CC30 *S. aureus* isolates

All twenty CC30 *S. aureus* genomes had in total 53,037 protein encoding genes, 43,919 (82.81%) of which were identified as their core genome. Of these core genes, 36,451 (83%) were found to have homologous functions with the COG genes, whereas 7,468 (17%) were not present in the COG database. The functional classification of the CC30 *S. aureus* core genome homologous with COG functions is shown in Table 5.

Table 5. *Functional classification of CC30 S. aureus core genes homologous with COG genes*

Functional Classification	Code	Number of genes	Percentage of core genome
Cellular processes and signalling			
Flagellum specific	NU	20	0.05%
Cell cycle control and cell division	D	360	0.82%
Intracellular trafficking and secretion	U/NU/NOU	400	0.91%
Signal transduction mechanisms	T	684	1.56%
Defence mechanisms	V	803	1.83%
Posttranslational modification, protein turnover, chaperones	O	1162	2.65%
Cell wall/membrane biogenesis	M/GM	1753	3.99%
Information storage and processing			
Chromatin structure and dynamics (Histone related)	BQ	20	0.05%
Replication, recombination and repair	L	1733	3.95%
Transcription regulators, repressors, anti-terminators	K/KT	2200	5.01%
Translation, ribosomal structure and biogenesis	J	2698	6.14%
Metabolism			
Secondary metabolites biosynthesis, transport, catabolism	Q	400	0.91%
Lipid transport and metabolism	I	920	2.10%
Nucleotide transport and metabolism	F	1142	2.60%
Coenzyme transport and metabolism	H	1462	3.33%
Energy production and conversion	C	1762	4.01%
Carbohydrate transport and metabolism	G/GT	2086	4.75%
Inorganic ion transport and metabolism	P	2389	5.44%
Amino acid transport and metabolism	E	3122	7.11%
Poorly characterized			
General function prediction only	R	3907	8.90%
Function unknown	S	3782	8.61%
Other categories			
Multi-functions	-	3646	8.30%

5.3.3 Important features in ILRI_Eymole1/1's core, shared and isolate-specific genes.

We identified 2,163 (78.51%) core genes, 507 (18.40%) variable/shared genes and 85 (3.09%) isolate-specific genes of the total protein encoding genes in camel *S. aureus* ILRI_Eymole1/1.

Isolate *S. aureus* ILRI_Eymole1/1 possessed genes encoding surface proteins known to be related to the adhesion of *S. aureus* to the epithelial cells, such as fibrinogen-binding protein ClfB (CEH27447: shared gene), Heme regulated surface protein IsdA (CEH26009: core gene), serine-aspartic acid repeat adhesin proteins SdrC and SdrE (CEH25318 and CEH25319, respectively: shared genes). It contained a gene encoding an extracellular adherence protein Eap (CEH26760) as part of its core genome. Eap is known to be important for *S. aureus* internalization and long time persistence in host cells (Haggar *et al.*, 2003).

The virulence factors in bacteria are either secreted into the extracellular micro-environment or injected directly into the host cell to develop pathogenicity to the host. Gram-positive bacteria are generally believed to have this secretion through a simpler mechanism due to their single surrounding membrane. However, a specialized secretory system has been reported in Gram-positive bacteria *M. tuberculosis*, *S. aureus*, *B. subtilis* and others (Burts *et al.*, 2005; Abdallah *et al.*, 2007). Camel *S. aureus* ILRI_Eymole1/1 core genome contains many secretory proteins known to have an essential role in the Ess/Esx/ESAT-6/type-VII secretory pathway. These core secretory proteins were the secretory antigen precursor protein SsaA (CEH25002), ESAT-6/Esx family secreted protein EsxA (CEH25003), putative secretion accessory protein EsaA (CEH25004), putative secretion system component EssA (CEH25005), putative secretion accessory protein EsaB (CEH25006), putative secretion system component EssB (CEH25007), FtsK/SpoIIIE family protein, and a putative secretion system component EssC (CEH25008).

As a result of protein repertoire comparison of CC30 *S. aureus* isolates, a total of 79 out of 85 isolate-specific genes in ILRI_Eymole1/1 were found located in the four putative phage insertions and two superantigen pathogenicity islands (SaPI). This suggests these insertions into camel *S. aureus* ILRI_Eymole1/1 are possibly due to horizontal gene transfer. Superantigen pathogenicity islands or SaPI are mobile genetic elements known to be associated with virulence and resistance in bacteria (Ubeda *et al.*, 2008). Among twenty CC30 *S. aureus* isolates used in this study, thirteen isolates contained SaPI1 island, two isolates had SaPI4 island and four of the isolates had SaPI2 and SaPI4 islands. The camel *S. aureus* isolate ILRI_Eymole1/1 had two islands SapIcam1 and SaPIcam2.

5.3.4 Phylogenetic relationship with human CC30 *S. aureus* isolates

More conserved core genes were extracted having identity >= 95% with at least 90% alignment length for all CC30 *S. aureus* isolates. Two additional *S. aureus* isolates, ST1 (Mu50: NC_002758) and an ST5 (N315: NC_002745) were also used in the analysis to be used as an outgroup for CC30 isolates, and phylogeny was established. The phylogenetic tree showed that camel ST30 *S. aureus* isolate ILRI_Eymole1/1 clustered with human CC30 *S. aureus* isolates (Figure 8). An unrooted tree was also constructed based on CC30 *S. aureus* isolates only, which showed that camel ST30 *S. aureus* isolate was genetically distinct from human *S. aureus* CC30 isolates due to SNP variations in conserved genes (Figure 9).

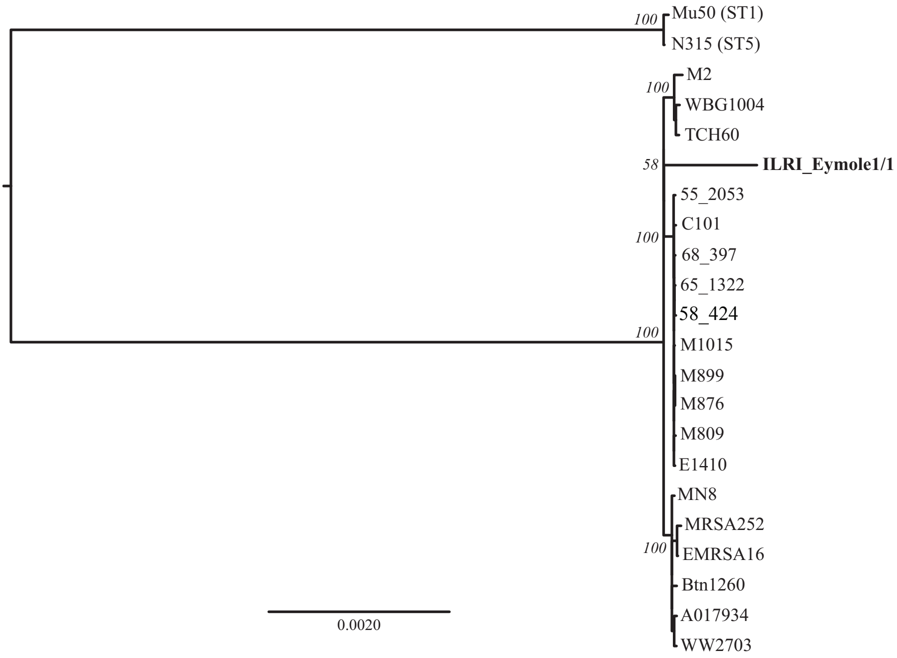


Figure 8. Maximum likelihood tree of the concatenated sequence of selected 283 core genes in 20 CC30 *S. aureus* isolates; one ST1 and one ST5 *S. aureus* isolate Mu50 and N315 respectively. General Time Reverse (GTR) model was used with 100 bootstrap replications. The bootstrap values are represented above the nodes. ST1 and ST5 are out grouped.

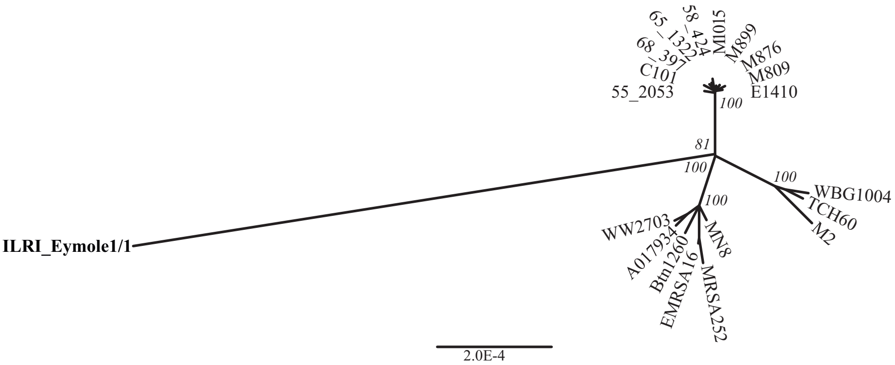


Figure 9. Maximum likelihood unrooted tree of 20 CC30 *S. aureus* isolates using set of 283 core genes. General Time Reverse (GTR) model and 100 bootstrap replications were used. The values indicated are the bootstrap values.

6 Conclusions

Availability of newly sequenced genomes of bacterial pathogens from camels and their comparative analysis with previously sequenced genomes of isolates obtained from cattle and human revealed many important findings.

- *CRISPR1* locus conserved in all *S. agalactiae* genomes could be potentially involved in the pathogenicity of all *S. agalactiae* isolates, and could serve as a universal vaccine candidate target for the treatment of GBS infections in multiple hosts. *CRISPR2* locus present in only camel GBS isolates 'ILRI005 and ILRI112' could be a potential pathogenicity associated locus in GBS originating infections in camels. It could be important for developing a host-specific vaccine or therapeutic approach for GBS infections in camels.
- Many other important virulence factors were found to be present in GBS core genome, such as type VII secretion system associated genes, competence operon, *lac* I operon and *cyl* locus.
- Many virulence factors were found to be shared among some of the *S. agalactiae* isolates, such as tetracycline resistance gene *tetM* known for acquiring antimicrobial resistance was identified in 09mas018883 (cattle) ILRI112 (camel), 2603V/R (human) GBS isolates. This gene was found associated with Tn*916* transposon in these three isolates. This finding is relevant for the development of strategies to combat antibiotic-resistance. Other important locus shared among some GBS isolates was *lac* II operon.
- Pan proteome analysis revealed that putative phage insertions in camel *S. agalactiae* isolates were isolate-specific, which is consistent with their exposure to less common kind of external phages, whereas those of the cattle and human *S. agalactiae* were shared among each other suggesting their close interactions and exposure to similar kind of external bacteriophages.

- Phylogenetic analysis revealed that two camel GBS isolates were paired together but were relatively distant from human and cattle GBS isolates. The cattle and human GBS isolates were relatively close to each other particularly the cattle *S. agalactiae* isolate 09mas018883 was clustered more closely to the human *S. agalactiae* isolates.
- Many genes encoding bacterial adhesion-related proteins were identified in camel *S. aureus* isolate ILRI_Eymole1/1; some were the part of *S. aureus* CC30 core genome while some were the part of variable genome. For example ClfB, SdrC, SdrE were only present in some of the *S. aureus* isolates, while IsdA and Eap were found as the part of the core genome. Likewise, genes encoding the secretory proteins of essential significance in type VII secretion system were present in the core genome. High proportion of ILRI_Eymole1/1's protein encoding genes (~97%) was found common to some or all of CC30 *S. aureus* isolates from human.
- Isolate-specific genes in novel *S. aureus* isolate from camel were found associated with putative phage insertions suggesting their acquisition through lateral gene transfer, like *S. agalactiae* camel isolates.
- Phylogenetic analysis using polymorphic core genes revealed that camel *S. aureus* isolate ILRI_Eymole1/1 falls within the human CC30 *S. aureus* isolate cluster but is genetically distinct based on SNP variations.

7 General Discussion and Future Perspectives

The availability of novel genome sequences for *S. agalactiae* and *S. aureus* isolates from camel, presented in this thesis provided the opportunity to explore their genome content in detail, and relate them with biological significance and virulence in various ways. The approach of combining the results of mapping and *de novo* assembly in Paper **I** and **II** was suitable not only for gap closure and sequence refinement but also allowed the identifications of large insertions in the genomes such as large insertion in camel *S. agalactiae* ILRI112's *tetM* gene associated transposon Tn*916*. Moreover, by closely examining the results of mapping and *de novo* assemblies in Mauve aligner, further enhanced our confidence level in genome areas where both kinds of assemblies generated the same result. It can also be advantageous to perform genome assembly of the same NGS data with two different assembler tools because sometimes one assembler can expose regions that another cannot due to different algorithms implemented. We have relied on a single kind of NGS data for any particular isolate's genome assembly due to limitation in resources. However we would encourage that a singe bacterial isolate should be sequenced using at least two different kinds of NGS platforms, to ensure the accuracy of SNPs in genomic data. In case of mapping assembly, one of the important points is to choose a good reference genome to align the reads data. We have selected reference genomes based on maximum percentage of data reads aligned to it, although it was challenging to quickly estimate which reference genome will do the job. We have performed a quick mapping assembly to estimate a good reference using a fast mapping assembly tool, Mosaik aligner (Lee *et al.*, 2014) and used the best suited reference genome to do mapping assembly using MIRA assembler that takes long time and high memory.

In Paper **III** and **IV**, we used core genome to determine the phylogenetic relationship of *S. agalactiae* and *S. aureus* isolates from camels with those from other hosts *i.e. S. agalactiae* from humans and cattle and *S. aureus* from humans. Although phylogenies based on few genes such as house keeping genes depicts a good phylogenetic relationship among isolates, but it does not depict genetic diversity at whole genome level. The core genome based phylogeny among bacterial isolates used in this thesis established phylogenetic relationship at broader level. The affordability of NGS data and relative convenience of generating more and more bacterial genome sequences will certainly increase the trend of using core genome data for deciphering intra-species as well as inter-species phylogenetic relationship among genomes. Both *S. agalactiae* and *S. aureus* isolates from camels were relatively distinct in their phylogenetic relationship to other hosts. However it is important to sequence more of these isolates from camels as well as humans, cattle, sheep and other community animals in Kenya. The large-scale phylogenetic analysis of Kenyan *S. agalactiae* and *S. aureus* isolates from these different hosts will in future provide further insights into patterns of their genetic diversity among each other.

The whole genome analysis of bacterial pathogens through comparative genomics described in this thesis provides a convincing strategy to identify virulence genes that can be used as potential vaccine targets. The phenomenon of gaining new genes or the loss of existing genes appeared prevalent in *S. agalactiae* and *S. aureus* pathogens as a protective immune response against foreign attacks such as exposure to bacteriophages. The protein repertoire comparison of *S. agalactiae* and *S. aureus* helped us to analyse core genes, shared genes and isolate specific genes in detailed perspective. It helped us to highlight similarities and differences in particular loci of various isolates, by observing either conservation or variation in their constituent number of genes. For example the same number of *cas* genes (four) were found to be present in *CRISPR1* locus of all *S. agalactiae*. Likewise many genes of the *cps* locus were identified as core genes, some as shared/variable genes and some as isolate specific genes or in other words insertion sequences with in the locus. The conservation and variation of genes has been discussed in many other loci in *S. agalactiae* isolates, such as *lac* I operon, *lac* II operon, *cyl* operon, competence operon and genes associated with type VII secretion system. These detailed differences (gain/loss of genes) could be significant for their specific pathogenic traits in their hosts. The experimental validation and expression analysis of the candidate loci identified in this research would be helpful in understanding their mode of action in detail.

CRISPR2 locus of potential pathogenicity was found to be present in both camel *S. agalactiae* isolates while absent in others. The question is whether, all GBS isolates associated with infections in camels would possess it? So it would be interesting to investigate this locus in GBS population from camels with infections. If these results were positive, it would suggest that the *CRISPR2* locus is the real cause of pathogenicity in camels. Knocking out the genes in this locus or inhibiting their expression and evaluating protective immunity in GBS isolates can further confirm the results. Further it would be interesting to examine the variations of *CRISPR2* locus in various camel GBS populations in future. The high proportion of the isolate-specific genomic make up of *S. agalactiae* and *S. aureus* isolates from camels was found to be horizontally transferred due to its clustering within large insertions such as putative phage insertions in *S. agalactiae* ILRI005 and *S. aureus* ILRI_Eymole1/1 and a large insertion in *S. agalactiae* ILRI112. Isolation and genome sequencing of additional *S. agalactiae* and *S. aureus* isolates from various infections of camels, cattle, sheep, goats and humans of the same region, and their thorough investigation would be advantageous to understand the host-pathogens interactions, host adaptation and zoonotic potential of these bacteria in a better way.

S. agalactiae isolates ILRI112 from camel, 2603V/R from human and 09mas018883 from cattle possessed Tn*916* like genetic element that also carried *tetM* gene (Supplementary Figure 1, paper **III**). Tn*916* family is responsible for resistance to different kind of antibiotics in various bacterial pathogens (Roberts & Mullany, 2011). We suggest a future study to investigate the source of transposon Tn*916* possessing *tetM* gene as a resistance gene. As initial step we identified it to be present in a few of the sequenced genomes *S. pneumonia* GA60132, *S. pneumonia* GA58981, *S. pneumonia* GA47502 and *S. gallolyticus subsp. gallolyticus* ATCC 43143, in addition to three mentioned *S. agalactiae* genomes. It would be interesting to further investigate the rate of infection or transfer of Tn*916* in *S. agalactiae*, using *in vitro* filter mating protocol (Werner *et al.*, 2011).

Current study showed the presence of *tetM* gene in three *S. agalactiae* isolates that indicated their acquisition of resistance against antibiotic 'tetracycline'. In Kenyan camels about 34% of GBS isolates possessed *tetM* (Fischer *et al.*, 2013). Its dissemination to susceptible isolates is highly probable and could produce super resistant strains. The acquisition of antibiotics resistance by pathogens is making several infectious diseases difficult to treat both in humans and animals. The rapid solution could be to start treating GBS infections with some available susceptible antibiotic. However, increasing resistance to antibiotics and shortage of new kinds of susceptible antibiotics restricts future use of antibiotics. One way to avoid use of antibiotics could be the use of phage therapy that is quite beneficial in treating infections that cannot be treated due to antibiotics resistance (Abedon *et al.*, 2011), however this also have certain challenges. Another alternative could be vaccine discovery. Vaccines can prevent infections and avoid complications associated with infections (Mishra *et al.*, 2012). Traditionally vaccinology-based methods were based on the expensive experimental procedures to screen only few known candidate features at a time, but the opportunity to access various pathogens' whole genome sequence has made it possible to investigate all potential vaccine targets. The advent of NGS technologies made the availability of accurate genomic data fast, cheap and convenient (Metzker, 2010). As a result, access to a large number of genome sequences for bacterial pathogens has introduced sequence based reverse vaccinology approaches through the application of comparative genomics, pan genome analysis, subtractive genomics, transcriptomics, proteomics, immunomics and structural genomics as ways to develop vaccines (Seib *et al.*, 2012). The availability of novel *S. agalactiae* and *S. aureus* genomes from camel and cattle and their detailed comparative genomics helped us to expose potential virulence genes that could be relevant for future vaccine discovery and the development of control measures.

References

Abdallah, A. M., van Pittius, N. C. G., Champion, P. A. D., Cox, J., Joen, L., Vandenbroucke-Grauls, C. M. J. E., Appelmelk, B. J. & Bitter, W. (2007). Type VII secretion - mycobacteria show the way. *Nature Reviews. Microbiology*, 5(november), pp. 883–891.

Abdelgadir, A. E. (2014). Mastitis in camels (Camelus dromedarius): Past and recent research in pastoral production system of both East Africa and Middle East. *Journal of veterinary medicine and animal health*, 6(7), pp. 208–216.

Abedon, S. T., Kuhl, S. J., Blasdel, B. G. & Kutter, E. M. (2011). Phage treatment of human infections. *Bacteriophage*, 1(2), pp. 66–85.

Abera, M., Habte, T., Aragaw, K., Asmare, K. & Sheferaw, D. (2012). Major causes of mastitis and associated risk factors in smallholder dairy farms in and around Hawassa, Southern Ethiopia. *Tropical Animal Health and Production*, 44(6), pp. 1175–1179.

Ahmad, S., Yaqoob, M., Bilal, M. Q., Muhammad, G., Yang, L.-G., Khan, M. K. & Tariq, M. (2012). Risk factors associated with prevalence and major bacterial causes of mastitis in dromedary camels (Camelus dromedarius) under different production systems. *Tropical animal health and production*, 44(1), pp. 107–12.

Aiken, A. M., Mutuku, I. M., Sabat, A. J., Akkerboom, V., Mwangi, J., Scott, J. A. G., Morpeth, S. C., Friedrich, A. W. & Grundmann, H. (2014). Carriage of Staphylococcus aureus in Thika Level 5 Hospital, Kenya: a cross-sectional study. *Antimicrobial resistance and infection control*, 3(1), p 22.

Al-Doughaym, A., Mustafa, K. & Mohamed, G. (1999). Aetiological study on pneumonia in camels (Camelus dromedarius) and in vitro antibacterial sensitivity pattern of the isolates. *Pakistan Journal of Biological Sciences*, 2(4), pp. 1102–1105.

Al-Juboori, A. A., Kamat, N. K. & Sindhu, J. I. (2013). Prevalence of some mastitis causes in dromedary camels in Abu Dhabi, United Arab Emirates. *Iraqi Journal of Veterinary Sciences*, 27(1), pp. 9–14.

Alhendi, A. A. B. (1999). Nasal microflora of camels (Camelus dromedarius) under two different conditions. *Pakistan Vet. J*, 19(4), pp. 164–167.

Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L. & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC genomics*, 12(1), p 402 BioMed Central Ltd.

Angiuoli, S. V & Salzberg, S. L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics (Oxford, England)*, 27(3), pp. 334–42.

Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. (2009). ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics (Oxford, England)*, 25(15), pp. 1968–1969.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. & Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, 9, pp. 75–89.

Bani Ismail, Z., Al-Rukibat, R., Al-Tarazi, Y. & Al-Zghoul, M. B. (2007). Synovial Fluid Analysis and Bacterial Findings in Arthritic Joints of Juvenile Male Camel (Camelus dromedarius) Calves. *Journal of Veterinary Medicine Series A*, 54(2), pp. 66–69.

Barlow, J. (2011). Mastitis Therapy and Antimicrobial Susceptibility : a Multispecies Review with a Focus on Antibiotic Treatment of Mastitis in Dairy Cattle. *J Mammary Gland Biol Neoplasia*, 16, pp. 383–407.

Boetzer, M. & Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome biology*, 13(6), p R56 BioMed Central Ltd.

Borchardt, S. M., DeBusscher, J. H., Tallman, P. A., Manning, S. D., Marrs, C. F., Kurzynski, T. A. & Foxman, B. (2006). Frequency of antimicrobial resistance among invasive and colonizing Group B streptococcal isolates. *BMC infectious diseases*, 6, p 57.

Bornstein, S. & Younan, M. (2013). Significant veterinary research on the dromedary camels of Kenya : Past and Present. *Journal of Camelid Science*, 6, pp. 1–48.

Burts, M. L., Williams, W. A., DeBord, K. & Missiakas, D. M. (2005). EsxA and EsxB are secreted by an ESAT-6-like system that is required for the pathogenesis of Staphylococcus aureus infections. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4), pp. 1169–74.

Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, 28(4), pp. 464–469.

Chevreux, B., Wetter, T. & Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information., GCB, Göttingen, Germany, 1999. pp. 45–56. GCB, Göttingen, Germany.

Christou, L. (2011). The global burden of bacterial and viral zoonotic infections. *Clinical microbiology and infection*, 17(3), pp. 326–30.

Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7), pp. 1394–403.

Dear, S., Durbin, R., Hillier, L., Marth, G., Thierry-Mieg, J. & Mott, R. (1998). Sequence Assembly with CAFTOOLS. *Genome research*, 8, pp. 260–267.

Detilleux, J. C. (2009). Genetic factors affecting susceptibility to udder pathogens. *Veterinary microbiology*, 134(1-2), pp. 157–64.

Deveau, H., Garneau, J. E. & Moineau, S. (2010). CRISPR/Cas system and its role in phage-bacteria interactions. *Annual review of microbiology*, 64, pp. 475–93.

Dogan, B., Schukken, Y. H., 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), pp. 5899–906.

Van Domselaar, G. H., Stothard, P., Shrivastava, S., Cruz, J. A., Guo, A., Dong, X., Lu, P., Szafron, D., Greiner, R. & Wishart, D. S. (2005). BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res*, 33(Web Server issue), pp. W455–9.

Duran, N., Ozer, B., Duran, G. G., Onlen, Y. & Demir, C. (2012). Antibiotic resistance genes and susceptibility patterns in staphylococci. *Indian J Med Res*, 135(March), pp. 389–396.

Enright, M. C., Day, N. P. J., Davies, C. E. & Peacock, S. J. (2000). Multilocus Sequence Typing for Characterization of Methicillin- Resistant and Methicillin-Susceptible Clones of Staphylococcus aureus. *Journal of clinical microbiology*, 38(3), pp. 1008–1015.

Fischer, A., Liljander, A., Kaspar, H., Muriuki, C., Fuxelius, H.-H., Bongcam-Rudloff, E., de Villiers, E. P., Huber, C. A., Frey, J., Daubenberger, C., Bishop, R., Younan, M. & Jores, J. (2013). Camel Streptococcus agalactiae populations are associated with specific disease complexes and acquired the tetracycline resistance gene tetM via a Tn916-like element. *Veterinary research*, 44(1), p 86 Veterinary Research.

Fitzgerald, J. R. (2012). Livestock-associated Staphylococcus aureus: origin, evolution and public health threat. *Trends in microbiology*, 20(4), pp. 192–8 Elsevier Ltd.

Galal, K., Hameed, A. & Sender, G. (2008). An association of BoLA alleles DRB3.2*16 and DRB3.2*23 with occurrence of mastitis caused by different bacterial species in two herds of dairy cows. *Animal Science Papers and Reports*, 26(1), pp. 37–48.

Gao, J., Yu, F., Luo, L., He, J., Hou, R., Zhang, H., Li, S., Su, J. & Han, B. (2012). Antibiotic resistance of Streptococcus agalactiae from cows with mastitis. *The Veterinary Journal*, 194(3), pp. 423–424 Elsevier Ltd.

Gautret, P., Benkouiten, S., Gaillard, C., Parola, P. & Brouqui, P. (2013). Camel milk-associated infection risk perception and knowledge in French Hajj pilgrims. *Vector borne and zoonotic diseases*, 13(6), pp. 425–7.

Glaser, P., Rusniok, C., Buchrieser, C., Chevalier, F., Frangeul, L., Msadek, T., Zouine, M., Couvé, E., Lalioui, L., Poyart, C., Trieu-Cuot, P. & Kunst, F. (2002). Genome sequence of Streptococcus agalactiae, a pathogen causing invasive neonatal disease. *Mol Microbiol*, 45(6), pp. 1499–513.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P. & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, 57, pp. 81–91.

Grissa, I., Vergnaud, G. & Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic acids research*, 35(Web Server issue), pp. W52–7.

Guinane, C. M., Ben Zakour, N. L., Tormo-Mas, M. A., Weinert, L. A., Lowder, B. V, Cartwright, R. A., Smyth, D. S., Smyth, C. J., Lindsay, J. A., Gould, K. A., Witney, A., Hinds, J., Bollback, J. P., Rambaut, A., Penadés, J. R. & 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, pp. 454–66.

Guindon, S. & Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5), pp. 696–704.

Haggar, A., Hussain, M., Lo, H., Herrmann, M., Norrby-Teglund, A. & Flock, J. (2003). Extracellular Adherence Protein from Staphylococcus aureus Enhances Internalization into Eukaryotic Cells. *Infection and immunity*, 71(5), pp. 2310–2317.

Heijne, G. Von (1992). Membrane Hydrophobicity Protein Structure Prediction Analysis and the Positive-inside. *J Molecular Biology*, 225, pp. 487–494.

Hillerton, J. E. (1999). Balancing mastitis and quality., Stonelergh, UK, 1999. pp. 31–36. Stonelergh, UK.

Hogeveen, H. (2005). Mastitis is an economic problem., Stoneleigh, UK, 2005. pp. 1–73. Stoneleigh, UK.

Holden, M. T. G., Feil, E. J., Lindsay, J. A., Peacock, S. J., Day, N. P. J., Enright, M. C., Foster, T. J., Moore, C. E., Hurst, L., Atkin, R., Barron, A., Bason, N., Bentley, S. D., Chillingworth, C., Chillingworth, T., Churcher, C., Clark, L., Corton, C., Cronin, A., Doggett, J., Dowd, L., Feltwell, T., Hance, Z., Harris, B., Hauser, H., Holroyd, S., Jagels, K., James, K. D., Lennard, N., Line, A., Mayes, R., Moule, S., Mungall, K., Ormond, D., Quail, M. A., Rabbinowitsch, E., Rutherford, K., Sanders, M., Sharp, S., Simmonds, M., Stevens, K., Whitehead, S., Barrell, B. G., Spratt, B. G. & Parkhill, J. (2004). Complete genomes of two clinical Staphylococcus aureus strains : Evidence for the rapid evolution of virulence and drug resistance. *Proc. Natl. Acad. Sci. U.S.A.*, 101(26), pp. 9786–9791.

Horvath, P. & Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. *Science*, 327(5962), pp. 167–70.

Huang, X. (1992). A contig assembly program based on sensitive detection of fragment overlaps. *Genomics*, 14(1), pp. 18–25.

Huson, D. H., Reinert, K. & Myers, E. W. (2002). The Greedy Path-Merging Algorithm for Contig Scaffolding. *Journal of the ACM*, 49(5), pp. 603–615.

Imperi, M., Pataracchia, M., Alfarone, G., Baldassarri, L., Orefici, G. & Creti, R. (2010). A multiplex PCR assay for the direct identification of the capsular type (Ia to IX) of Streptococcus agalactiae. *Journal of microbiological methods*, 80(2), pp. 212–214.

Jain, B., Tewari, A., Bhandari, B. B. & Jhala, M. K. (2012). Antibiotic resistance and virulence genes in Streptococcus agalactiae isolated from cases of bovine subclinical mastitis. *Vet. arhiv*, 82(5), pp. 423–432.

Jones, G. M. (2009). Understanding the Basics of Mastitis. *Virginia State University, USA*, 404(233), pp. 1–5.

Kagunyu, A. W. & Wanjohi, J. (2014). Camel rearing replacing cattle production among the Borana community in Isiolo County of Northern Kenya, as climate variability bites. *Pastoralism: Research, Policy and Practice*, 4(1), p 13.

Kaufmann, B. (1998). *Analysis of pastoral camel husbandry in Northern Kenya*. p 194 Center for Agriculture in the Tropics and Subtropics, University of Hohenheim. Weikersheim: Margraf. Verlag, Germany.

Khan, M. & Khan, A. (2006). Basic facts of mastitis in dairy animals: a review. *Pakistan Vet. J*, 26(4), pp. 204–208.

Kulig, H., Kmieć, M. & Wojdak-Maksymiec, K. (2010). Associations between Leptin Gene Polymorphisms and Somatic Cell Count in Milk of Jersey Cows. *Acta Veterinaria Brno*, 79(2), pp. 237–242.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biol*, 5(2), p R12.

Ladhani, S. (2004). Bacteraemia due to Staphylococcus aureus. *Archives of Disease in Childhood*, 89(6), pp. 568–571.

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T. & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research*, 35(9), pp. 3100–3108.

Lambert, P. A. (2005). Bacterial resistance to antibiotics: modified target sites. *Advanced drug delivery reviews*, 57(10), pp. 1471–85.

Lancefield, R. C. (1933). A serological differentiation of human and other groups of hemolytic streptococci. *J. Exp. Med*, 57(1), pp. 571–595.

Langille, M. G. I. & Brinkman, F. S. L. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics*, 25(5), pp. 664–665.

Lee, W.-P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P. & Marth, G. T. (2014). MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PloS one*, 9(3), p e90581.

Leyva-Baca, I., Schenkel, F., Sharma, B. S., Jansen, G. B. & Karrow, N. A. (2007). Identification of single nucleotide polymorphisms in the bovine CCL2, IL8, CCR2 and IL8RA genes and their association with health and production in Canadian Holsteins. *Animal genetics*, 38(3), pp. 198–202.

Liddy, H. & Holliman, R. (2002). Group B Streptococcus highly resistant to gentamicin. *Journal of Antimicrobial Chemotherapy*, 50(1), pp. 142–143.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, 251364.

Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE : a program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.*, 25(5), pp. 955–964.

Maina, E. K., Kiiyukia, C., Wamae, C. N., Waiyaki, P. G. & Kariuki, S. (2013). Characterization of methicillin-resistant Staphylococcus aureus from skin and soft tissue infections in patients in Nairobi, Kenya. *International journal of infectious diseases*, 17(2), pp. e115–9 International Society for Infectious Diseases.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y., Chen, Z., Dewell, B., Du, L., Fierro, J. M., Gomes, X. V, Goodwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Maria, L. I., Jarvie, T. P., Jirage, K. B., Kim, J., Knight, J. R., Lanza, R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Kenton, L., Lu, H., Makhijani, V. B., Mcdade, K. E., Mckenna, M. P., Myers, W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Greg, A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Richard, F. & Rothberg, J. M. (2005).

Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature*, 437(7057), pp. 376–380.

Mekibib, B., Furgasa, M., Abunna, F., Megersa, B. & Regassa, A. (2010). Bovine Mastitis : Prevalence , Risk Factors and Major Pathogens in Dairy Farms of Holeta. *Veterinary World*, 3(9), pp. 397–403.

Ménigaud, S., Mallet, L., Picord, G., Churlaud, C., Borrel, A. & Deschavanne, P. (2012). GOHTAM: a website for "Genomic Origin of Horizontal Transfers, Alignment and Metagenomics". *Bioinformatics (Oxford, England)*, 28(9), pp. 1270–1271.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1), pp. 31–46 Nature Publishing Group.

Miller, J. R., Koren, S. & Sutton, G. (2010). Assembly algorithm for Next-Ganeration Sequencing data. *Genomics*, 95(6), pp. 315–327.

Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., Shaw, P. D. & Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data. *Briefings in bioinformatics*, 14(2), pp. 193–202.

Mishra, R. P. N., Oviedo-Orta, E., Prachi, P., Rappuoli, R. & Bagnoli, F. (2012). Vaccines and antibiotic resistance. *Current opinion in microbiology*, 15(5), pp. 596–602 Elsevier Ltd.

Nakamura, P. A. M., Schuab, R. B. B., Neves, F. P. G., Pereira, C. F. A., Paula, G. R. De & Barros, R. R. (2011). Antimicrobial resistance profiles and genetic characterisation of macrolide resistant isolates of Streptococcus agalactiae. *Mem Inst Oswaldo Cruz, Rio de Janeiro*, 106(2), pp. 119–122.

Nishito, Y., Osana, Y., Hachiya, T., Popendorf, K., Toyoda, A., Fujiyama, A., Itaya, M. & Sakakibara, Y. (2010). Whole genome assembly of a natto production strain Bacillus subtilis natto from very short read data. *BMC genomics*, 11, p 243.

Obied, A. I. & Bagadi, H. O. (1996). Mastitis in Camelus dromedarius and the somatic cell content of camels' milk. *Research in Veterinary Science*, 61, pp. 55–58.

Ogorevc, J., Kunej, T., Razpet, A. & Dovc, P. (2009). Database of cattle candidate genes and genetic markers for milk production and mastitis. *Animal genetics*, 40(6), pp. 832–51.

Oliver, S. P., Murinda, S. E. & Jayarao, B. M. (2011). Impact of antibiotic use in adult dairy cows on antimicrobial resistance of veterinary and human pathogens: a comprehensive review. *Foodborne pathogens and disease*, 8(3), pp. 337–55.

Peltola, H., Soderlund, H. & Ukkonen, E. (1984). SEQUAID: a DNA sequence assembly program based on a mathematical model. *Nucl. Acids Res.*, 12(1), pp. 307–321.

Peters, J. & von den Driesch, A. (1997). The two-humped camel (Cumelus buctrianus): new light on its distribution, management and medical treatment in the past. *Journal of Zoology*, 242, pp. 651–679.

Petersen, A., Stegger, M., Heltberg, O., Christensen, J., Zeuthen, A., Knudsen, L., Urth, T., Sorum, M., Schouls, L., Larsen, J., Skov, R. & Larsen, A. (2013). Epidemiology of methicillin-resistant Staphylococcus aureus carrying the novel mecC gene in Denmark corroborates a zoonotic reservoir with transmission to humans. *Clinical microbiology and infection*, 19(1), pp. E16–22.

Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10), pp. 785–786 Nature Publishing Group.

Pop, M., Phillippy, A., Delcher, A. L. & Salzberg, S. L. (2004). Comparative genome assembly. *Briefings in bioinformatics*, 5(3), pp. 237–248.

Poyart, C., Jardy, L., Quesne, G., Berche, P. & Trieu-Cuot, P. (2003). Genetic Basis of Antibiotic Resistance in Streptococcus agalactiae Strains Isolated in a French Hospital. *Antimicrobial Agents and Chemotherapy*, 47(2), pp. 794–797.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13(1), p 341 BMC Genomics.

Regassa, A., Golicha, G. & Tesfaye, D. (2013). Prevalence , risk factors , and major bacterial causes of camel mastitis in Borana Zone , Oromia Regional State , Ethiopia. *Trop Anim Health Prod*, 45, pp. 1589–1595.

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. *Infect Genet Evol.*, 11(6), pp. 1263–1275.

Richter, M. & Rossello, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci USA*, 106(45), pp. 19126–19131.

Riley, D. R., Angiuoli, S. V, Crabtree, J., Hotopp, J. C. D. & Tettelin, H. (2012). Using Sybil for interactive comparative genomics of microbes on the web. *Bioinformatics*, 28(2), pp. 160–166.

Roberts, A. P. & Mullany, P. (2011). Tn916-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. *FEMS microbiology reviews*, 35(5), pp. 856–871.

RUpp, R. & BOichard, D. (2003). Review article Genetics of resistance to mastitis in dairy cattle. *Veterinary research*, 34, pp. 671–688.

Sahl, J. W., Matalka, M. N. & Rasko, D. A. (2012). Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments. *Applied and environmental microbiology*, 78(14), pp. 4884–92.

Schleifer, K.-H. & Bell, J. A. (2009). Genus I. Staphylococcus Rosenbach 1884, 18AL. In: De Vos P, Garrity G, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, W. W. (Ed) *Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 3*. pp. 392–401. New York.

Schroeder, J. W. (2012). Bovine Mastitis and Milking Management., North Dakota State University, Fargo, North Dakota, USA, 2012. North Dakota State University, Fargo, North Dakota, USA.

Seib, K. L., Zhao, X. & Rappuoli, R. (2012). Developing vaccines in the era of genomics: a decade of reverse vaccinology. *Clinical microbiology and infection*, 18(5), pp. 109–116.

Sender, G., Korwin-Kossakowska, A., Pawlik, A., Hameed, K. G. A. & Oprządek, J. (2013). Genetic basis of mastitis resistance in dairy cattle – a review. *Annals of Animal Science*, 13(4), pp. 663–673.

Singer, A. J. & Talan, D. A. (2014). Management of skin abscesses in the era of methicillin-resistant Staphylococcus aureus. *The New England journal of medicine*, 370(11), pp. 1039–47.

Smyth, D. S., Feil, E. J., Meaney, W. J., Hartigan, P. J., Tollersrud, T., Fitzgerald, J. R., Enright, M. C. & Smyth, C. J. (2009). Molecular genetic typing reveals further insights into the diversity of animal-associated Staphylococcus aureus. *Journal of medical microbiology*, 58(Pt 10), pp. 1343–53.

Sori, H., Zerihun, A. & Abdicho, S. (2005). Dairy Cattle Mastitis In and Around Sebeta, Ethiopia Title. *Intern J Appl Res Vet Med*, 3(4), pp. 332–338.

Sprague, L. D., Al-Dahouk, S. & Neubauer, H. (2012). A review on camel brucellosis: a zoonosis sustained by ignorance and indifference. *Pathogens and global health*, 106(3), pp. 144–9.

Stein, L. (2001). Genome annotation: from sequence to biology. *Nature reviews. Genetics*, 2(7), pp. 493–503.

Stiles, D. (1987). Camel vs Cattle Pastoralism: Stopping desert spread., 1987. pp. 15–21.

Sung, J. M.-L., Lloyd, D. H. & Lindsay, J. A. (2008). Staphylococcus aureus host specificity: comparative genomics of human versus animal isolates by multi-strain microarray. *Microbiology (Reading, England)*, 154(7), pp. 1949–59.

Tamiru, F., Alemu, S. & Tsega, A. (2013). Aerobic Microorganisms Isolated from Mastitic Bovine Milk and Their Antimicrobial Susceptibility Profiles , Ethiopia. *J Agric and Environ Sci*, 13(7), pp. 920–925.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution*, 30(12), pp. 2725–9.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1), pp. 33–36.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Naomi, L., Angiuoli, S. V, Crabtree, J., Amanda, L., Durkin, A. S., Deboy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit, I., Peterson, J. D., Hauser, C. R., Jaideep, P., Nelson, W. C., Madupu, R., Lauren, M., Dodson, R. J., Rosovitz, M. J., Sullivan, S. A., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, L., Zhou, L., Zafar, N., Khouri, H., Dimitrov, G., Watkins, K., Kevin, J. B., Connor, O., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Dennis, L., Telford, J. L., Wessels, M. R., Rappuoli, R., Heim, C. M., Thase, M. E., Daniel, N., Rush, A. J., Schatzberg, A. F., Ninan, P. T., Mccullough, P., Weiss, P. M., Dunner, D. L., Barbara, O., Page, C., Charvin, D., Vanhoutte, P., Borrelli, E. & Caboche, J. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae : Implications for the microbial "pan-genome." *Proc Natl Acad Sci USA*, 102(45), pp. 13950–13955.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Eisen, J. A., Peterson, S., Wessels, M. R., Paulsen, I. T., Nelson, K. E., Margarit, I., Read, T. D., Madoff, L. C., Wolf, A. M., Beanan, M. J., Brinkac, L. M., Daugherty, S. C., DeBoy, R. T., Durkin, A. S., Kolonay, J. F., Madupu, R., Lewis, M. R., Radune, D., Fedorova, N. B., Scanlan, D., Khouri, H., Mulligan, S., Carty, H. A., Cline, R. T., Van Aken, S. E., Gill, J., Scarselli, M., Mora, M., Iacobini, E. T., Brettoni, C., Galli, G., Mariani, M., Vegni, F., Maione, D., Rinaudo, D., Rappuoli, R., Telford, J. L., Kasper, D. L., Grandi, G. & Fraser, C. M. (2002). Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V Streptococcus agalactiae. *Proc Natl Acad Sci USA*, 99(19), pp. 12391–12396.

Ubeda, C., Maiques, E., Barry, P., Matthews, A., Tormo, M. A., Lasa, I., Novick, R. P. & Penadés, J. R. (2008). SaPI mutations affecting replication and transfer and enabling autonomous replication in the absence of helper phage. *Molecular microbiology*, 67(3), pp. 493–503.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C. & Médigue, C. (2006). MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res USA*, 34(1), pp. 53–65.

Wareth, G., Murugaiyan, J., Khater, D. F. & Moustafa, S. A. (2014). Subclinical pulmonary pathogenic infection in camels slaughtered in Cairo, Egypt. *Journal of infection in developing countries*, 8(7), pp. 909–13.

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E., Vonstein, V., Warren, A., Will, R., Wilson, M. J. C., Yoo, H. S., Zhang, C., Zhang, Y. & Sobral, B. W. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic acids research*, 42, pp. D581–91.

Webber, M. A. & Piddock, L. J. V (2002). The importance of efflux pumps in bacterial antibiotic resistance. *Journal of Antimicrobial Chemotherapy*, 51(1), pp. 9–11.

Werner, G., Freitas, A. R., Coque, T. M., Sollid, J. E., Lester, C., Hammerum, A. M., Garcia-Migura, L., Jensen, L. B., Francia, M. V, Witte, W., Willems, R. J. & Sundsfjord, A. (2011). Host range of enterococcal vanA plasmids among Gram-positive intestinal bacteria. *The Journal of antimicrobial chemotherapy*, 66(2), pp. 273–82.

Wertheim, H. F. L., Melles, D. C., Vos, M. C., Leeuwen, W. Van, Belkum, A. Van, Verbrugh, H. A. & Nouwen, J. L. (2005). Subscription Information : Review The role of nasal carriage in Staphylococcus aureus infections. *Lancet Infect Dis*, 5(December), pp. 751–762.

Whiley, R. & Hardie, J. (2009). Genus I. Streptococcus Rosenbach 1884, 22AL. In: De Vos P, Garrity G, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, W. W. (Ed) *Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 3*. pp. 655–684. New York.

Wright, G. D. (2011). Molecular mechanisms of antibiotic resistance. *Chemical Communications*, 47, pp. 4055–4061.

Yeruh, I., van Straten, M. & Elad, D. (2002). Entropion, Corneal Ulcer and Corneal Haemorrhages in a One-Humped Camel (Camelus dromedarius). *Journal of Veterinary Medicine, Series B*, 49(8), pp. 409– 410.

Younan, M., Ali, Z., Bornstein, S. & Müller, W. (2001). Application of the California mastitis test in intramammary Streptococcus agalactiae and Staphylococcus aureus infections of camels (Camelus dromedarius) in Kenya. *Preventive veterinary medicine*, 51(3-4), pp. 307–16.

Younan, M. & Bornstein, S. (2007). Papers & Articles Lancefield group B and C streptococci in East African camels (Camelus dromedarius). *Veterinary Record*, 160, pp. 330–335.

Younan, M., Bornstein, S. & Gluecks, I. V (2007). Peri-arthricular abscesses in camel calves in North Kenya. *J Camel Pract Res* , 14, pp. 161–164.

Yuan, Z., Chu, G., Dan, Y., Li, J., Zhang, L., Gao, X., Gao, H., Li, J., Xu, S. & Liu, Z. (2012). BRCA1: a new candidate gene for bovine mastitis and its association analysis between single nucleotide polymorphisms and milk somatic cell score. *Molecular biology reports*, 39(6), pp. 6625–31.

Zerbino, D. R. & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5), pp. 821–9.

Zhang, L. P., Gan, Q. F., Ma, T. H., Li, H. D., Wang, X. P., Li, J. Y., Gao, X., Chen, J. B., Ren, H. Y. & Xu, S. Z. (2009). Toll-like receptor 2 gene polymorphism and its relationship with SCS in dairy cattle. *Animal biotechnology*, 20(3), pp. 87–95.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. (2011). PHAST: A Fast Phage Search Tool. *Nucl. Acids Res.*, 39, pp. W347–W352.

Zubair, S. (2010). Whole genome assembly , annotation and bioinformatics analysis of streptococcus agalactiae isolated from cow suffering of mastitis. 322, pp. 1–35.

Zubair, S., Villiers, E. P. De, Fuxelius, H. H., Andersson, G., Johansson, K., Bishop, R. P. & Bongcam-Rudloff, E. (2013a). Genome Sequence of Streptococcus agalactiae Strain 09mas018883, Isolated from a Swedish Cow. *Genome Announc*, 1(4), pp. e00456–13.

Zubair, S., Villiers, E. P. De, Younan, M., Andersson, G., Tettelin, H., Riley, D. R., Jores, J., Bongcam-Rudloff, E. & Bishop, P. (2013b). Genome Sequences of Two Pathogenic Streptococcus agalactiae Isolates from the One-Humped Camel Camelus dromedarius. *Genome Announc*, 1(4), pp. e00515–13.

Acknowledgements

I thank Almighty Allah Who guided, supported and blessed me in all fields of my life, including this PhD.

I would start acknowledging my four fathers in my PhD ☺

> Erik Bongcam-Rudloff, my main supervisor for always expecting higher from me, which kept me struggling hard during the course of my PhD. His comments on my thesis were valuable. I further appreciate his enthusiastic, welcoming and open attitude towards others.
>
> Göran Andersson, my co-supervisor who responded always whenever I called him in difficult times. His comments on my manuscripts and thesis were valuable and aided me to improve the manuscripts.
>
> Etienne de Villiers and Richard Bishop my co-supervisors at Kenya, whose comments on my manuscripts helped me to refine my research findings, and improve the readability of the manuscripts. I further acknowledge Etienne for his comments on my thesis.

I would acknowledge Hervé Tettelin from Institute for Genome Sciences, USA whose work on previously published GBS genomes was source of inspiration for me.

I would acknowledge Joerg Jores, Anne Jores and Mario Younan from Kenya for their comments on my work.

I would acknowledge Juliette Hayer for her comments on my thesis.

I would acknowledge Higher Education Commission (HEC) of Pakistan, for its financial support in my PhD.

I would acknowledge SLU, the members of SLU Global Bioinformatics Center (SGBC), Swedish Institute (SI), ILRI Kenya and Uppmax's computational support.

I would not miss the opportunity to acknowledge Helena Pettersson for her smiley and welcoming behaviour whenever I met her. A big smile for her ☺

I would acknowledge Pakistani community whose presence in Uppsala made me feel I am not too away from my homeland Pakistan.

I would love to acknowledge my gorgeous and superb parents, sisters and the only brother whose love and prayers were sufficient for me in spite of thousands of miles between Sweden and Pakistan. May Allah give them long, happy and healthy life!

I would acknowledge my mother-in-law and father-in-law for their prayers and support.

Last but not the least, I would acknowledge my dedicated husband Mudassir, and my lovely daughters Emaan and Fakiha who are source of peace and satisfaction in my life. May Allah bless them as well!

I wish you all a peaceful life!